MORAL SEQUENCING AND INTERVENING TO PREVENT HARM

by

BENJAMIN DAVID COSTELLO

A thesis submitted to the University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

Department of Philosophy
School of Philosophy, Theology, and Religion
College of Arts and Law
University of Birmingham
July 2018

ABSTRACT

This thesis will utilise the literature on the distinction between doing harm and allowing harm to develop a novel system of moral sequencing that can be applied to general moral problems to decide if, when, and how an agent should intervene to prevent harm from occurring to another agent. Off the back of this discussion, this thesis will offer a way of determining the responsibility of certain agents for their actions within a moral sequence. These motivations will be at the centre of the discussions in this thesis and will be accomplished by ultimately updating the system of moral sequencing so that it makes sense of practical cases of inner-agent change, which itself provides a justification for intervening earlier, and potentially diminishing the responsibility for agents who have experienced an inner-agent change, in a moral sequence.

For Emily, Anthony, and Diane

ACKNOWLEDGEMENTS

I sincerely thank Iain Law for his expertise, for offering stimulating and fruitful discussions, and for tolerating me during my most fickle stages of writing. I also thank Lisa Bortolotti for her advice on early drafts of my arguments.

I am grateful to my parents, Anthony and Diane, for all they have done for me and for always providing a listening ear. They fostered my love of philosophy, gave me the confidence to persist in my academic endeavours, and inspired me to pursue my ambitions.

Finally, I thank my wife, Emily, for her endless love and encouragement. She always supported me when I was worried, motivated me when I lost interest, and made sure I had a fresh cup of coffee when I needed it most.

TABLE OF CONTENTS

INTRODUCTION

Jones is known for glassing people at his local bar. The Police have been tipped-off that Jones will be visiting the bar this evening; they believe that once he is there Jones will glass a man, Smith, with whom he has previously quarrelled. At what point during the night should the Police intervene to prevent Jones from glassing Smith? Can Jones be detained before entering the bar, only when he has a drink in his hand, only when he walks towards Smith with a glass-in-hand, or only when Jones throws the glass at Smith? Moreover, should we attribute responsibility to Jones for his actions? And what if the Police decided not to intervene and, as a result, Smith was seriously harmed—should we say they are at least partially responsible for Smith's injuries?

Providing answers to the questions above relies on constructing a novel structure of moral sequencing that can be used to assess general moral problems. Moral sequencing will allow us to determine if and when we should intervene to prevent the occurrence of harm and will permit us to make an assessment of an agent's responsibility for his actions.

The motivation for this thesis is therefore to develop a novel system of moral sequencing that can be applied to general moral problems to decide if and when an agent should intervene to prevent harm from occurring to another agent; and, off the back of this discussion, provide an account of the responsibility that can be attributed to certain agents for their actions.

Chapter 1 will start by providing an overview of the literature on the distinction between doing harm and allowing harm in order to tease-out terminology and concepts that are

relevant to and that will be employed in the novel system of moral sequencing that I will begin to create in chapter 2. Chapter 2 will start to construct this new system of moral sequencing to extend the success of the literature on the doing/allowing distinction—most notably building on the successes of Fiona Woollard's work in which the precursors to moral sequencing can best be seen—to encompass a fuller understanding of moral sequencing that can be employed outside of the doing/allowing distinction. At this juncture the reader will have an understanding of the framework of moral sequencing, the roles different agents play in those sequences, and the different type of parameters involved. Chapter 3 will build on this by discussing how an agent should go about making a moral decision in a moral sequence and will outline the kind of moral decision that is normatively favourable over other kinds of moral decisions. This will link directly into chapter 4 which will show how an agent can use the moral decision-making process and apply it to a particular moral decision, namely deciding if and when to intervene in a moral sequence to prevent harm. This chapter will argue that an intervention is justifiable only if an intervention takes place beyond a threshold of harm, which is gauged by considering three (primary, secondary, and tertiary) narratives that account for the probability of harm eventuating, the knowledge an intervener has about other agents and the sequence, whether intervening is necessary, whether the intervention is proportionate to the threatened harm, and whether any agents harmed by an intervention are liable to that harm. Chapter 5 will complete the picture of moral sequencing presented thus far by posing and answering questions related to determining the responsibility of agents for the harms threatened to the victim of a moral sequence. Chapter 6 will at first appear to take a side-step since it will lead with presenting a potential problem for the system of moral sequencing developed throughout the thesis. It will demonstrate that moral sequencing cannot *prima facie* account for a motivation to see that agents who undergo a radical change in their personality during a moral sequence

should not be held responsible for their intra-sequence actions. This issue will be laid to rest by showing how personality changes can be accounted for in a revised system of moral sequencing. Chapter 7 will consider two upshots of revising the system of moral sequencing in light of the discussions in chapter 6. It will show why accounting for personality changes during a moral sequence will: (a) provide us with a reason (required in light of the problem case introduced in chapter 6) to *diminish* the responsibility previously attributed to an agent for their intra-sequence actions; and (b) provide a justification for intervening *earlier* in a moral sequence than would otherwise have been justified to prevent harm occurring to an agent.

CHAPTER 1:

THE HISTORY OF AND BACKGROUND TO MORAL SEQUENCING

This chapter will delve into (a) a review of the current literature that feature the foundations of moral sequencing and (b) an assessment and critique of, what I consider to be, the most currently developed account of moral sequences, *viz.* Fiona Woollard's sequencing.

The strengths and weaknesses of Woollard's sequences will be determined, and the structure of moral sequencing to date will, in chapter 2, be deconstructed and reassembled in a way that permits moral sequencing to be used outside of the constraints of other authors' sequencing (*viz.* outside of the doing/allowing distinction).

After reviewing the accounts and terminology on which my novel moral sequencing will be built, chapter 2 will use Fiona Woollard's sequences as models and will outline ways in which the problems that I identify with her sequences can be rectified. It is the subsequent construction of my novel system of moral sequencing that will be developed, discussed, and advocated throughout this thesis.

1.1. THE USEFULNESS OF MORAL SEQUENCING

Moral sequencing can be employed to diagrammatically depict an agent's actions in a discrete time-frame, starting with an agent initiating a moral sequence and ending in either the occurrence of harm or an intervention to prevent that harm from occurring. In a moral sequence, an agent's actions are mapped-out, sequence-event by sequence-event, to predict

what that agent has done and what that agent might do. What an agent "might" do is determined by a continual assessment of the possible contingencies of harm up until what I call "the threshold of harm" is reached, after which an intervention can be justified. These predictions can be continually updated in real-time as the sequence progresses, and the proximity to the threshold of harm can be updated and recalculated with each new sequence-event. By constructing a moral sequence, one is in a position to establish whether an intervention to prevent harm is appropriate and, if it is, when it is most justifiable. After these assessments are made, moral sequencing provides the relevant mechanism for attributing responsibility to an agent for their actions in a particular moral sequence.

1.2. THE CURRENT PICTURE

Although the terminology 'moral sequencing' is not widely used in the philosophical literature, its roots can be found in the work of authors who have written on areas that either implicitly or explicitly employ moral sequencing. Topics such as the *Doctrine of Double Effect* and the *Distinction between Doing Harm and Allowing Harm* are built on or involve moral sequencing broadly construed. This section will present an overview of the current philosophical literature on the distinction between doing harm and allowing harm, including a discussion on the distinction between killing and letting die, a widely discussed sub-topic of the broader doing/allowing distinction that has received much attention both in theoretical and practical philosophy. The two are related in so much as 'killing is an instance of *doing*, or directly causing an event to occur, while letting die is an instance of *allowing* an event to occur' (McMahan, 1993: 250). The doing/allowing distinction is not only discussed in the philosophical literature, it has philosophical import (and is significant) in other areas too, particularly in legal, medical, and religious contexts, and can be brought to bear on real-life

situations such as abortion and euthanasia (e.g. Foot, 1994; Rachels, 1975; Rachels, 1989; Sullivan, 1977)[1].

It is worth noting that this distinction is, in some literature, described as the distinction between making (or making happen) and allowing (or letting happen) (Bennett, 1993; 1995), acting and refraining (Fitzgerald, 1967), action and inaction (Quinn, 1989), action and omission (Lichtenberg, 1982), amongst others (Norcross, 1994: 9), although there are, on some readings, subtle differences between them. There is, for example, a debate surrounding whether 'initiating' and 'sustaining' a sequence should be regarded as instances of 'doing' (acting), and whether 'enabling' a sequence to continue and 'forbearing to prevent' a sequence from stopping should be regarded as instances of 'allowing' (omitting) (see Woollard, 2015), and whether there is a difference between doing and action, and allowing and inaction[2]. To discuss the subtleties in difference between these accounts, and to assess whether the doing/allowing distinction should be defined and understood as one of the alternatives mentioned, would be to detract from the drive of this chapter, namely to understand the system of moral sequencing in the context of the doing/allowing distinction. The doing/allowing distinction is discussed only to contextualise the importance of moral sequencing in the philosophical literature, and not to necessarily support any author's conclusions on how the doing/allowing distinction should be understood. I concede that it

---

[1] See Steinbock and Norcross (1994: 1–23) for a good overview of the legal, medical, and religious aspects of the doing/allowing distinction.

[2] I agree with Woollard (2015: 11) that 'the conflation of the action/inaction distinction and the doing/allowing distinction has unfortunate consequences. It has led to unnecessary criticism of potential analyses […] [and] a proliferation of false counterexamples'. For a good discussion of the distinction between doing/allowing and action/inaction, drawing on Jeff McMahan (1993; 1998) and Jonathan Bennett (1981; 1995), see Woollard (2015: 8–11).

is important to carefully demarcate terminological distinctions, and that there is much to be said for understanding how these terminological differences impact on the Doctrine of Doing and Allowing; but these discussions are only relevant to an account of the Doctrine of Doing and Allowing and are not germane to moral sequencing[3].

After the scene has been set, and the current literature on the doing/allowing distinction has been presented, it will become clear how Fiona Woollard's work—which I take to be the most currently developed system of moral sequencing—can provide a firm basis from which to understand and develop moral sequencing, but devoid of its attachment to the broader doing/allowing distinction.

Bonnie Steinbock (1994: 24) uses the poignant case of Kitty Genovese[4], a 28-year-old woman who was stabbed to death near her home in New York, to highlight what many might consider the 'common-sense morality' (c.f. McMahan, 1993) or 'everyday moral intuition' (c.f. Foot, 1984) that 'it is worse to kill a person than to let him or her die'. During Kitty's murder, thirty-eight people heard her scream, but not one of them intervened, called the police, or acted in a way that demonstrated any interest in the commotion. For this reason, Steinbock (1994: 24) (I think rightly) accuses at least some of these thirty-eight people (depending on their awareness of, and proximity to, the incident) of letting Kitty die. But

---

[3] For instance, Jonathan Bennett's (1995) account of 'making' and 'allowing' is especially important and relevant to the Doctrine of Doing and Allowing. However, such discussions are only pertinent to the debates surrounding how best to understand the Doctrine of Doing and Allowing, and have little import into discussions of moral sequencing, which, instead of ascertaining whether there is a (morally significant, etc.) difference between doing harm and allowing harm, is concerned with ascertaining the link between agency, action, intervention, responsibility, and (non-pre-existing) threats.

[4] It is worth noting that the facts of this case have been disputed; see Cook (2014) for a thorough account of the case and its controversies. However, Steinbock's presentation of the case is adequate for current purposes.

the question remains: even if we accuse a bystander of letting Kitty die, is this less, equally, or more morally significant than if the bystander had killed Kitty himself? And does this matter, philosophically speaking?

Much of the current literature on the doing/allowing distinction presents a thesis—either arguing that there is or that there is not a morally relevant distinction between doing harm and allowing harm—in which a pair of cases are examined to determine whether the factual differences in the cases (but where both lead to the same harmful consequence) are morally significant. This is determined by drawing a conclusion about whether the two cases are morally equivalent (i.e. that the conduct in each case is morally equivalent to the conduct in the other case) or are morally different (i.e. that the conduct in each case is more or less morally bad than the other case). The debate can therefore be roughly divided into two positions[5]: those who hold that the distinction *is* morally significant (e.g. Quinn, 1989)) and those that hold that the distinction *is not* morally significant (e.g. Rachels, 1975; Tooley, 1994). It is worth noting that some commentators have argued that underlying the debate on the moral significance of the doing/allowing distinction is 'a battle between consequentialism and absolutism' (Steinbock, 1994): an absolutist might argue that the distinction is morally significant since killing, but not necessarily allowing to die, is always morally wrong; a consequentialist might argue that, if the only factor of difference in the pair of cases is that one involves an agent acting and the other involves an agent refraining

[5] The debate is, of course, more complicated and intricate than this dual position suggests; some authors, for instance (Rachels, 1975; Tooley, 1994), argue that this distinction is not morally significant, but that some other similar distinction is. However, since this section is primarily concerned with presenting an overview of the current literature concerning the moral significant of this distinction, a commentary on and analysis of these alternative distinctions will not be discussed.

from acting, then, since the consequence of the pair of cases is the same (*viz.* harm occurs), the doing/allowing distinction is not morally significant. However, since the drive of this chapter is to outline the current picture of moral sequencing and establish a new system of moral sequencing, the consequentialist/absolutist debate will not be discussed any further[6].

1.2.1. FOOT AND THE RESCUE CASES

Philippa Foot's work (1984; 1985; 1994) provides the foundations for the system of moral sequencing that I will present in chapter 2, and is the first instance in the debate on the doing/allowing distinction in which the relevant terminology (i.e. 'initiated', 'sustaining', 'forbearing to prevent', 'sequence', etc.) and core tenets (i.e. the importance of agency) of moral sequencing are featured.

Foot (1984) investigates what it means for an agent to be the cause of harm to another agent[7]. Foot suggests that one must make a distinction between 'what one does or causes and what one merely allows' (Foot, 1994: 273) and drawing this distinction can be achieved if it is understood that 'one person may or may not be 'the agent' of harm that befalls someone else' (Foot, 1984: 178). On Foot's account, person *A* is 'the agent' of person *B*'s harm if and only if *A initiated* or *sustained* a sequence that caused harm; doing harm relies on ascertaining that the harm *originated from* or was *sustained by A* (Foot, 1984: 179). For

---

[6] For a good discussion of the consequentialist/absolutist debate in the doing/allowing distinction, see Bennett (1966).

[7] It is worth noting that although this thesis is primarily concerned with the occurrence of harm or, to use the terminology in the literature, 'pre-existing threats' in (moral) sequences (Woollard, 2008; 2015), and despite the fact that Foot primarily discusses the doing/allowing distinction in relation to negative outcomes (in sequences), the doing/allowing distinction is also relevant in discussions of positive and neutral outcomes (see Woollard (2008: 263) and Bennett (1995: 123)).

Foot, a harmful sequence is initiated when an agent is the origin of a harmful sequence, and a harmful sequence is sustained when an agent prolongs a sequence that would have come to a halt had the agent not acted. An agent allows the occurrence of harm, under Foot's definition, by either *enabling* the occurrence of harm ('the removal of some obstacle which is, as it were, holding back a train of events' (Foot, 1994: 273)) or *forbearing to prevent* the occurrence of harm ('For this we need a sequence thought of as already in train, and something the agent could do to intervene. (The agent must be able to intervene but does not do so.)' (Foot, 1994: 273)). Foot (1984: 179) presents a pair of cases in 'Killing and Letting Die' to help solidify her claim:

> *Rescue I*
>
> 'Suppose there are a group of people [*saving group*] hurrying to save five people [*distressed five*] who are imminently threatened by an ocean tide. There is not a moment to spare, so when the group of people hear of a single person [*distressed individual*] who also needs rescuing from another disaster, the group regretfully cannot rescue this person as it will prevent them from saving the other five. Therefore, they leave the individual to die'.

Rescue II

> ‘Once again a group of people [*saving group*] are hurrying to save five people [*distressed five*] from imminent threat of the ocean tide. In this version of the story, the lone individual [*distressed individual*] is trapped on the path. If the group of people are to rescue the five, they will have to drive over the individual. If the group do not drive over the individual the other five will drown but the individual will be safe. The group of people however decide to drive over the individual and save the other five’.

Foot draws the following conclusions from the two cases above. In Rescue I, the distressed five are saved because the saving group *allow* the distressed individual to die; the saving group merely *allow harm*. In Rescue II, however, the distressed five are saved because the saving group are the *cause* of the death of the distressed individual; the saving group *do harm*. The saving group act morally permissibly in Rescue I as they do not cause the death of the distressed individual; by saving the distressed five, the saving group merely allow harm to befall the distressed individual. But in Rescue II, the saving group act morally impermissibly as they cause the death of the distressed individual; by saving the distressed five, the saving group cause harm to the distressed individual.

For Foot, the importance of accounting for agency in the doing/allowing distinction is shown in that ‘it makes all the difference whether those who are going to die [or be harmed] if we act in a certain way will die [or be harmed] as a result of a sequence that we originate or one that we allow to continue, it being of course something that did not *start* by our agency’ (Foot, 1984: 180). So, according to Foot, a person can only be said to do harm if he

originates or sustains a sequence with a harmful consequence. I agree that a moral sequence, or, for current purposes, the doing/allowing distinction, should be understood as necessarily containing a relationship between an agent and the harm that occurs (and it is worth noting, for the discussion in chapter 2, that Fiona Woollard (2015: 22) agrees). However, as I will argue in chapter 2, we should extend our understanding of doing harm to those that initiate a moral sequence or forbear to prevent harm (and to those who "snowball" a moral sequence, c.f. §2.2.2.2.). It also seems sensible to claim that when we are presented with a situation in which harm has occurred—which is inevitably surrounded by a number of possible factors and considerations that could have been causal or even deciding factors in the occurrence of harm—we ascertain what information is relevant to explaining the sequence of events that led to harm. Indeed, this is what will drive my discussion of moral sequencing properly understood—including sequence parameters—in chapter 2.

By analysing sequence-events s^1 and s^2 in sequence s, it is possible to explain how s^3 occurred; the agent, *viz.* the originator of the sequence, can be identified. In Foot's words, 'we think of particular effects as the result of particular sequences […] This idea is implied in coroners' verdicts telling us what someone died of' (Foot, 1984: 178–179). If a coroner investigating the death of a man can show that 'the subject died by poisoning and it was I who put the poison into his drink, then I am the agent of his death' (Foot, 1984: 179). Foot's insistence on accounting for agency in the doing/allowing distinction ultimately equips an assessor, in this case the coroner, with the ability to see a multitude of past events and 'pick out the fatal sequence and go on to ask who initiated it' (Foot, 1984: 179). On Foot's account, doing harm is therefore more morally significant than allowing harm, since if it were not for an agent's actions, the harmful outcome would not have occurred. The man died because I poisoned him; I took the poison out of my pocket, poured it into the man's

glass when he wasn't looking, and gave the poisoned drink to the man, who then consumed it. But 'Is killing, in itself, worse than letting die?' (Rachels, 1975: 79). Rephrased for our purposes, the question stands: Is doing harm, in itself, worse than allowing harm?

1.2.2. RACHELS AND THE WICKED UNCLE CASES

Consider the following two *Wicked Uncle Cases*[8] presented by James Rachels (1975: 78–80):

> *Wicked Uncle Smith*
>
> 'Smith stands to gain a large inheritance if anything should happen to his six-year-old cousin. One evening while the child is taking his bath, Smith sneaks into the bathroom and drowns the child, and then arranges things so that it will look like an accident.'

[8] I use the term 'Wicked Uncle Cases' in line with its common use in the literature, and despite the fact that the six-year-old boys described in the cases are Smith's and Jones' young cousins and not their nephews. I agree with Fiona Woollard's (2015: 12) speculation that 'the error occurred because in many families a child would refer to his or her adult cousin [or even a close family friend] as 'Uncle''.

Wicked Uncle Jones

> ‘Jones also stands to gain if anything should happen to his six-year-old cousin. Like Smith, Jones sneaks in planning to drown the child. However, just as he enters the bathroom Jones sees the child slip and hit his head, and fall face down in the water. Jones is delighted; he stands by, ready to push the child’s head back under if it is necessary, but it is not necessary. With only a little thrashing about, the child drowns all by himself, “accidentally,” as Jones watches and does nothing’.

We can draw some initial conclusions from the Wicked Uncle Cases. Parallels can be drawn with Rescue I and Rescue II. In Rescue I, the saving group allowed harm to the distressed individual by allowing him to die, and Wicked Uncle Jones *allows* his six-year-old cousin to be harmed by not lifting his cousin’s head out of the water. Likewise, in the same way that the saving group did or caused harm to the distressed individual by driving over him in Rescue II, Wicked Uncle Smith *does* or *causes* harm to his six-year-old cousin by drowning him. The case of Wicked Uncle Jones is therefore, even on Foot’s account, a case of allowing harm, and the case of Wicked Uncle Smith is a case of doing (or causing) harm. This is where Rachels departs from Foot. If we make the reasonable assumption that both Smith’s and Jones’ reasons for acting are identical (the personal gain of inheriting a large fortune) and that Jones would have acted as Smith did had his cousin not slipped and hit his head, and we make the further reasonable assumption that the consequence is identical in both cases (the six-year-old cousin dies), then, Rachels argues, the fact that the two cases differ in how the harmful consequence occurred (*viz.* Smith did harm whereas Jones allowed harm) has no bearing on the moral difference between the two cases. In short, ‘Smith killed

the child, whereas Jones "merely" let the child die. That is the only difference between them' (Rachels, 1975: 79). From this, Rachels concludes that '[i]f the difference between killing and letting die [or doing harm and allowing harm] were in itself a morally important matter, one should say that Jones's behavior was less reprehensible than Smith's' (Rachels, 1975: 79–80) since Jones did nothing except walk into the bathroom and watch, 'with delight', as his cousin drowned; Jones merely allowed harm by refraining from lifting his cousin's head out of the water. Yet Jones could have, had he wanted to, acted to prevent his cousin from drowning; it is arguably this realisation that compels Rachels to call any appeal to Jones' case (embodying allowing harm/letting die) as demonstrating that allowing harm is less morally reprehensible in itself than doing harm a 'grotesque perversion of moral reasoning' (Rachels, 1975: 80). Wicked Uncle Smith and Wicked Uncle Jones are therefore alike except for the fact that the former is a case of killing (doing harm) and the latter is a case of allowing to die (allowing harm). So once all the relevant factors have been eradicated bar the fact that one is a case of doing harm and the other a case of allowing harm, both Smith's and Jones' behaviour are equally morally bad, and neither doing harm nor allowing harm is more morally reprehensible than the other. Thus, according to Rachels, the Wicked Uncle Cases show that the doing/allowing distinction is not morally significant (and that if one believes that doing harm is morally worse than allowing harm, or *vice versa*, then other factors other than the fact that one is a case of doing harm and the other a case of allowing harm must also be at play).

1.2.3. AGAINST RACHELS: KAGAN AND THE ADDITIVE FALLACY

Although an evaluation of Rachels' argument is not pertinent to the discussion of moral sequencing *per se*, it is worth noting that Rachels' conclusions have been contested. There

are too many objections to mention in detail[9], but there is one that is appropriate to the current discussion. Beauchamp (1979) claims that the fact that factor *f* is morally relevant in context c^1 does not necessitate that *f* is or must be morally relevant in c^n. Shelly Kagan (1988) builds on this and directs his line of attack against Rachels' use of the 'contrast strategy'[10,11]. Kagan accuses Rachels of committing the 'additive fallacy'; Rachels' argument relies on the *additive assumption* that 'the status of the act is the net balance or sum which is the result of adding up the separate positive and negative effects of the individual factors' (Kagan, 1988: 259). A contradiction can arise when using the contrast strategy for two pairs of cases. Take a pair of cases in which one case is a case of doing harm (A^d) and the other is a case of allowing harm (A^a), the two cases differ only in so far as A^d is a case of doing harm (indicated by the superscript '*d*') and A^a is a case of allowing harm (indicated by the superscript '*a*'). Kagan's claim is that if instances of doing harm (^{d}) and allowing harm (^{a}) are introduced into another pair of cases, say B^d and B^a, where the

---

[9] Frances Myrna Kamm (1983), for instance, argues that the doing/allowing distinction sometimes is and sometimes is not morally significant. Kamm presents four 'conceptual components' of allowing harm or letting die that are not necessarily components of killing or doing harm (Kamm, 1983: 301) to show that 'introducing a property conceptually true of letting die [or allowing harm] into a case of killing [or doing harm] might make the particular killing [or act of harming] more easily justified than killing [or doing harm] in a case which lacked the property'. For a full account, see Kamm (1983; 1986; 2007). For a good overview of Kamm's account, see Woollard (2015).

[10] Woollard provides a good explanation of the contrast strategy (2012: 461): 'This strategy involves putting forward contrasting pairs [i.e. Rescue I and Rescue II, and Wicked Uncle Smith and Wicked Uncle Jones], which are identical in all features except that one involves doing harm while the other involves merely allowing harm. The idea is that we will see a moral difference between the cases if and only if the doing/allowing distinction is morally significant in itself. If the cases are morally equivalent, then the doing/allowing distinction is not morally significant in itself'.

[11] For Kagan, the doing/allowing distinction is a 'dangling distinction' (Kagan, 1989: 14); it is a distinction that explains any moral intuition that doing harm is more morally significant that allowing harm, whilst remaining unexplained by whatever moral theory we hold. For a good response to Kagan, and for a good defence of the contrast strategy, see Heidi Malm (1992).

two cases of *B* differ only in so far as B^d is a case of doing harm (indicated by the superscript '*d*') and B^a is a case of allowing harm (indicated by the superscript '*a*'), then the conclusion of the moral significance or moral insignificance of the doing/allowing distinction for A^d and A^a is not *necessarily* the same conclusion for B^d and B^a.

A good example of Kagan's account in action, and an example that demonstrates how the contrast strategy can generate contradictory outcomes, is Woollard's (2012: 461) comparison of the Mountain Rescue Cases[12] and the Wicked Uncle Cases. Woollard compares these two pairs of cases to show that 'it is possible to find pairs of contrast cases which do seem morally inequivalent' (Woollard, 2012: 461). By showing that, in the Mountain Rescue Cases, 'it is clearly permissible to refuse to stop and free Charlie in the first case and impermissible to push the boulder towards Charlie in the second case' (Woollard, 2012: 461), Woollard seeks to undermine Rachels' contrast strategy that supposedly shows that the results of the Wicked Uncle Cases extend to other pairs of cases in which the only significant differential[13] is that one case in the pair is a case of doing harm

---

[12] It is important to note that Woollard's Mountain Rescue Cases are different from Foot's (1984) cases of Rescue I and Rescue II. The Mountain Rescue Cases can be described as follows: 'In the first case, you are driving Alastair and Bryan to the hospital for life-saving treatment. You see that Charlie is trapped on the hillside. A boulder is rolling towards him and he will be crushed to death by it unless you save him. You could save him, but it would delay you so that it would be too late to save Alastair and Bryan. You drive on. In the second case, the boulder is blocking the route to the hospital. The only way to get to the hospital is to push the boulder towards Charlie, who is trapped on the hillside. You push the boulder. In both cases, you must choose whether Charlie is crushed to death by the boulder or Alastair and Bryan die from their injuries' (Woollard, 2012: 461).

[13] It has been pointed out to me that there are many factual differences between the Mountain Rescue Cases and the Wicked Uncle Cases. This is true, but these are not the differentials that Woollard is referring to. Woollard grants that these factual differences are unavoidable, however they do not detract from the claim that the only *significant* difference, relevant to the current discussion, is that one case in the pair is a case of doing harm and the other is a case of allowing harm.

and the other is a case of allowing harm. Even if we agree with Rachels that the pair of Wicked Uncle Cases demonstrates that the doing/allowing distinction is not morally significant (and that the two cases are therefore morally equivalent), other pairs of cases show how the converse conclusion can be drawn; the Mountain Rescue Cases exemplify how the doing/allowing distinction does make a moral difference and is in fact morally significant (and that the two cases are not morally equivalent). We are therefore left with the conclusion that Rachels is wrong and that 'the contrast strategy does not work because it wrongly assumes that if a factor makes a moral difference anywhere', like in the Wicked Uncle Cases, then 'it will make the same moral difference everywhere', i.e. it will make a moral difference in the Mountain Rescue Cases (Woollard, 2012: 461)[14].

1.2.4. FOOT AND THE MORALLY SIGNIFICANT DIFFERENCE BETWEEN DOING AND ALLOWING: NEGATIVE RIGHTS/DUTIES AND POSITIVE RIGHTS/DUTIES

This now leaves the door open to return to Foot's account. In addition to there being a *moral distinction* between doing harm and allowing harm—namely that being an agent of harm requires initiating or sustaining a series of events that results in harm—there is a *morally significant difference* between doing and allowing harm (or rather, for the purposes of Foot's paper, between killing and allowing to die). According to Foot (1984), the moral significance of the doing/allowing distinction can be ascertained by establishing a connection between certain agents and certain rights, *viz.* linking a doing-agent and an

[14] There have been a number of other criticisms and objections leveled at Rachels. For an overview, see Woollard (2015).

allowing-agent with either a 'positive duty/right' or a 'negative duty/right'. Foot explains that 'there are rights to noninterference […] and there are also rights to goods or services', and that, although these rights can be overridden in exceptional circumstances[15], 'it takes more to justify an interference than to justify the withholding of goods or services' (Foot, 1984: 180–181). Doing harm typically involves violating another's right to non-interference, and this right is more morally binding than allowing harm, which typically only involves violating another's right to goods or services (i.e. the 'organ transplant procedures' (Harris, 1975: 81)). The right to non-interference is a *negative right*; agent *A* has a *negative duty* to not harm agent *B*. (Property rights are an example of a negative right: 'others have in ordinary circumstances a [negative] duty to not interfere with our property' (Foot, 1984: 181)). The right to goods or services is a *positive right*; agent *A* has a *positive duty* to not allow agent *B* to be harmed.

Foot (1984: 181) claims that, since the 'violation of a right to non-interference consists in interference, which implies breaking into an existing sequence and initiating a new one', denying someone a right to goods or services is easier to justify and that right is less morally binding than denying someone a right to non-interference. So, whilst letting die (allowing harm) involves violating someone's right to goods or services, killing (doing harm) involves violating someone's right to non-interference. This will be important for my discussion in chapter 2, since violating a negative right, on Foot's account, implies that a new sequence has been initiated, whilst violating a positive right does not cause the initiation of a new sequence, rather it enables it to continue to a harmful outcome or forbears to prevent the

---

[15] Foot cites Elizabeth Anscombe's example of preventing the spread of a fire by destroying someone's house as an instance of what she considers to be 'exceptional circumstances' in which 'the right is overridden' (Foot, 1984: 181).

sequence from reaching the harmful outcome[16]. For Foot, then, the moral significance of the doing/allowing distinction lies in the fact that negative duties are stronger and more morally binding than positive duties. Foot's conclusion is therefore that initiating or sustaining a sequence (doing harm) is less morally permissible than enabling a sequence to continue or forbearing to prevent a harmful outcome (allowing harm), and this is explained by appealing to the morally significant difference between doing harm and allowing harm. The morally significant difference between Rescue I (a case of allowing harm) and Rescue II (a case of doing harm), for instance, lies in the fact that in Rescue I the saving group's position is not one in which they have to decide whether to negate the distressed individual's negative right (to non-interference, i.e. to not be harmed); the saving group merely fail to fulfil their positive duty to save the distressed individual. Whereas in Rescue II, the saving group are asked to, and do, negate the distressed individual's negative right (to non-interference); the saving group fail to fulfil their negative duty to not harm the distressed individual.

The ramifications of Foot's account for moral sequencing is that driving over the distressed individual in Recue II violates his right to non-interference and, by doing so, initiates a new sequence, and this new sequence results in doing harm to (killing) another agent. In Rescue I, however, because the distressed individual was already under a pre-existing threat of harm (from 'another disaster') the saving group do not initiate or sustain a sequence, and the

---

[16] Foot (1984: 181) does mention that there are 'cases in which the right to non-interference exists and is not overridden, but where the right to service either does not exist or is overridden': 'it often happens that whereas someone's rights stand in the way of our interference, we owe him no service in relation to that which he would lose if we interfered. We may not deprive him of his property, though we do not have to help him secure his hold of it'.

saving group's act of allowing harm (death) to befall the distressed individual is not the cause of his harm (death).

What is missing from Foot's account, however, is a compelling theory of rights that can explain why negative rights should precede positive rights. Without such an explanation Foot's account is left wanting; if asked *why* negative duties are more morally significant than positive duties, Foot's only conceivable response would be "Because doing harm is morally worse than allowing harm", but then this leads to a circular argument—"But why is doing harm morally worse than allowing harm?" "Because negative duties are more morally significant than positive duties!" And so on[17].

1.2.5. QUINN AND THE PRECEDENCE OF NEGATIVE RIGHTS

Warren Quinn (1989) is sympathetic to Foot's claim that negative rights (the right to non-interference) are more important, morally speaking, than positive rights (the right to goods or services), and criticises Rachels' claim about the moral insignificance of the

---

[17] It is worth noting that Frances Myrna Kamm (1986: 5–11) proffers two explanations for why doing harm might be considered intrinsically worse than allowing harm: (a) there might be some morally bad feature intrinsic to doing harm that is not intrinsic to allowing harm; or (b) there might be some morally bad feature intrinsic to doing harm that is not intrinsic to allowing harm, but where this morally bad feature can emerge is in cases of allowing harm. This same point is also discussed by Quinn (1989: 289, footnote 7), in which he states that 'the idea that intrinsically nonequivalent parts must always make an overall evaluative difference when embedded in identical contexts seems wrong'. Quinn (1989: 289, footnote 7) appeals to the Wicked Uncle Cases to show that even if (b) is true, the moral equivalence of Wicked Uncle Smith and Wicked Uncle Jones does not demonstrate that all cases of doing harm and allowing harm are morally equivalent, since Wicked Uncle Jones' letting his cousin die 'might be a special case in which letting die has the bad feature essential to killings [doing harm] but not lettings die [allowing harm]'. Since the intrinsic moral value of doing harm and allowing harm is not pertinent to this chapter, this discussion will end here. For a full account, see Kamm (1986).

doing/allowing distinction. Importantly, Quinn (1989: 306) provides an explanation for why negative rights take precedence over, and are more morally important than, positive rights. He describes negative rights (in contrast to positive rights, which 'are claim rights to aid or support') as 'claim rights against harmful intervention, interference, assault, aggression, etc.'. Quinn (1989: 308) argues that there has to be a precedence of negative rights, otherwise one's body—or 'one might say his person'—'is not in any interesting moral sense *his*'. If one's negative rights are not held in the highest moral regard and do not precede positive rights, then it would be morally acceptable to use another person's body as one sees fit.

The problem, in Fiona Woollard's terms, is that '[i]f morality does not include such a constraint [against violating another's negative rights], then it treats the victim's body and mind as common property' (Woollard, 2012: 2)[18]. To use an example from John Harris' cases of the 'organ transplant procedures'[19] (Harris, 1975: 81), in order to prevent others from harvesting my organs it is essential that negative rights precede any other rights, including positive rights; if negative rights do not precede positive rights, then 'we do not have a genuine belonging [of our body] but mere association' (Woollard, 2015: 106). Woollard gives a good example to solidify this claim. If others are able to use my car to drive or to dismantle it and use the parts to fix other cars then the car cannot be said to be

---

[18] Woollard (2015: 6) provides a more refined version of Quinn's claim: 'without the constraint against doing harm a person's body may be damaged whenever this is necessary to prevent greater harm occurring to others. His interests count for no more than anyone else's in determining what may be done to his body. His body is treated as common property rather than as genuinely belonging to him'.

[19] Harris (1975: 81) describes a situation in which 'Y needs a new heart and Z new lungs. They point out that if just one healthy person were to be killed his organs could be removed and both of them be saved'.

mine, and the car is quite plainly a commonly owned car (Woollard, 2015: 106). Thus, to prevent my body or my car from becoming common property, I must have a 'privileged status' over my body/car for it to 'genuinely belong to me' (Woollard, 2015: 107)[20].

Quinn therefore provides a case against Rachels' claim that the doing/allowing distinction is morally insignificant, and provides an answer to a lingering question left by Foot's account of the moral importance of rights to non-interference and the moral difference between doing harm and allowing harm, namely "Why do negative rights take precedence over positive rights?" Quinn's answer, as we have seen, is simply that the morally significant difference between doing harm and allowing harm is tied-up in the fact that harming someone necessarily infringes on his negative right (to not be harmed), and this right must take precedence over his positive right (to claim aid or support), otherwise he cannot be said to have ownership of his own body[21].

However, as both Fiona Woollard (2015: 6–7; 2015: 106–107) and Frances Howard-Snyder (2011) identify, Quinn's account and defence of the precedence of negative rights is lacking:

---

[20] Woollard (2015: 106, emphasis added) actually goes as far as to say that 'I need *normative protection* that gives me a privileged status with respect to this body'. For Woollard, the protection of this privileged status must extend to encompass the idea that property (whether it be an object like my car or my own body) belongs to me regardless of whether others rob me of it; even if someone does rob me of my possession(s), my possession(s) still belong to me, all that has happened is that 'my privileges of possession have been infringed'. See Woollard (2015: 106–111) for a full account of normative protection and normative imposition. See also Woollard (2015: 187–205) for a defence of 'the Body Claim' and how having 'full-fledged agency' (Woollard, 2015: 196) is impossible if others do not recognise that my body belongs to me.

[21] Quinn's work and the importance of negative rights are discussed further in §4.1.

> ‘[Quinn] has not shown why we should have constraints against *doing* harm. For all Quinn has said, any other set of constraints would do just as well’ (Woollard, 2015: 7).

> ‘Quinn’s is a funny sort of defense of negative rights. Unless I'm missing something, it doesn't pick out any special feature of negative rights that makes them specially worth respecting’ (Howard-Snyder, 2011).

What both Woollard and Howard-Snyder identify is that, even though Quinn’s account enshrines negative rights in a way that importantly enables an agent’s body to belong to them, he fails to demonstrate why having constraints against *doing* harm is important, and why this *particular* set of (negative/positive) rights should take precedence over other sets of rights. Indeed, we might wish to divide rights in other ways: i.e. ‘the rights of children and the rights of adults, rights concerning the upper half of the body and rights concerning the lower half, etc.’, and then, whenever these come into conflict, we may wish to prioritise one over the other (Howard-Snyder, 2011). The former worry (of Woollard) is more important to the current discussion than the latter worry (of Howard-Snyder), since Quinn’s account gives us no solid grounds on which to build a case for the moral significance of doing harm; all Quinn gives us is an, albeit convincing, argument in support of the claim that negative rights are important.

1.2.6. BENNETT, POSITIVE AND NEGATIVE FACTS, AND AN AGENT'S POSITIVE/NEGATIVE RELEVANCE TO AN UPSHOT

Jonathan Bennett's (1967; 1981; 1993; 1995) account of the distinction between doing harm and allowing harm[22] is that there is a distinction between situations in which an agent's behaviour is 'positively relevant' to an 'upshot' and situations in which an agent's behaviour is 'negatively relevant' to an 'upshot'. Importantly, Bennett's distinction does not distinguish two kinds of action [i.e. positive acts and negative acts]: 'there are no negative actions' (Bennett, 1993: 77)[23]. Instead, the distinction correlates to a distinction between 'positive facts' and 'negative facts', which further correlate to a distinction between 'positive propositions' and 'negative propositions'. Positive facts are more informative than negative facts since positive facts describe/reveal something that *is*, whereas negative facts only describe/reveal something that *is not*: Jones throws a glass at Smith (a positive fact); Police Officer did not intervene to stop Jones glassing Smith (a negative fact). Essentially, negative facts do not provide information about the actual state of affairs; negative facts only reveal what is not a state of affairs, and therefore the use of these facts is limited to eliminating what is the case. So whilst positive facts reveal something informative about a situation and an agent's action (we know that Jones did glass Smith), negative facts do not reveal as much information about a situation or an agent's action (we do not know where Police Officer was nor what he was doing when Smith was glassed by Jones). Bennett's account of positive/negative facts is therefore useful for assessing how much information is

---

[22] Importantly, and as I have stated previously, Bennett (1995) does not draw a direct distinction between doing and allowing *per se*, but rather between action and omission or more specifically making and allowing. However, for the purposes of this chapter, I will discuss Bennett in relation to the doing/allowing distinction.

[23] See Woollard (2015: 38–39) for a good overview of why discussions of negative acts are problematic.

revealed or left out (i.e. knowing that Jones did or did not glass Smith) and how many possible other actions or states of affairs are left unexplained (i.e. knowing where Police Officer was when Jones glassed Smith). To avoid the problems associated with trying to offer a general account of positive facts and negative facts, Bennett's distinction between positive and negative facts relate only to facts about an agent's behaviour. Importantly, this relates the positive/negative fact distinction to an agent's possible movements. Analysing the distinction in this way diverts attention away from ascertaining whether a fact is informative *per se* and towards whether a fact is informative about the movements of an agent. The distinction between positive and negative facts can then be mapped on to a further distinction between positive propositions and negative propositions (Bennett, 1995: 91–95). Woollard (2015: 41) provides a succinct outline of positive/negative propositions and their relation to positive/negative facts:

> 'A proposition is a *negative proposition* about the conduct of an agent if and only if most possible movements of the agent's body are such that if he had moved that way, the proposition in question would not have been true.'

> 'A proposition is a *positive proposition* about the conduct of an agent if and only if most possible movements of the agent's body are such that if he had moved that way, the proposition in question would have been true.'

It is at this point that Bennett's (1995: 92–96) discussion of an agent's 'behaviour space' becomes most useful. Bennett represents an agent's behaviour space, *viz.* the ways in which

an agent could have mobilised himself, as a square (these are rectangles in Woollard's diagrams below), where a proposition stating that the agent moved in a certain way is represented as a point in the (square/rectangular) behavioural space (see Bennett (1995: 91) for a full account of his diagrammatical representation of behaviour space). A more simplified, but equally useful, method of visualising an agent's behaviour space is presented by Woollard (2015: 42, Figure 3.1 and Figure 3.2) in the following figures:

Figure 1: Woollard's diagram of Bennett's behaviour space showing that 'P is a negative proposition'

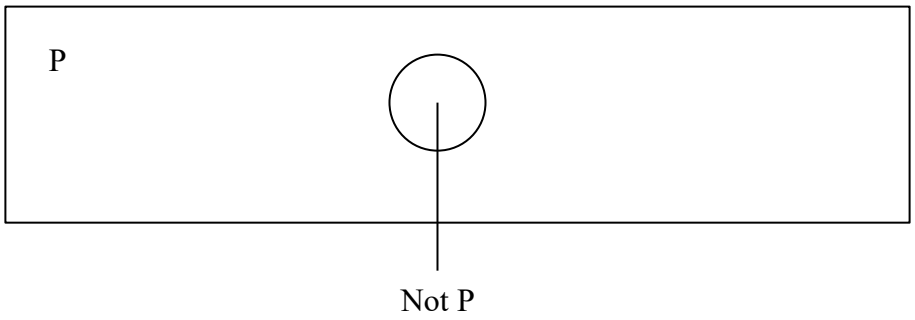
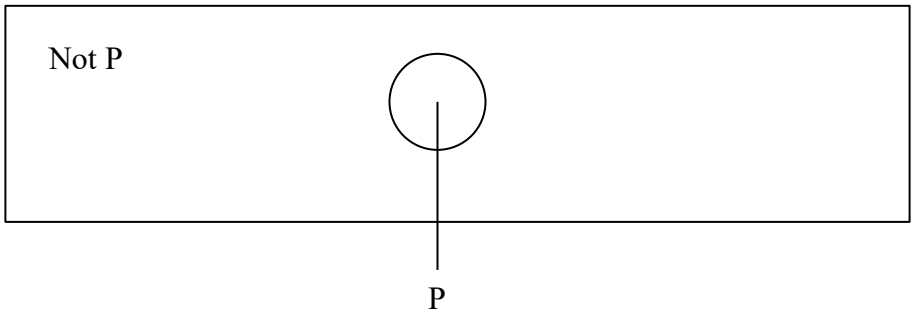


Figure 2: Woollard's diagram of Bennett's behaviour space showing that 'P is a positive proposition'



So, according to Bennett, proposition P will have a larger subspace of an agent's behaviour space than the subspace of proposition Not P when P is a negative proposition about an

agent's behaviour (represented in Figure 1), and proposition Not P will have a larger subspace of an agent's behaviour space than the subspace of proposition P when P is a positive proposition about an agent's behaviour (represented in Figure 2). When applied to the case of Jones glassing Smith, that Jones glasses Smith is a positive fact about Jones' behaviour since the proposition 'Jones glassed Smith' occupies a small region of Jones' behaviour space; the majority of movements available for Jones would not have made the proposition 'Jones glassed Smith' true. That Police Officer did not intervene to prevent Jones from glassing Smith is a negative fact about Police Officer's behaviour since the proposition 'Police Officer did not intervene to prevent Jones from glassing Smith' occupies a large region of Police Officer's behaviour space; the majority of movements available for Police Officer would have made the proposition 'Police Officer did not intervene to prevent Jones from glassing Smith' true[24].

What we are left with is a way of assessing the positive/negative fact distinction in terms of facts about the relevance of an agent's behaviour to an outcome (or rather, using Bennett's terminology, an 'upshot'). An agent's behaviour is *positively relevant* to an upshot if and only if the majority of behaviour open to that agent *would not* have resulted in that upshot. An agent's behaviour is *negatively relevant* to an upshot if and only if the majority of possible behaviour open to that agent *would* have resulted in that upshot[25]. Jones' behaviour is positively relevant to the upshot of Smith being glassed since Jones could have behaved in a number of ways that would not have resulted in him glassing Smith; Jones could have

---

[24] For a detailed (and technical) discussion and analysis of Bennett's behavior space, see Woollard (2015: 211–226, Appendix 1).
[25] See Woollard (2015: 38–51) for a good overview and discussion of Bennett's account (and the related discussion of an agent's 'behaviour space' and the 'space of possibilities' (Bennett, 1995: 92–96)).

drank his drink, ordered something else from the bar, gone to the toilet, spoken to the bartender, and so on; these behaviours, and the plethora of actions that Jones could have performed, were available to him, yet he chose to glass Smith. If, however, Jones failed to push Smith out of the way of a glass that was thrown at him then Jones' behaviour would be negatively relevant to the upshot of Smith being glassed, since the majority of Jones' actions (apart from pushing Smith out of the way or catching the glass mid-air) would have resulted in Smith being glassed.

From this, Bennett argues that there is no morally significant difference between positive facts and negative facts since 'if someone moves in a way that causes or makes probable some bad upshot, nobody would think that the moral status of his conduct depends on how many other movements by him would have done the same' (Bennett, 1995: 102)[26]. On Bennett's account, then, it seems that doing harm and allowing harm carry equal moral weight; one is not morally worse than the other: if agent *B*'s behaviour *P* causes upshot *U*, and *B* could have behaved in a number of ways that was not *P* (*viz.* ¬*P*, i.e. action *Q*) but which still resulted in *U*, then, because *U* occurs whether *B* behaves in *P* or ¬*P*, we can conclude (according to Bennett) that whether *B* behaves in *P* or ¬*P* is morally insignificant; whether *B* behaves in *P* or ¬*P*, *U* occurs. In other words, if in a case of doing harm (Case *D*) agent *B*'s behaviour *P* results in upshot *U*, and if in a case of allowing harm (Case *A*) *B*'s behaviour *Q* also results in *U*, then, because *U* occurs in both Case *D* and in Case *A*, we can

---

[26] It is worth noting that Judith Jarvis Thomson (1996: 550–551) agrees with Bennett that the positive/negative fact distinction is not morally significant, but she does add that Bennett has 'not earned' the 'stronger conclusion' that there is no moral significance between making and allowing (which, for current purposes, should be seen loosely as adhering to the doing/allowing distinction): 'he has at most earned the conclusion that the existing literature does not succeed in establishing that there is [a morally significance between making and allowing]'. For a full discussion of this, see Thomson (1996).

conclude (according to Bennett) that there is no morally significant difference between Case *A* and Case *D*[27].

However, as Frances Howard-Snyder (2011) points out, these conclusions are 'surprising, even shocking' because: (a) there are a number of examples that can be given, and that have been given, that demonstrate that there is a morally significant difference between doing harm and allowing harm (c.f. Woollard's (2012: 461) comparison of the Mountain Rescue Cases and the Wicked Uncle Cases); (b) we tend to (at least intuitively) think that the method by which an upshot is brought about has (at least some) bearing on the moral significance of that action (c.f. Quinn's (1989: 295–296) discussion of Bennett's (1995: 97) case of Henry); and (c) we are often drawn to the intuition that action is a positive fact and immobility is a negative fact. A number of authors—such as Frances Howard-Snyder (2011), Daniel Dinello (1971), Don Locke (1982), and Warren Quinn (1989)—have provided counter-examples and have cast a shadow over Bennett's account and his conclusions. One of the most widely discussed issues is what Bennett (1993; 1995: 96–100) calls the 'immobility objection'. Arguments in relation to the immobility objection differ in approach, and can be roughly divided into a linguistic approach and a moral approach. The linguistic approach can be characterised by appealing to the intuition that, in contrast to Bennett's account, an agent's immobility does not necessitate that an agent was positively relevant to the upshot. An example of the linguistic approach to the immobility objection is presented by Matthew H. Kramer (2014: 80–81):

---

[27] It is worth noting that Bennett (1981: 69) also states that '[i]f someone is prima facie to blame for conduct which had a disastrous consequence, the blame could not conceivably be lessened just by the fact that most of his alternative ways of behaving would have had the same consequence'.

Hiram I

'Suppose that, if Hiram remains completely motionless in the sealed room where he is standing, a fine metallic dust in the air will settle upon the floor. Some of the dust will fall onto a tiny electronic device and will close a circuit, triggering an explosion'.

Hiram II

'By contrast, if Hiram moves his body […] he will prevent the fine dust from settling and will thereby avert an explosion'[28].

In both of these situations, remaining immobile is the only behaviour available to Hiram that will result in the explosion. In Hiram I, if Hiram remains immobile then the upshot of the explosion would occur; on Bennett's account (1995: 98), Hiram's behaviour would be considered an action (or a case of doing). In Hiram II, if Hiram moved in any of the number of ways available to him then the upshot would not have occurred, and each of the possible ways of mobilising himself would, on Bennett's account, count as an omission (or allowing). The Hiram Cases therefore provide an example in which immobility is an action or a case of doing (as in Hiram I) and in which mobility is an omission or a case of allowing (as in Hiram II). However, as Kramer (2014: 81) notes, some of Bennett's critics think that 'we should be troubled' by this result. They claim that this is because, in Hiram I, we intuitively want to claim that if Hiram remains immobile then he allows the dust to settle, and therefore omits from behaving in a way that would prevent the explosion; and in Hiram II, we intuitively want to claim that if Hiram moves then the explosion has been averted not

[28] Kramer's (2014) examples are similar to examples given by other authors, particularly Bennett (1981: 66; 1995: 97) and Quinn (1989: 295), discussed in this section.

by his omission but by his action of not staying still. However, Bennett (1993: 83–85) would admit that if Hiram does remain immobile then he allows the dust to fall, but that these sorts of counter-examples serve to demonstrate 'how the detailed meaning of the word 'allow' is a poor guide in our present problem area' (Bennett 1993: 83). Bennett does not aim to provide an account of doing harm and allowing harm (or for his purposes, and more specifically, an account of making and allowing) that explains and neatly fits each and every possible case and situation. This is because the intuitions that are usually drawn on to underpin one's convictions vary according to the particular case and are formed by a variety of factors that do not transpose on to a single distinction. The problem, in Bennett's (1993: 84) words, is that 'people are guided by a clean, deep concept, but only imperfectly, because they sometimes drift away from it and use the terminology […] in ways that have no solid conceptual support'. Bennett thinks that the critics who provide this sort of counter-example are ultimately led 'astray' (Bennett 1993: 83) by the common use of the term 'allow', and this only compounds the intuition that immobility is joined to allowing and mobility or action to doing[29].

The moral approach can be characterised by appealing to some moral intuition that Bennett's account and his distinction between positive and negative relevance does not adequately capture the morally significant distinction demanded by common-sense morality. Quinn's (1989: 295–296) response to Bennett's (1995: 97) example of the case of

---

[29] Here it is worth noting that Howard-Snyder (2011) presents a compelling argument for why 'immobility is not necessarily incompatible with positive relevance to an upshot'. Although this is an interesting development to the debate, its discussion is not directly relevant to the purpose of this chapter. For more information, see Howard-Snyder's (2011) 'mild earthquake' example and the ensuing debate (involving Bennett's (1995) 'assassins' example in which Bennett's response (in correspondence with Howard-Snyder) is discussed).

Henry is at the forefront of arguments that take the moral approach to the immobility objection. Bennett (1995: 96) states that his account 'implies that *He moves* is [a] negative [proposition] and that *He does not move* is [a] positive [proposition]'; 'Agent did not move' is a positive proposition about Agent's behaviour, and 'Agent moved' is a negative proposition about Agent's behaviour. Bennett (1981: 66; 1995: 97) asks us to consider the following case of Henry, which he says is 'a kind of example that has repeatedly been brought against my analysis'; this is the case that Quinn (1989: 295) draws on to support his claim that Bennett's account 'gets certain cases intuitively wrong':

> *Henry I*
>
> 'Henry is in a sealed room where there is fine metallic dust suspended in the air. If he keeps stock still for two minutes, some dust will settle in such a way as to close a tiny electrical circuit which will lead to some notable upshot U [e.g. an explosion that causes the death of Bill]. Thus, any movement from Henry, and U will not obtain; perfect immobility, and we shall get U'[30].

Bennett's analysis of Henry I is that Henry's immobility makes U obtain, e.g. the death of Bill (or, for consistency with this chapter's discussion of doing harm and allowing harm, Henry does harm to Bill). By remaining immobile Henry is positively relevant to the harm that come to Bill, and, since Bennett states that immobility counts as causing harm, Henry harms Bill (Bennett, 1995: 98). Quinn (1989: 296) continues:

---

[30] This is taken from Bennett (1995: 97), although Bennett first introduced this example (in a slightly different format) in *The Tanner Lectures on Human Values* (Bennett, 1981: 66).

Henry II

> ‘But suppose Henry could save five only by staying where he is—suppose he is holding a net into which five are falling’.

Quinn (1989: 296) states that, in Henry II, Henry ‘might then properly refuse to move even though it means not saving Bill […] [since] his agency in Bill’s death would in that case seem negative’. Quinn (1989) argues that Henry II provides a case in which Henry can remain immobile even though this will result in the death of Bill; according to Quinn, Bennett misses a case in which Henry’s immobility is negatively relevant to Bill’s death. But this is plainly wrong. Henry—or more specifically Henry’s immobility—is still positively relevant to Bill’s death. Here, Quinn seems to confuse Bennett’s account of positive and negative relevance, importantly both morally neutral concepts, with his own account of positive and negative agency, both of which are morally-laden terms. Quinn in fact ends up making a moral judgement about Henry’s responsibility for an upshot. By using the term ‘negative’ to represent ‘negative agency’ instead of ‘negative relevance’, and by in effect giving Henry a reason for remaining immobile, Quinn seems to be attempting to establish that Henry’s responsibility is in some way lessened or diminished. Quinn has clearly lost focus of his target—making a moral judgement of Henry’s involvement is an entirely different task to determining whether Henry’s behaviour is positively or negatively relevant to Bill’s death. Quinn therefore fails to establish that Bennett ‘gets certain cases intuitively wrong’.

Quinn (1989: 296) continues this mistake when arguing that ‘Bennett also misses the opposite case’:

Henry III

> ‘Suppose the device will go off only if Henry makes some move or other’. (And, presumably, by moving Henry lets the dust fall and set in the position that his body was sheltering.)

Quinn is correct in identifying that, in Henry III, Bennett would say that if Henry moves then he is negatively relevant to Bill’s death. However, Quinn argues that Henry is positively relevant to Bill’s death if he detonates the device by moving. It seems that Quinn wants to argue that Henry should not save the five if this results in an explosion that kills Bill; in doing so, Henry’s action would be positively relevant to the death of Bill. But Quinn has, once again, made a mistake here. It seems that, like in Henry II, Quinn is either conflating or confusing positive agency and positive relevance to an upshot. We are therefore left with an account by Quinn that, at best, misunderstands Bennett’s account or conflates his positive/negative agency with Bennett’s positive/negative relevance, or, at worst, tries (and fails) to smuggle-in the assumption that Bennett’s account has moral significance.

1.2.7. DONAGAN AND ‘THE COURSE OF NATURE’: QUIESCENCE, INTERVENTION, AND ABSTENTION

Alan Donagan (1977) provides a different account of an agent’s relevance—be it positive or negative—to an upshot, in which there is a consideration of what upshot would have occurred in ‘the course of nature’ had that agent been quiescent, or would have occurred had that agent ‘abstained from intervening in the course of nature’. Donagan’s terminology can be best understood through Bennett’s descriptions and examples of the three ways in which an agent can relate to a state of affairs (Bennett, 1993: 86–87): an agent can be

quiescent in the course of nature, and in such cases agency or intentional action is not involved (e.g. an agent's head moved because a brick hit it); an agent can *intervene* in the course of nature, and in such cases an agent's actions are relevant to the upshot (e.g. an agent's head moved because he nodded); or an agent can *abstain* from intervening in the course of nature, and in such cases an agent's actions are not relevant to the upshot (e.g. an agent's head moved because, feeling the onset of a suppressible sneeze, he decided not to supress the sneeze and to let the sneeze take hold). The difference between the terminology is subtle—especially between quiescence and abstention—but important. And although Donagan does not directly talk of 'quiescence', his account—as we shall see—lends itself to identifying 'quiescence' as distinct from a discussion of 'intervention' and 'abstention'. That said, one might question the extent to which quiescence and abstention are terminologically distinct from each other; one might, for example, question the difference between moving one's head because it was moved by a heavy brick (quiescence) and moving one's head because of one's refusal to supress a sneeze (abstention). The distinction lies in that Donagan's account takes into consideration what *would have happened* (in 'the course of nature') had an agent not acted. Donagan (1977: 42) claims that an action 'is a deed done in a particular situation or set of circumstances […] [consisting] partly of matters external to the agent […] and partly of his own bodily and mental states', and it is the agent's very ability to affect the course of nature by acting that grants the agent the ability to either intervene in the course of nature or abstain from intervening. By being quiescent, an agent simply follows the course of nature; by intervening, an agent does not allow the course of nature to follow; and by abstaining from intervening, an agent's actions allow the course of nature to follow. In other words, if an agent's actions affect the course of nature then this constitutes an intervention, and if an agent's actions do not affect the course of nature then this constitutes an abstention (and, to finish the triad of distinctions presented by Bennett

(1993: 86–87), if an agent does not act and thus the course of nature follows then this constitutes quiescence). So, to settle the mind of anyone who questions whether there is a difference between quiescence and abstention: on Donagan's account, or more specifically on Bennett's portrayal of Donagan's account, the agent whose head moved due to being hit by a brick is quiescent only in so far as the movement of his head was a part of and flowed with the course of nature, and the agent whose head moved due to refraining from supressing a sneeze abstained from intervening only in so far as the movement of his head could have been prevented and thus chose to intervene in the course of nature[31].

The intervention/abstention distinction[32] applies to the distinction between doing harm and allowing harm in the following way. In Donagan's (1977: 43, emphasis added) words, when an agent intervenes 'he can be described as *causing* whatever would not have occurred had

---

[31] There is much to be said on Donagan's account, especially on the distinction between quiescence and abstention. For starters, Donagan's notion of quiescence, intervention, and abstention depends on Donagan identifying and utilising a philosophically sound understanding of 'the course of nature'—the very idea of which is deeply problematic (i.e. are human agents part of the course of nature, *viz.* isn't my head moving a part of the course of nature in as much as a cat's head movement is, and if not, why is it not?). There are other worries too: why talk of or worry about the course of nature at all? If my head moves and causes harm to another, does it make a difference, morally speaking, whether my head moves by quiescence, intervention, or abstention? If my head knocks someone's tooth out, does talking about what would have happened in the course of nature vindicate the quiescent agent ("Look, my head just moved, it was out of my control. I'm sorry!")? However, as interesting as these questions are, and as pertinent as they are to understanding whether Donagan's account is philosophically coherent, they are not pertinent to this chapter's purpose of providing an overview of the current picture of moral sequencing, and so will not be discussed.

[32] Quiescence will not be discussed any further for two reasons. Firstly, Donagan (1977) does not explicitly discuss quiescence; Bennett draws a distinction between quiescence and intervention and abstention to introduce how 'agency can relate to a state of affairs in any of [these] three ways', which is useful for establishing Donagan's account. Secondly, the distinction between doing harm and allowing harm relies on establishing that an agent plays an active role in a sequence with a harmful upshot (*viz.* that an agent is in some way relevant to a harmful upshot), and, since a quiescent agent is essentially a passive agent in 'the course of nature', quiescence adds little to the current debate.

he abstained', and when an agent abstains he is '*allowing* to happen whatever would not have happened had he intervened'. Donagan's account can therefore be applied to Bennett's discussion of an agent's actions being positively/negatively relevant to an upshot in the following ways. When ascertaining whether an agent is positively or negatively relevant to an upshot, one must determine what the upshot would have been had the agent 'abstained from intervening in the course of nature' (Donagan, 1977). If upshot *U* obtained because agent *A* abstained from intervening in the course of nature, then *A* allowed *U* (*A* is passive in the occurrence of *U*) and *A* is negatively relevant to *U*. If an upshot *U* obtained because agent *A* did not abstain from intervening (*viz. A* intervened in the course of nature), then *A* caused *U* (*A* is active in the occurrence of *U*), then *A* is positively relevant to *U*.

However, as some authors have commented, there are some serious problems with Donagan's account. Norcross (1994: 13–14) mentions how, if agent *A*'s action *P* caused harmful upshot *U* and *A* could have acted in a way *Q* (or simply ¬*P*) that would have ensured that *U* did not occur, then, because *U* would have occurred regardless of whether *A* performed *P* or *Q*, whether *A* performs *P* or *Q* seems to be irrelevant; *U* occurs regardless. This echoes the concerns I outlined with Bennett's account, where intervening and abstaining from intervening seem to be morally equivalent, since we have a case in which, regardless of whether an agent intervenes or abstains from intervening in the course of nature, the same upshot occurs. This ultimately casts a shadow over the enterprise of portraying the distinction between doing harm and allowing harm as a distinction between an agent being active or passive in the occurrence of harm.

Howard-Snyder (2011) has a different concern, and asks us to consider a man who, previously asleep on the ground, wakes to see a boulder rolling towards him. Able to easily

reposition himself to avoid the boulder, he has two options: move, but allow the boulder to continue on its path and kill several children; or remain where he is and prevent the boulder from reaching the children. He stays put and sustains serious injury. The boulder comes to a halt. The problem, in Howard-Snyder's (2011, emphasis added) words, is that 'Donagan's account […] seems to imply that he merely *allows* the rock to stop, since, had he remained asleep, the rock would have struck and been stopped by his body'. Donagan's account is therefore problematic for two reasons. We might want to say that the man *caused* the boulder to stop; the man's quiescence is not an issue that one intuitively brings to bear in such a discussion, for doing so relies on claiming that what is morally significant is what would have happened in the course of nature had the man not acted.

1.2.8. MCMAHAN AND BARRIERS

In *Killing, Letting Die, and Withdrawing Aid* (1993), Jeff McMahan discusses the importance of barriers in understanding the distinction between doing harm and allowing harm. McMahan's account is case-centred rather than concept-centred; in other words, McMahan does not focus on assessing the concepts of doing harm and allowing harm themselves, but rather focuses on particular cases involving doing harm and allowing harm. He approaches the distinction between doing harm and allowing harm in this way since, he argues, 'the empirical criteria [*viz.* the cases he discusses] determine a way of applying the concepts that we recognize as having moral significance' (McMahan, 1993: 250). In this section, my aim is not to provide an argument in support of or against McMahan's account or conclusions; rather, by discussing McMahan's account, I will demonstrate the importance of accounting for barriers in the doing/allowing distinction debate. This will prove to be

important for my later discussion of moral sequencing in chapter 2, which will feature barriers similar to those discussed by McMahan.

McMahan's overarching claim is that some, but not all, cases of removing a barrier to harm should be considered to be a case of doing harm; in some cases, removing a barrier to harm is a case of allowing harm. Assessing whether the removal of a barrier counts as a case of doing harm or allowing harm relies on establishing whether (a) the agent installed the barrier (McMahan, 1993: 255), (b) the barrier is 'self-sustaining'[33] (McMahan, 1993: 256), and (c) the barrier is 'operative' or 'as-yet inoperative'[34] (McMahan, 1993: 261). In McMahan's words:

> 'if a person requires or is dependent for survival on further aid from or protection by an agent, and if the person dies because the agent fails to provide further aid or withdraws his own aid either while it is in progress or before it becomes operative, and if the agent is not causally responsible for the person's need for aid or protection, then the agent lets the victim die' (McMahan, 1993: 261).

McMahan employs a number of cases to support his account and to underpin his claim that if an agent installs a barrier and if that barrier is as-yet inoperative or is not self-sustaining, then, if that agent removes that barrier, the agent has allowed harm to occur. Otherwise the

---

[33] A barrier can be said to be 'self-sustaining' if and only if the persistence of the barrier is not reliant on an agent's action to sustain the presence of that barrier.
[34] A barrier can be said to be 'operative' if and only if the barrier itself is preventing harm from occurring.

agent has done harm. Consider four of McMahan's cases, two of which, on McMahan's account, are cases of doing harm and two cases of allowing harm:

Respirator

'A person is stricken with an ailment that would normally be fatal but is given mechanical life-support to sustain him until the condition can be cured. While the patient is on a respirator, his enemy surreptitiously enters the hospital and turns the machine off. The patient dies' (McMahan, 1993: 254).

The Pipe Sealer

'An earthquake cracks a pipe at a factory, releasing poisonous chemicals into the water supply. Before a dangerous amount is released, a worker seals the pipe. But a year later he returns and removes the seal. As a result, numerous people die from drinking contaminated water' (McMahan, 1993: 256).

The Dutch Boy

> ‘A little Dutch boy, seeing that the dike is beginning to crack, valiantly sticks his finger in the crack to prevent the dike from breaking and flooding the town. He waits patiently but after many hours no one has come along who can help. Eventually succumbing to boredom and hunger, the boy withdraws his finger and leaves. Within minutes the dike bursts and a flood engulfs the town, killing many’ (McMahan, 1993: 257).

The Impoverished Village

> ‘Having given one’s accountant full power of attorney, one learns that because of a misunderstanding he is preparing to sign away 10% of one’s income to be sent to the [impoverished] village. One phones to instruct him not to [send the money]’ (McMahan, 1993: 259)[35].

The first two are cases of doing harm. In Respirator, the barrier is the life-support machine, which is both self-sustaining (since its sustenance does not rely on the actions of any agent) and operative (since the switched-on life-support machine is preventing harm from coming to the person attached to it). By turning off the life-support machine, the ‘enemy’ can, on McMahan’s account, be said to have done harm to (killed) the person on life-support. In The Pipe Sealer, the barrier is the seal on the pipe, which is both self-sustaining (since its permanence is not reliant on the actions of any agent) and operative (since the seal is

---

[35] Jonathan Bennett (1981: 89) originally discusses this example. McMahan calls this case The Impoverished Village 3 (McMahan, 1993: 259) to differentiate it from two similar cases, The Impoverished Village and The Impoverished Village 2, which he discusses in the same paper.

preventing harm from coming to those who drink the water). By removing the seal, the worker can, on McMahan's account, be said to have done harm to (killed) those who drank the water. The final two are cases of allowing harm. In The Dutch Boy, the barrier is the boy's finger in the crack, which is operative (since his finger is preventing the flood) but is not self-sustaining (since the prevention of harm relies on the boy keeping his finger in the crack). By removing his finger, the boy can, on McMahan's account, be said to have allowed harm to come to those killed by the flood. In The Impoverished Village, the barrier is the person's money, which is self-sustaining (since the sum of money will inevitably help save the impoverished villagers and, if transferred, would not require any agent to sustain the barrier) but is as-yet inoperative (since the money has not yet been transferred and is therefore not yet preventing the death of the impoverished villagers). By preventing the transfer of the money, the person who called his accountant can, on McMahan's account, be said to have allowed harm to come to the impoverished villagers.

Although I find McMahan's account of self-sustaining and operative barriers appealing[36], it is not without problems. As Norcross (1994: 20) identifies, the terms 'self-sustaining', 'operative', and 'as-yet inoperative' are 'vague' matters, 'leading to uncertainty as to how to classify some cases'. Norcross' claim is that the vagueness of each term can lead to problems in some cases. For example, in Respirator, McMahan (1993: 266) asks us to agree with his 'intuitively right' conclusion that this is a case of doing harm for the reasons provided above. However, as McMahan (1993: 266) himself mentions, there is a 'lack of clarity about whether or not a life-support machine counts as a self-sustaining form of aid' precisely because life-support machines *do* require monitoring and maintenance. Although

---

[36] To understand why I find McMahan's construal of barriers attractive and to see how it is an important component in moral sequences, see §2.2.2.

Norcross (1994: 20) does not think this point would necessarily affect making a judgement in Respirator, it is a point worth noting because it does challenge the appropriateness of McMahan's terminology. After all, I do not know of any self-monitoring, self-maintaining life-support machines, and as such one could put forward an argument claiming that the barrier, the life-support machine, does require monitoring/maintenance and thus is not self-sustaining. This is indeed an attractive criticism, but not one that I will pursue.

An equally interesting criticism, again from Norcross (1994: 20–21), concerns whether the seal on the pipe in The Pipe Sealer is operative or is in fact as-yet inoperative. Consider the following case:

> *The Plumber*
>
> '[A] plumber installs a totally reliable maintenance-free water filter designed to remove a specific type of toxic particle from a household water supply. A year later, he removes the filter, and soon afterwards the residents die from poisoned water. [However it transpires that] [t]hroughout the year in which the filter is in place, there are no toxic particles in the water coming into the house, so the filter doesn't actually remove any. Soon after the plumber removes the filter, purely coincidentally, toxic particles show up for the first time' (Norcross, 1994: 20–21).

What started out to be a case similar to The Pipe Sealer ends up being a dissimilar case, with different implications. A question that arises from The Plumber is whether the filter should be considered to be operative or as-yet inoperative when the plumber removed it. The

answer could, on McMahan's account, change whether the plumber can be said to have done harm or merely allowed harm to occur. Because the filter had not actually caught any toxic particles and had therefore not acted as a barrier to harm when the plumber removed it, the filter could be considered to have been as-yet inoperative rather than operative. This change in the operativeness of the filter would, on McMahan's account, change our judgement of the plumber: if the plumber removed the filter *before* the toxic particles entered the water supply, then the filter was as-yet inoperative, and the plumber has only allowed harm to come to the residents. But if the plumber removed the filter *just after* the toxic particles entered the water supply, then the filter was operative, and the plumber has done harm to the residents. The Plumber case illustrates the difficulty in determining the operativeness of a barrier, and how one's judgement of whether harm has been done or allowed to occur is subject to knowing whether the barrier is *actually operative*. Perhaps we should say that an 'operative barrier' on McMahan's account is simply a 'barrier that is actually operative' and an 'as-yet inoperative barrier' is a 'barrier that is not actually operative'. Might this alleviate the issue at hand? It would, if only there were a fool-proof method for determining whether the barrier was, at the time it was removed, actually operative or not. Although this could be accomplished in many cases—the plumber or a third-party could have tested for toxic particles before the filter was installed, for instance—it would require a continual assessment and knowledge of whether the barrier was at the time preventing harm from occurring. It is entirely conceivable, for instance, that toxic particles could enter and exit the water supply at any time, and this would continually change whether the barrier was operative or as-yet inoperative. It is therefore not philosophically or conceptually rigorous enough to link our judgement of the plumber to the operativeness of the barrier at the time it was installed, and can pose problems for judging whether the plumber has harmed or allowed harm depending on when the barrier was removed. For if the plumber removed the

filter at time t^1 when the barrier was actually operative and was preventing toxic particles from poisoning the residents, then, on McMahan's account, we would say that the plumber had done harm to the residents. But if the plumber had waited just a few more minutes and removed the filter at t^2 when the barrier was not actually operative and was not preventing toxic particles from poisoning the residents, then, on McMahan's account, we would say that the plumber had allowed harm to come to the residents. It therefore seems that McMahan's account forces us to make a judgement on whether harm has been done or allowed to occur based on a rigid system that cannot reliably track the operativeness of the barrier and arguably does not track common-sense morality; the few minutes between t^1 and t^2 that ultimately determine whether the plumber has done harm or allowed harm to occur would not likely, in our ordinary thinking, deter us from saying that the plumber at t^2 had killed the residents.

What if we say, as Norcross (1994: 21) mentions, that the filter was *operative all along*, and so, even though it didn't actually catch any toxic particles, it would have filtered them had the water been contaminated? This 'operative all along' clause *assumes* that a barrier was actually operative at all times, and would therefore prevent a barrier from ever being considered as-yet inoperative. However, understanding barriers in this way would require a complete overhaul of McMahan's account, in which different mechanisms would have to be in place to determine whether the removal of a barrier should be considered to be a case of doing harm or allowing harm. Perhaps a more pressing concern for McMahan would be that this 'operative all along' way of understanding barriers would have implications for other cases. If The Plumber case is compared to other cases discussed by McMahan, then the status of the barrier could change. Consider the following case, presented by McMahan (1993: 262):

Burning Building 2

> '[After placing a net under a man who has jumped from a burning building] the firefighter immediately notices that two other persons have jumped from a window several yards away. He therefore repositions the net so that it catches the two. The first jumper then hits the ground and dies'.

In this case, the barrier (the net) is self-sustaining (since once it has been positioned its persistence does not depend on the actions of any agent) and is as-yet inoperative (since jumper has not yet hit the net and so is not yet preventing harm from occurring). On McMahan's account, by repositioning the net, the firefighter allows the first jumper to die. But if we see the barrier as being operative all along, then the firefighter could be said to have killed the first jumper; the net was placed under the first jumper, and would have caught and saved him, and by removing this barrier that was operative all along, the firefighter caused harm to the first jumper.

Other authors have similar worries, most notably Kai Draper (2005) and Fiona Woollard (2015). Draper looks at The Impoverished Village case to illustrate the difficulty in determining the operativeness of aid; is the aid operative when the recipient 'receives the cheque, cashes it, purchases food, or begins to eat?' (Draper, 2005: 209). The lack of clarity over operativeness and how it should be understood and applied can evidently lead to problems for the cases that McMahan relies on for his account. Amongst other things, Woollard takes issue with the role of the agent in establishing the barrier: 'the agent […] did not need to do anything to set the barrier up' (Woollard, 2015: 68) since it was the accountant who is responsible for transferring the money to the villagers. It is therefore

strange to see McMahan hold that it is the agent (here, the person to whom the money belongs) who establishes the barrier, since it is ultimately the accountant who is responsible for sending aid. Our hand is therefore forced, as Woollard (2015: 68–69) mentions, to accept the fact that the agent provides the barrier since it is his money that is providing aid, and without the money the accountant's role would be insignificant. Although this is a fair point, this case points towards having to account for dual-agency or perhaps even multi-agency in sequences like The Impoverished Village, where the installation of a barrier is a multi-faceted concept relying on the actions of more than one agent. Providing an account that accepts this approach would require overhauling McMahan's account and will not be discussed further, but what is important is that even the role of an agent—or rather more than one agent—must withstand a level of scrutiny that McMahan's current account cannot currently withstand.

1.2.9. WOOLLARD AND THE LIMITATIONS AND INADEQUACIES OF CURRENT ACCOUNTS

So far, this chapter has presented an overview of some of the most influential accounts of the distinction between doing harm and allowing harm broadly construed. These accounts make up the essential history and background of moral sequencing, as it is within these accounts that the foundations on which moral sequencing is built can be glimpsed. However, what I consider to be the most interesting account, and the account that brings the most cards to the table, is yet to be properly attended to, namely Fiona Woollard's account of the doing/allowing distinction.

In *Doing and Allowing Harm*, Woollard (2015) draws her influences from some of the authors discussed above, and draws from and builds on her earlier work (Woollard, 2008; 2012; 2013). However, Woollard's drive is to pick up where the other authors depart, and to provide answers to questions that some authors either do not answer or do not answer adequately. Woollard (2015: 20) argues that: Foot's analysis is 'incomplete' since it leaves 'vital concepts unexplained and thus undefended'; McMahan leaves 'the idea of actively removing a barrier itself unexplored'; Quinn's account is 'not adequately defended, for he did not explain how the authority of ownership connects to the distinction between doing and allowing itself'; and other authors 'have provided important insights, but […] none have yet met Bennett's challenge'. Woollard's (2015) monograph essentially bridges the gap between these omissions. In Part I (Woollard, 2015: 3–94), Woollard presents an account detailing how the doing/allowing distinction should be understood. Woollard adds to the debate by suggesting that 'harm will count as something the agent has done if it is the upshot of a series of substantial facts leading from the agent's behaviour, i.e. if there is an unbroken series leading from the agent's behaviour to the upshot' (Woollard, 2015: 34) and that 'an agent merely allows harm if and only if the agent's behaviour is relevant to a harm but does not count as part of the sequence leading to that harm, because the chain of facts linking the agent's behaviour to the harm is broken by some non-substantial fact' (Woollard, 2015: 208). Woollard spills much ink on defining and explaining what constitutes a 'substantial fact', and she builds her argument on Bennett's (1995) account of positive facts to achieve this (see §1.2.6.). To summarise Woollard's views, positive facts can be divided into those that are 'specificity positive' and those that are 'scalar positive' (Woollard, 2015: 55). The former, Woollard (2015: 54–55) argues, is the kind of positive fact discussed by Bennett (1995); a specificity positive fact is informative and 'tells us that something specific was the case, pinning us down to a small set of alternatives' (Woollard, 2015: 54). The latter,

Woollard (2015: 55) argues, 'tells us that we are above a certain point on a natural scale'—that "The temperature exceeds 28 degrees Celsius" is positive, that "The temperature is below 28 degrees Celsius" is negative (Woollard, 2015: 54). However, negative facts can sometimes be substantial facts too. Woollard (2015: 57–58) claims that a negative fact is substantial if that fact is not considered a 'normal presupposition' that is usually considered to be common knowledge. To clarify this position Woollard (2015: 58) uses the example of how we assume that the air around us contains oxygen: if agent *A* dies from asphyxiation because agent *B* 'use[s] a machine that sucks the oxygen from the room', then because the fact that there is no oxygen in the room goes against our normal presupposition that there is oxygen in the room, that there is no oxygen in the room (a negative fact) is a substantial fact and so *B* can be said to have done harm to *A*. In addition, Woollard builds on McMahan's (1993) account of barriers in order to show how typically non-substantial facts are sometimes substantial to a harm and the stronger claim that '[w]e need the idea of normal presuppositions to classify such cases correctly' (Woollard, 2015: 59), i.e. the 'car theft case' (Woollard, 2015: 58–59).

However, since Woollard's defence of and conclusions on the distinction between doing harm and allowing harm are not directly relevant to this chapter's purpose of providing the background to moral sequencing, I will not assess or critique her account or argument here. Instead, the rest of this chapter will attend to Woollard's (2008) earlier paper on *Doing and Allowing, Threats and Sequences* in which she presents an innovative way of understanding sequences using visually rich diagrams that provide a schematic way of representing the doing/allowing distinction, and which provide the foundations on which I will build my account of moral sequencing.

1.2.10. THE EMERGENCE OF MORAL SEQUENCING PROPERLY UNDERSTOOD

Building on Philippa Foot's usage (1994)[37], Woollard (2008; 2015) uses the terms 'initiating', 'sustaining', 'enabling', and 'forbearing to prevent' to provide a system of sequencing that can be used to gauge the difference between doing harm and allowing harm. Person *A* 'initiates' a sequence if *A* acts in a way that brings about *x,* where *x* would not have occurred if it were not for *A* bringing about *x*. 'Forbearing to prevent', on the other hand, requires us to think of 'a sequence […] as already in train, and something the agent could do to intervene. (The agent must be able to intervene but does not do so.)' (Foot, 1994: 273). So, whilst 'initiating' relies on *A* bringing about the occurrence of something which would otherwise not have happened, 'forbearing to prevent' involves *A* being able to intervene in a sequence initiated by an 'unknown other', but fails to do so. The former can be seen as equating to killing (e.g. *A* drives his car into *B*) and the latter allowing to die (e.g. *A* had the chance to push *B* out the way of a moving car but chose not to). However, one still needs to distinguish between 'cases in which [...] a sequence would have stopped [*sustaining*], and cases in which [...] the sequence would have been stopped [*enabling*]' (Woollard, 2008: 266). Therefore, *A* can be said to be 'sustaining' a sequence if *A* acts in a way that facilitates *y*. In process *R*, in which *R* would have halted before the occurrence of *y, A* does *x* to prolong *R* in such a way as to facilitate *y* which, without *x,* would not have occurred (e.g. *C*'s car would have come to a halt before killing *B*, but *A* decided to drive his car into the rear end of *C*'s car, causing *B* to be hit). Additionally, *A* can be said to be 'enabling' a sequence to continue if *A*'s inaction could have prevented *x* from happening to

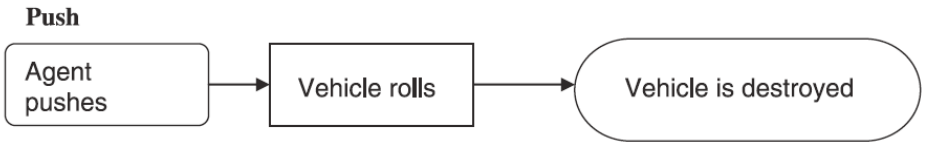
---

[37] It is important to note that these terms are not coined by Foot and are used in the work of some of the authors discussed earlier in this chapter. However, Woollard employs these terms in a way similar to and consistent with Foot's usage.

B (e.g. *A*'s car was parked in such a way that *C*'s car would have hit *A*'s car, preventing *B* from sustaining harm. However *A* decided to move his car, affording *C* the path required to hit *B*).

'Initiating' therefore relies on one bringing about a state of affairs that did not exist prior to the agent's actions, whilst 'enabling', 'sustaining', and 'forbearing to prevent' involve there already being a pre-existing threat that one permits to continue[38]. Woollard describes the relationship between 'initiating'/'sustaining', and 'enabling'/'forbearing to prevent' by reference to the former group's *dependency* on an agent, and the latter group's *independency* of an agent. In sequences involving 'initiating' or 'sustaining', the sequence 'is in some relevant sense *dependant* upon the agent', whilst in the case of 'enabling' and 'forbearing to prevent', the sequence seems to be 'in some relevant sense *independent* of the agent' (Woollard, 2008: 266). Consider the following diagrams, as given by Woollard (2008: 267–270):

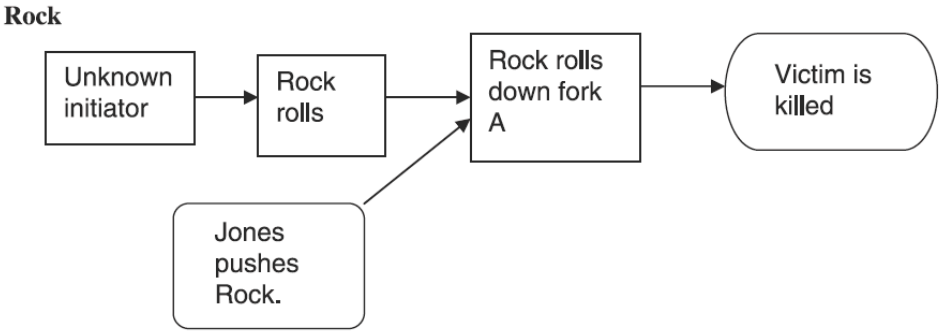
Figure 3: Woollard's Push sequence



[38] In order to draw a moral distinction between doing and allowing, Woollard groups 'initiating' and 'sustaining' as cases of doing, and groups 'enabling' and 'forbearing to prevent' as cases of allowing. However, for the purpose of understanding the mechanics of Woollard's sequences, one need only understand these terms in relation to their role in sequences.

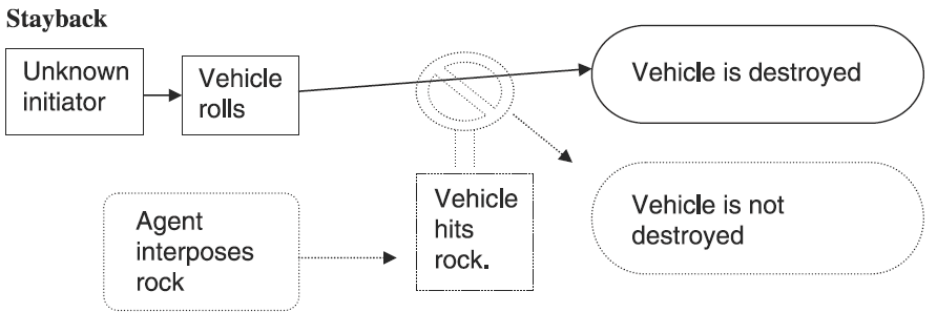
Here, Push clearly exemplifies a case of 'initiating', for an agent starts a sequence that results in the Vehicle being destroyed. The sequence is *dependent* upon Agent's actions, *viz.* the act of pushing, without which there would have been no sequence.

Figure 4: Woollard's Rock sequence



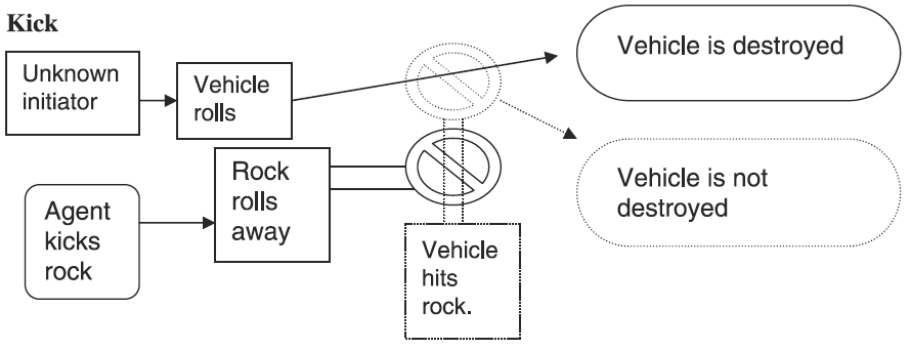
In Rock, 'some unknown initiator has set a large steel ball [*viz.* Rock] rolling and Jones gives it the extra momentum required to crush and kill Victim' (Woollard, 2008: 270). As such, Rock is a clear case of 'sustaining'. By pushing the rock, Jones 'move[d] forward a sequence that would otherwise have stopped' (Woollard, 2008: 270); without Jones' push, Victim would not have died. The sequence is therefore *dependent* on Jones' action of pushing, without which the rock would not have had the momentum to kill Victim.

Figure 5: Woollard's Stayback sequence



In Stayback, ‘Agent does not interpose the rock, so the vehicle does not hit the rock and the sequence continues’ (Woollard, 2008: 269), and for this reason Stayback illustrates a case of ‘forbearing to prevent’. The sequence was initiated by an unknown initiator and would have proceeded in such a way as to culminate in the vehicle being destroyed. Because the agent’s actions would have only prevented the vehicle from being destroyed, the initiation of the sequence does not depend on the agent’s actions, and so the sequence is *independent* of the agent (although the conclusion of the sequence is arguably determined by the agent’s action/inaction).

Figure 6: Woollard’s Kick sequence



In Kick, the ‘vehicle is rolling to a point where there is a rock that can bring it to a halt. Agent kicks away the rock, and the vehicle rolls to its destruction’ (Woollard, 2008: 268). Kick is a clear case of ‘enabling’, for in this sequence Agent removes the only obstacle preventing the unfolding of the sequence, or ‘train’, of events. Similar to Stayback, the Kick sequence is started by an unknown initiator, that is, it is initiated *independently* of Agent’s actions.

What Woollard wants us to take from her diagrams is that ‘when an agent initiates or sustains a sequence, his behaviour is part of that sequence’ in as much as the sequence-

outcome is dependent on that agent's actions; but when that agent 'merely allows or enables the sequence his behaviour is not part of that sequence', the sequence-outcome is independent of that agent's actions (Woollard, 2008: 270). But can Woollard's understanding of sequences be used to understand broader moral problems, other than the doing/allowing distinction?

1.3. CONCLUDING REMARKS

This chapter presented an overview of the literature on the distinction between doing harm and allowing harm in order to tease-out terminology and concepts that are relevant to and that will be employed in the novel system of moral sequencing that I will present and outline in the next chapter.

Each section introduced new terminology and concepts which jointly form the foundations for constructing a new system of moral sequencing that can be applied to general moral issues. In particular, the concepts of agency, sequencing, initiating, sustaining, enabling, forbearing to prevent, intervention, interposition, barriers, responsibility within a sequence, and negative rights all have a role to play in the following chapters.

Having grappled with the various approaches discussed throughout this chapter, the reader will hopefully be primed to be receptive to the moral sequencing I will present in chapter 2.

CHAPTER 2:

MORAL SEQUENCING

> "Creating a new theory is not like destroying an old barn and erecting a skyscraper in its place. It is rather like climbing a mountain, gaining new and wider views, discovering unexpected connections between our starting points and its rich environment. But the point from which we started out still exists and can be seen, although it appears smaller and forms a tiny part of our broad view gained by the mastery of the obstacles on our adventurous way up."
>
> — Albert Einstein and Leopold Infeld*, The Evolution of Physics* (1938: 159).

This chapter will argue for the philosophical value and pragmatic usefulness of employing what I call *moral sequencing*. This chapter seeks to (a) provide a new system of moral sequencing that builds on the literature, (b) use Fiona Woollard's sequences as models, (c) outline ways in which the problems that I identify with the current picture of moral sequencing, including Woollard's sequences, can be rectified, and (d) construct two sequence archetypes that can be, and in later chapters will be, employed and added to in order to properly account for justifiably intervening in a moral sequence and attributing responsibility. This will be achieved so that, in chapter 6, the archetypes can be revised and added to so that the personality of agents can be accounted for in light of the problem case discussed there.

2.1. WHAT IS A MORAL SEQUENCE?

A moral sequence is a linear process that depicts (both in prose and diagrammatically) an agent's actions in a discrete time-frame, starting with an agent initiating what I call a *non-pre-existing threat of harm* (NPET) and ending when either that NPET is fulfilled (and the harm that it threatened occurs) or another agent intervenes to prevent the NPET from being fulfilled (and prevents harm from occurring). In a moral sequence, an agent's actions are mapped-out, sequence-event by sequence-event—in real-time, in retrospect, or both—to predict what that agent might do (in real-time) with the view to intervening to prevent harm from occurring to another agent, and/or to ascertain what an agent has done (in retrospect) with the view to determining whether that agent (or agents) is (or are) responsible for the sequence-outcome. In both real-time and retrospective moral sequences, the agents under scrutiny (for whom intra-sequence and post-sequence evaluations will be made) are the agents that initiate a moral sequence, intervene to prevent harm, and forbear to prevent harm.

However, the way in which sequence-events are determined and mapped-out differ depending on the type of moral sequence. In the case of real-time moral sequences, sequence-events (*viz.* what an agent "might" do) are predicted based on a continual assessment of the possible courses of action and the way(s) in which they contribute to crossing what I will call *the threshold of harm* (see chapter 4). Sequence-events can be continually updated in real-time as the sequence progresses, and the proximity to the threshold of harm can be updated and recalculated with each new sequence-event. These predictions continue to help establish the probability of harm eventuating (see chapter 4). An agent (who I will later call a Deliberator, see §2.2.1.1.) has grounds for justifying intervening in the moral sequence in order to prevent the NPET from being fulfilled (rendering the Deliberator who I will later call an Intervener, see §2.2.1.1.) only if this

threshold of harm is passed; equally, the Deliberator might not intervene (and thus become who I will later call a Forbearer, see §2.2.1.1.), in which case the NPET is fulfilled (and harm occurs). In the case of retrospective moral sequences, sequence-events are determined solely by the actions performed between the initiation of a sequence and either the intervention that prevented harm from occurring or the occurrence of harm.

This new structure of moral sequencing will lend itself to being employed in general moral situations; by constructing a moral sequence, an agent (namely a Deliberator or who I will later call a Moral Assessor) is in a position to gauge whether an intervention to prevent harm is/was appropriate and, if it is/was, when it is/was most justifiable. Moral sequences are therefore philosophically valuable and pragmatically useful in a number of ways. Real-time moral sequences are valuable and useful for intra-sequence considerations, namely to:

(1) *Predict* whether harm will occur if there is no intervention (discussed in chapter 3);

(2) *Determine* when one can justifiably intervene (discussed in chapter 4); and

(3) *Attribute* any responsibility to agents for their actions (discussed in chapter 5).

Retrospective moral sequences are valuable and useful for post-sequence considerations, namely to:

(4) *Legitimise* intervening (discussed in chapter 4);

(5) *Accommodate* an agent's personality (discussed in chapter 6); and

(6) *Establish* whether an agent's responsibility for their actions can be diminished (discussed in chapter 7).

2.2. UTILISING THE LITERATURE AND CONSTRUCTING MORAL SEQUENCING

Although Woollard's sequences are, without doubt, useful for understanding the doing/allowing distinction, I will argue that they fall short of encompassing a full understanding of moral sequences. Woollard's sequencing provides often ambiguous definitions of sequence parameters (initiating, sustaining, enabling, and forbearing), and delivers a narrow understanding of interferences that cannot be used in broader moral sequences. This section will introduce the "missing ingredients" necessary to complete the picture of moral sequencing. My moral sequencing, which will be outlined over the course of this chapter, builds on the terminology and concepts found in the doing/allowing distinction literature (presented in chapter 1) and, more specifically, in Woollard's sequencing by allowing sequences to be assessed not only in retrospect but also in real-time, as the sequence progresses (chapter 3) and by accounting for interventions in a way that facilitates making a decision concerning a justifiable juncture to intervene in a moral sequence (chapter 4). This will help to build an understanding of an agent's responsibility for their actions (chapter 5). After, a problem case will be introduced to challenge the moral sequencing framework and to refine moral sequencing to account for a crucial component: personality (chapter 6).

2.2.1. SEQUENCE COMPONENTS

This section will, in §2.2.1.1., outline the types of agents that are relevant to moral sequencing to demarcate agency-related terminology that will enable us, in §2.2.2. and throughout the following chapters, to discuss moral sequencing—and indeed different types of moral sequences—consistently and in the most appropriate way. This section will also,

in §2.2.1.2., critically examine the terminology related to interference—namely intervention, interposition, and interruption—to avoid any ambiguities concerning how an agent might interfere in a moral sequence that permeate the literature and to provide a case for why the moral sequencing presented in this thesis will focus only on cases of intervention.

2.2.1.1. AGENCY

Agents can be divided into those that have intra-sequence relevance and those that have post-sequence relevance. We might also describe agents that are outside of a moral sequence, including those who are not aware of a moral sequence as it is happening, as extra-sequence agents. To elaborate, there are those intra-sequence agents who: initiate a moral sequence by initiating a NPET (*Initiator*); are under threat of harm in a moral sequence (*Victim*); and consider installing a barrier to prevent harm (*Deliberator*) and then either do (*Intervener*) or do not (*Forbearer*) install a barrier. There are those post-sequence agents who make post-sequence normative moral evaluations based on the actions of intra-sequence agents (*Moral Assessor*). There are also those agents who may partake in a moral sequence but who do not initiate the moral sequence or intervene or forbear to prevent harm. I call these agents *Bystanders* in a moral sequence. The ways in which these Bystanders are involved in a moral sequence can vary depending on the moral sequence itself and is, admittedly, a grey area. Although the term may connote that Bystanders voluntarily, knowingly, or otherwise willingly partake in the moral sequence, I do not think such a limitation should be imposed on our understanding of Bystanders in moral sequences; for example, Bystanders may unknowingly be involved in a moral sequence (an *Unknowing Bystander* may accidentally intrude into a moral sequence that forms part of the sequence

but is otherwise not an act of initiating or intervening) or a Bystander may involuntarily become involved (an *Innocent Bystander* may be coerced or forced to participate—e.g. by being used, without their consent, by an Intervener in an intervention to avert harm to Victim)[39]. And, finally, there are those extra-sequence agents who, for reasons including being ignorant of a moral sequence (during the occurrence or unfolding of a moral sequence) or are outside of the considerations of a moral sequence (during post-sequence events such as being a spectator in court), are not involved in the moral sequence and therefore fall outside the sequence parameters of moral sequencing.

A *Victim* of a moral sequence is an agent who is under threat of harm, or would have been under threat of harm had an effective intervention not taken place, as a direct result of an agent (whom we shall shortly call an *Initiator*) initiating a NPET and thus a moral sequence. Harm does not necessarily have to befall the Victim, nor does the full threat have to reach the Victim—another agent (who we shall later call an *Intervener*) might install a barrier to prevent this harm from occurring or whose barrier might reduce the severity of harm to the Victim but not prevent harm in its totality. In order for an agent to be considered a Victim, I propose that a Victim is simply an agent to whom the initiated NPET will directly cause harm. It is, of course, entirely plausible that there are multiple Victims for any given moral sequence (a single NPET might cause direct harm to a number of different agents).

---

[39] Bystanders by definition stand on the periphery of a moral sequence and are quasi-intra-sequence agents since they only become an intra-sequence agent under certain conditions, for example if an Intervener decided to use them without their consent in their intervention to prevent harm befalling Victim. For this reason, Bystanders will take a backseat until the discussion of intervention in chapter 4.

In moral sequences, there are certain types of agents that are objects of normative moral evaluation: those who initiate a moral sequence and those who consider intervening, do intervene, or do not intervene in a moral sequence to prevent harm. Different moral questions arise in relation to each type of agent.

An *Initiator* is an agent who initiates a NPET and thus a moral sequence by acting in a way that creates a threat of harm to another agent (*Victim*) that did not exist until the Initiator acted. It is therefore a question of responsibility that arises in relation to an Initiator's action(s). In other words, an Initiator is an object of normative moral evaluation in so far as she is responsible for initiating the NPET. The extent to which an Initiator is responsible for initiating a moral sequence, including questions relating to intentionally and unintentionally initiating a moral sequence, will be discussed in chapter 5.

A *Deliberator* is an agent who is spatio-temporally and physically available to intrude into a moral sequence to install a barrier to prevent harm befalling a Victim and is aware that there is a threat of harm to a Victim that will be fulfilled without their intervention[40]. The term "Deliberator" is a placeholder until that agent either: installs a barrier to harm (but where the success of this barrier preventing the NPET from resulting in harm to the Victim is not necessarily guaranteed) or does not install a barrier to harm. If the Intervener installs a barrier in a way that (successfully or unsuccessfully) prevents the threat of harm to the Victim from being fulfilled or if the barrier is only partially effective and reduces the harm

---

[40] Those who are spatio-temporally and physically available to install a barrier but who are ignorant of the moral sequence or the threat of harm to the victim are extra-sequence agents that we might call *Potential Interveners*. Whether these agents should be an object of normative moral evaluation (e.g. whether they are responsible for their ignorance) may be a component of moral sequencing, but it is an issue that will not be addressed in this thesis.

sustained by the Victim, then that agent can simply be said to be an *Intervener*. An Intervener is an object of normative moral evaluation since the intervener (a) intruded into a moral sequence and (b) attempted to prevent harm from befalling the Victim (regardless of whether the barrier was effective, partially effective, or not effective in preventing a Victim from being harmed). A moral evaluation is required to determine whether the Intervener was justified in intervening and, based on the legitimacy of the intervention, further moral evaluations might be made. For instance, if an intervention was legitimate, the Intervener might be praised. If the Deliberator does not install a barrier to harm (but was spatio-temporally and physically available to do so and was aware that the harm to the Victim could have potentially been either reduced or eradicated by installing a barrier), then that agent can be said to be a *Forbearer* (since they forbear to prevent harm from occurring). A Forbearer is an object of normative moral evaluation since the Forbearer could have sought to prevent, but did not prevent, harm befalling the Victim; in this case, a moral evaluation is required to determine whether the Forbearer is responsible for not intervening and whether the Forbearer is responsible for the harm that befell the Victim[41]. How a Deliberator decides to become an Intervener (*viz.* how one decides to intervene in a moral sequence) will be discussed in chapter 3; whether an Intervener can be said to have been justified in intervening in a moral sequence will be discussed in chapter 4; and the extent to

---

[41] Although I use the singular in all three types of agents (Deliberators, Interveners, and Forbearers), it is important to note that there may be multiple agents of those types in a moral sequence. For instance, if I am sitting in a beer garden and witness a kerfuffle, I, along with other patrons, may take notice and, as the incident unravels and voices become increasingly raised, consider whether to intervene. At this moment, myself and the other patrons are Deliberators. If the incident escalates to physical violence, I might decide to approach those involved and attempt to break-up the fight (install a barrier to prevent harm) by restraining an aggressor. By doing so, I would become an Intervener. If the other onlooking patrons do not act in a similar way, they can be said to be Forbearers.

which a Forbearer is responsible for forbearing to prevent harm befalling Victim will be discussed in chapter 5.

Although moral discussions pertaining to the actions of the Intervener would be moot if the Initiator has not initiated a moral sequence in which this Intervener was tasked with deciding if, whether, and when to intervene, it is important to note that normative moral questions and evaluations arising as a result of the actions of the Initiator, Intervener, and Forbearer can be divorced; whether the Initiator is responsible for initiating the moral sequence and whether the Intervener/Forbearer could have and should have intervened in a moral sequence are separate.

Moral Assessors are tasked with asking normative moral questions and making normative moral evaluations related to the actions of Initiators, Interveners, and Forbearers. By this I simply refer to those agents—including you, the reader of this thesis and someone who I assume is interested in issues of morality and correctly attributing responsibility, or a judge—who morally assess the actions of these agents in the moral sequences in which they initiate, intervene, or forbear in order to determine appropriate normative moral conclusions. This scrutiny will be a journey on which I will guide the reader throughout the course of this thesis. *Moral Assessors* are therefore relevant to, but importantly outside of, moral sequences—they are those agents (from laypeople to judges) who make post-sequence moral conclusions.

2.2.1.2. INTERFERENCE: A CASE FOR INTERVENTION

Woollard (2008; 2012; 2013; 2015) uses both the terms 'intervene' and 'interpose' when talking about sequencing, although these terms are never defined (we have to assume that their meaning and employment track our everyday use and understanding of the terms), and they appear to be used interchangeably. However, these terms can be employed in a way that reflects an important distinction between types of *interference* in a moral sequence. I will define an interference in a moral sequence as an event that intrudes into a moral sequence and that does not originate from the agent who initiated the sequence (or, to be more precise, the agent who initiates the relevant non-pre-existing threat of harm (NPET)) (see §2.2.2.1. for the definition of a NPET and chapter 4 for a discussion of the role of NPETs in moral sequences). In order to explain and motivate this thesis' focus on a type of interference, namely intervention, this section will define some relevant terminology and explain why the scope will be narrowed to intervention alone.

Interferences can be divided into two event categories, interventions and interpositions, depending on whether the event in question is agential (in the case of the former) or non-agential (in the case of the latter):

Intervention

An intervention is an *agential event* consisting of installing[42] a barrier to harm that intrudes into a moral sequence and which originates directly from the actions of an agent other than the agent who initiates the moral sequence; the agent intervening in a moral sequence knowingly imposes himself on and is motivated to affect the outcome of the moral sequence.

Interposition

An interposition is a *non-agential event* consisting of installing a barrier to harm that intrudes into a moral sequence and which originates from an environmental event that is not directly connected to the actions of an agent; through causes unrelated to the action(s) of any agent, an environmental event interposes in a moral sequence to disrupt the progression of agential events.

In other words, here I offer a distinction between two types of interferences, interventions and interpositions, where interventions are events that are directly caused by an agent and interpositions are events that are not directly caused by an agent. Under this terminology, an intervention would be, for example, where Deliberator saw a horse galloping towards Victim and, seeing that harm would befall Victim, decides to impose himself on the moral

---

[42] The concept of "installing" a barrier may bring with it connotations of "building a wall" or suchlike. However, I use this in a broader way and as a catch-all phrase to denote any barrier being put in place, whether knowingly (intervenes) or unknowingly (interrupts), by both agents and non-agents alike. Both an agent pushing an empty car in the path of a rolling boulder (agential) and the wind causing a tree to fall on the path of the rolling boulder (non-agential) are instances of installing a barrier.

sequence by making loud noises so as to scare and divert the horse; and an interposition would be, for example, where a minor earthquake makes the horse fall and break its leg short of reaching Victim.

Drawing this distinction is useful in as much as it allows us to distinguish between interferences that are, and interferences that are not, caused by an agent. An agent can interfere in a moral sequence by intervening in that sequence, but an agent cannot interpose a moral sequence, for the occurrence of an interposing event in a moral sequence is simply a by-product of the moral sequence taking place in the natural world. This is an important distinction to make since I wish to isolate those cases in which an Intervener knowingly imposes himself on, and is motivated to affect, the outcome of the moral sequence from those cases in which an Intervener is not—the former, I will argue, are more salient.

However, problems arise when assessing interferences that seem to be caused by natural events (c.f. Donagan's discussion of the course of nature in §1.2.7.), but whose cause could be attributed to the actions of an agent; the worry here is that *prima facie* interpositions could turn out to be interventions. To clarify this possible concern, consider the following case of interference in a moral sequence. An agent *A* is driving his car towards agent *B*, who, fraught with shock, is unable to move out of the way. A rock rolls down a hill adjacent to the road down which *A* is driving. The rock rolls onto the road and stops in-between *A*'s car and *B*. *A* crashes into the rock and *B* remains unharmed. The rock therefore interferes in the moral sequence. But is this an intervention or an interposition? This seems to depend on whether the rock was mobilised by an agent, or, more precisely, whether or not the rock's rolling can be directly connected to the actions of an agent. There are two possibilities: either the rock was pushed by an agent *C*, or the rock rolled due to natural causes that cannot be

directly attributed to the actions of an agent. In the first instance, if *C* pushed the rock, one might say that *C* intervened in the sequence; *C*'s action of pushing caused the rock to roll. In the second instance, authorship of the event is not attributed to an agent, and the object or force itself can be said to have interposed the sequence; a natural event, such as a minor earthquake, caused the rock to roll. However, the line between a purely natural interference, devoid of any agential causation, and an interference that is caused by the actions of an agent, is sometimes blurred. One might, for instance, argue that the minor earthquake that triggered the rock to roll can be attributed to agential action(s). One might say that nearby hydraulic fracking caused the earthquake, and so the rolling of the rock can be directly connected to the agent(s) who operated the fracking equipment. The same could be said for a number of other seemingly natural events, such as flooding. If a flood interfered in a moral sequence, instead of simply chalking this up to a natural event (an interposition), one might attempt to draw a link between human consumption of fossil fuels (causing rising sea levels as a result of climate change) and the occurrence of the flood. In such cases, we seem to be left with an obscure understanding of the role agency plays in interpositions, and we are left questioning the extent to which an interference that is *in some way* caused by an agent is an interposition[43].

To deal with this uncertainty, I offer the following clarificatory remarks. Even though such cases seem to cast a shadow over whether an interference with an agential component can be classed as an interposition, the problem actually stems from determining, or rather

---

[43] Determining whether an agent or agents can be said to have caused a natural event, such as a flood or an earthquake, can be extremely difficult, if not impossible, to ascertain. It is not, however, my task here to comment on or attempt to determine whether a natural event can ever be said to have been caused by agential influence(s). I mention this here only to illustrate a criticism that may be advanced concerning the terminological distinctions I discuss.

clarifying, what it means for an intervention to be 'directly connected' to the action(s) of an Intervener. This can be addressed by attributing salience to the Intervener knowingly intruding into the sequence and the Intervener's motivation for acting. If we understand an intervention as being 'directly connected' to the actions of an Intervener if, and only if, that Intervener knowingly intrudes into a moral sequence and that Intervener acts as a result of his motivation to affect the sequence-outcome, then the terminological uncertainty fades away. Understanding this *direct-connectedness requirement* in this way tracks the ordinary use of the term 'intervention', where the intervening agent is, or seems to be, deliberately interested in preventing a harmful sequence-outcome for the sequence in which they intervene. It also allows us to distance our understanding of this sort of agential interference from another agential interference. It is here that I will propose a third type of interference, *interruption*, necessary to understand a type of agential event in which the acting agent either unknowingly intrudes into a moral sequence or is not motivated to affect or alter the sequence-outcome.

Interruption

> An interruption is an *agential event* consisting of installing a barrier to harm that intrudes into a moral sequence, but where the direct-connectedness requirement is broken when either that agent unknowingly intrudes into the moral sequence (*the ignorant condition*) or that agent knowingly intrudes into the moral sequence but is not motivated to affect the outcome of the moral sequence (*the non-motivated condition*)—both of these conditions are not jointly necessary conditions but both are individually sufficient conditions.

So an agent who absent-mindedly kicks the rock, oblivious to the moral sequence involving *A* and *B*, fulfils the ignorant condition and cannot be said to have intervened in the sequence. Equally, that agent's kicking of the rock might satisfy the non-motivated condition if his action of kicking the rock was not motivated by affecting the sequence-outcome (i.e. preventing *B* from being hit by *A*'s car) but was instead motivated by some other factor(s) (i.e. to alleviate boredom or to see if he had the strength to move the rock). If an agent fulfils the ignorant condition and/or the non-motivated condition of an interruption, then that agent can only be said to have *interrupted* the sequence. Understanding the term 'interrupted' in this way is consistent with its ordinary use: one may interrupt my bath by walking into the bathroom, even though one did not know I was bathing; equally, one might interrupt my bath without the motivation to prevent me from enjoying my bath but instead because of a motivation to wash their hands in the sink (or both). This is also consistent with the discussion of agency in §2.2.1.1. in which an Intervener must first be a Deliberator; an agent that interrupts (we might call them an *Interrupter*) cannot have been a Deliberator due to the ignorant condition and the non-motivated condition.

Drawing a distinction between these three types of interference is important for moral sequencing as it helps to isolate the most significant type of interference for moral sequencing, namely intervention. Intervention is the most significant type of interference when it comes to moral sequencing, and this thesis and my moral sequence archetypes (presented in §2.3.) will discuss and feature intervention over interposition and interruption due to its salient direct-connectedness requirement being able to provide an answer to the prominent question "When is intervening in a moral sequence justified?" Since agents are often confronted with deciding if and when to intervene in a moral sequence to prevent a harmful sequence-outcome, interventions are most relevant and will provide the most

fruitful route of philosophical enquiry. The decision-making process involved in deciding if and when to intervene in a moral sequence will be discussed in chapter 3. It is also important to understand an intervention in the way described above since it is a type of sequence-event (discussed in §2.2.3.).

But how does or how can an agent intervene in a moral sequence? To answer this, I borrow terminology from the literature, particular from McMahan (1993) on 'barriers'. In short, if a Deliberator *installs a barrier to harm*, or more specifically installs a barrier to prevent the NPET that was initiated by Initiator from harming Victim, then that Deliberator becomes an Intervener (representing the fact that he has intervened). What counts as a barrier is broad; anything that an agent can put in place to prevent harm from occurring to Victim would constitute a barrier. The next section on sequence parameters will, during the course of its discussion, provide a more robust explanation of barriers, how they fit into moral sequencing, and what roles they play—in particular, the removal of a barrier (see the discussion of enabling in §2.2.2.3.) and the refusal to install a barrier (see the discussion of forbearing in §2.2.2.4.). Peripheral issues will also be attended to, such as a failed intervention (in the case of a fully ineffective barrier) or a sub-optimal intervention (in the case of a partially effective barrier), in order to provide a well-rounded account of the role of barriers. This will all directly lead on to a discussion of how a Deliberator decides to intervene (in chapter 3) and a discussion of when an Intervener is most justified in intervening (in chapter 4).

2.2.2. SEQUENCE PARAMETERS

This section will utilise the literature on the doing/allowing distinction discussed in chapter 1 to build on the key terminology that has been useful in those sequences and provide more robust parameters that better fit more general moral sequences.

Woollard argues that in sequences in which an agent 'initiates' or 'sustains' a sequence, that sequence 'is in some relevant sense *dependant* upon the agent' (Woollard, 2008: 266); conversely, in a sequence in which an agent 'enables' the sequence to continue or 'forbears to prevent' harm, that sequence is 'in some relevant sense *independent* of the agent' (Woollard, 2008: 266). However, moral sequencing does not employ this reasoning and this section will demonstrate why Woollard's claims are misguided.

Throughout this section the concept of a barrier will be discussed in relation to each of the sequence parameters in order to demonstrate how barriers form part of the conceptual makeup of each sequence parameter and in order to provide a holistic conceptual framework for barriers (including their installation or forbearance, *viz.* intervening and forbearing to prevent harm in a moral sequence).

2.2.2.1. INITIATING A MORAL SEQUENCE

When does a moral sequence begin? Or, to employ the terminology of moral sequencing, when is a moral sequence initiated? If one looks to the literature discussed in chapter 1 to provide an answer, then one would walk away empty-handed. Even Woollard, who I believe has presented the most developed system of moral sequencing to date, does not provide firm sequence parameters. To recap Woollard's position (using her terminology), 'initiating'

relies on one bringing about a potentially harmful state of affairs that did not exist prior to the agent's actions (whilst 'sustaining', 'enabling', and 'forbearing to prevent' involve there already being a pre-existing threat that one permits to continue). From the off-set, then, a sequence for Woollard can either be initiated by a known agent or by an "unknown initiator"; only in Push (see Figure 3 in §1.2.10.) does a known agent initiate the sequence; in Rock (see Figure 4 in §1.2.10.), Stayback (see Figure 5 in §1.2.10.), and Kick (see Figure 6 in §1.2.10.) the sequence is initiated by an unknown initiator.

Although I agree with Woollard that moral sequences broadly construed must account for both known and unknown initiators, I will limit the moral sequencing presented and discussed in this thesis to those moral sequences that are initiated by an (agential) known initiator. Much time would have to be spent identifying cases of agential and non-agential unknown initiators (whether a sequence was initiated by an agent or other non-agential causes, e.g. flood) in order to identify whether talk of attributing responsibility for a sequence-outcome to the initiator of a moral sequence is appropriate, and further ascertaining the identity of an agential unknown initiator of a moral sequence in order to attribute responsibility to the appropriate agent. Since the drive of this thesis is to provide an answer to when it is most justifiable to intervene in a moral sequence to prevent harm from occurring and whether responsibility can be attributed to those involved in a moral sequence (amongst other questions), limiting the discussion to moral sequences being initiated by an agential known initiator will, in future chapters, encourage a philosophically engaging discussion without worry of addressing and accounting for fringe cases (i.e. how do we identify an unknown initiator, does it make sense to attribute responsibility to an agential unknown initiator, and so on) that are not necessarily salient to moral sequencing.

Importantly, then, although in the strictest sense a moral sequence can be initiated by an agential or non-agential known or unknown initiator, for the reasons outlined above this thesis will limit its discussion to moral sequences that are initiated by a known agent (hereafter referred to simply as an agent—and, in the case of initiating, such initiating agents will be referred to as Initiators). A sequence can only be initiated when a threat of harm to another agent (a Victim) is introduced by an Initiator, and where this threat didn't exist prior; I call this a *non-pre-existing threat of harm* (NPET).

> *Non-Pre-Existing Threat of Harm*
>
> A NPET is a direct threat of harm to another agent (Victim) that originates as a result of the Initiator's actions and where this threat did not pose a direct threat to Victim until Initiator acted.

With this in mind, I offer the following definition and conceptual clarification of how a moral sequence is initiated.

> *Initiating a Moral Sequence*
>
> A moral sequence is initiated when an agent (Initiator) brings about a non-pre-existing threat of harm (NPET).

I therefore agree with Woollard that initiating relies on one bringing about a state of affairs that did not exist prior to the agent's actions and that initiating 'is in some relevant sense *dependant* upon the agent' (Woollard, 2008: 266)—it is dependent on an agent (an Initiator) bringing about a NPET. However, the concept of initiating needed to be refined to isolate how a moral sequence begins. What we are concerned with are those circumstances in which

an Initiator changes the *status quo* to bring about a threat to a Victim—doing so, bringing about a NPET, is what we are concerned with in moral sequencing, as are questions related to intervening to prevent harm, forbearing to prevent the harm, and who, if anyone, is responsible for the harm sustained by Victim (or the attempted harm that would have been sustained by Victim if an intervention occurred).

2.2.2.2. SUSTAINING A MORAL SEQUENCE

The notion of 'sustaining' a sequence is problematic. In Rock, which Woollard takes to be a case of sustaining, Jones pushes a rock that was originally rolled by an unknown initiator and which, without Jones' push, would have otherwise stopped. Because of Jones' push, the rock maintains momentum and kills Victim. For Woollard, Rock is a case of sustaining because Jones' action of pushing the rock sustains the sequence that leads to Victim's death. However, it is simply not the case that Jones' action can be said to have sustained the (moral) sequence; without Jones' push, the rock would have come to a natural stop and would not have killed Victim. The unknown initiator (whoever orchestrated moving the rock in the first place) therefore cannot be said to have brought about a NPET, since the rolling rock would have lost momentum and posed no threat to Victim. It was only because Jones pushed the rock, giving it the momentum it needed, that Victim became endangered. We must therefore say that by pushing the rock Jones initiated a NPET and thus initiated a moral sequence. Jones did not sustain a moral sequence; Jones initiated his own moral sequence.

In response, one might say that it was only by combining his push with the force imparted by the unknown initiator that Jones was able to contribute to the harm done to Victim; without the unknown initiator initially mobilising the rock, Jones' extra push would not

have contributed to the harm inflicted on Victim (and we might further say that Jones would have lacked the strength to get the rock rolling again had it come to a complete stop). Therefore, Jones did not initiate a NPET and Jones did not initiate a moral sequence; Jones sustained the unknown initiator's moral sequence. However, this cannot be the case for one simple reason: the unknown initiator did not initiate a moral sequence in the first place since the rock would have failed to reach and harm Victim. The already rolling rock might be a necessary condition for the harm to be inflicted on Victim (since the initial momentum was necessary for the harm to be inflicted on Victim), but it cannot constitute a sufficient condition (since the rock would not have reached Victim without Jones' extra push). But Jones' extra push of the already rolling rock is both a necessary and a sufficient condition for the harm to be inflicted on Victim (since the rock would not have reached and would not have caused harm to Victim without this extra push). This is why we can say with philosophical certitude that, by giving the rock an extra push, Jones initiated a NPET and thus initiated a moral sequence.

A further response to my claim that Jones in fact initiates a NPET and thus a moral sequence is that the type of sustaining Woollard has in mind is a situation in which the rock would reach Victim, but where Jones gave it that extra little push to make sure it did indeed reach Victim. Although this clearly isn't the case Woollard has in mind—since she acknowledges that by acting Jones "move[d] forward a sequence that would otherwise have stopped' (Woollard, 2008: 270)—it is worth consideration. If Rock is adapted to create another case, Rock 2, in which "some unknown initiator has set a large steel ball rolling and Jones gives it *an extra but unnecessary* push to ensure that the ball crushes and kills Victim" then we might say that Jones has helped to ensure that the rock reaches Victim, and so Jones has helped to sustain that sequence. However, Jones has neither initiated the sequence nor

sustained it (in any relevant sense). Jones' extra but unnecessary push means that he has not brought about a NPET and any talk of sustaining the sequence is redundant since the rock would reach Victim regardless of Jones' involvement in the moral sequence. One might wish to say that Jones is an accessory to or complicit in the moral sequence, but such an assessment lies beyond the parameters of the moral sequencing I am proposing. *Prima facie*, then, there is no room for talk of sustaining a moral sequence in moral sequencing.

However, it is quite plausible that, although the extra momentum given to the rock by Jones was unnecessary for the rock to cause harm to Victim, the extra momentum might affect the *severity* of the harm sustained by Victim. Consider another case, Rock 3: "some unknown initiator has set a large steel ball rolling and Jones gives it *an extra push* that *increases the severity of harm* to Victim". In an example of Rock 3 in which Jones' extra push of the rock causes the death of Victim, but where without the extra momentum given to the rock by Jones the rock would have still reached Victim but only maimed and not killed him, we might say that Jones sustained the moral sequence. However, this seems to be a separate category to sustaining; it is a case of what I will call *snowballing* a moral sequence.

> *Snowballing*
>
> Where an agent acts in a way that increases the severity of harm that will occur as a direct result of the initiation of the NPET.

Moral sequences are therefore 'in some relevant sense *dependant* upon the agent', but not in the ways cashed-out by Woollard; moral sequences are in some relevant sense dependent upon the agent that brings about a NPET and thus initiates a moral sequence, but we cannot say the same about agents who are involved in any sort of sustaining since the concept of

sustaining (as presented by Woollard) cannot feature in moral sequences for the reasons outlined above. Rock cannot be a case of sustaining a moral sequence since Jones' pushing of the rock that would otherwise have stopped is itself the initiation of a NPET; Jones therefore initiates his own moral sequence. Rock 2 cannot be a case of sustaining a moral sequence either since Jones' pushing of the rock was neither necessary for the rock to kill Victim nor did Jones initiate a new moral sequence since the pushing of the rock cannot constitute initiating a NPET. However, although Rock 3 is *prima facie* a case of sustaining, upon further inspection, it appears to be a different case altogether; the sustaining of a sequence has a suppressed premise, namely the notion that the severity or level of harm remains constant (since in the case of Rock discussed by Woollard the rock crushes and kills Victim). If Jones were to sustain (in Woollard's terms) a moral sequence by ensuring that the same severity of harm occurred, we would be thrown back to either Rock (but in which Jones' push initiates a NPET and thus a new moral sequence) or Rock 2 (but in which Jones' push in no way affected the outcome of harm but nonetheless was a contributing factor), neither of which, I have argued, can constitute cases of sustaining in moral sequencing. Instead, Rock 3 involves the idea that the severity of harm increases as a result of Jones' involvement; by pushing the rock, Jones *snowballs* the moral sequence, turning a NPET that was initiated by another agent into a more harmful NPET. Snowballing is a curious case in moral sequencing since Jones does not initiate a NPET but his actions nonetheless result in (increased) harm to Victim. However, the moral sequencing that I will present in this thesis will not focus on cases of snowballing in order to focus on the most salient features of moral sequencing and to ensure that a consistent and uncomplicated discussion of the practical and philosophical implications and ramifications of moral sequencing can be addressed in later chapters—that said, the issue will be returned to in chapter 5 when considering whether

responsibility can be attributed to the actions of intra-sequence agents, including those who snowball a moral sequence.

2.2.2.3. ENABLING A MORAL SEQUENCE TO CONTINUE

Any talk of a moral sequence being 'in some relevant sense *independent* of the agent' (Woollard, 2008: 266) requires qualification. Woollard thinks that both 'enabling' and 'forbearing to prevent' fall into this category. I will now discuss each case in turn.

The case of enabling is challenging for moral sequencing as currently portrayed. Woollard uses Kick (see Figure 6 in §1.2.10.) to illustrate 'enabling' in a sequence: 'The vehicle is rolling to a point where there is a rock that can bring it to a halt. Agent kicks away the rock, and the vehicle rolls to its destruction' (Woollard, 2008: 268). By kicking away the rock, Agent removes the only barrier preventing the destruction of the vehicle; by removing that barrier, Agent 'enables' both the sequence to continue and the vehicle's destruction. This concept of 'enabling' a sequence to continue and McMahan's concept of removing a barrier to harm are very similar, and so I will discuss both and their relevance to moral sequencing in tandem here. A keen reader may wish to glance back at §1.2.8. in which McMahan and the concept of removing a barrier to harm was first discussed. Since, in Kick, the rolling vehicle is on course to hit and be halted by the rock in its path (which presumably would damage but not 'destroy' the vehicle), it is only the removal of the rock by Agent that enables the vehicle to continue rolling, ultimately leading to its destruction. If we modify Kick to Kick 2 in which the removal of the barrier causes harm (e.g. leading to the death of Victim rather than or in addition to the destruction of the vehicle), then the concept of removing the (self-sustaining, operative) barrier must count as initiating a sequence since

Agent brings about a NPET. The removal of a barrier to harm is therefore a sufficient (but not necessary) condition for a NPET; every instance of removing a (self-sustaining or not self-sustaining) operative barrier initiates a new NPET and therefore a new moral sequence. Importantly, McMahan's terminology ('self-sustaining', 'operative', and 'as-yet inoperative') is also helpful for (a) establishing whether a NPET has occurred since the removal of an operative barrier counts as bringing about a NPET, and (b) identifying a neglected type of barrier, namely a dispositional barrier, whose removal I will shortly argue also constitutes initiating a NPET[44].

On the subject of self-sustaining barriers, whether a barrier is self-sustaining or not self-sustaining does not make a difference in the moral sequencing framework; whether the rock is in situ and immobile (self-sustaining) or whether an agent is holding the rock in place (not self-sustaining), the removal of the barrier by either pushing it to mobilise it (in the case of a self-sustaining rock) or relinquishing your grip on it and thus allowing it to move (in the case of a not self-sustaining rock) is an instance of bringing about a NPET and thus initiating a moral sequence.

On the subject of as-yet inoperative barriers, we might *prima facie* think that the removal of an as-yet inoperative barrier would not count as initiating a moral sequence nor bringing about a NPET in moral sequencing since the barrier is, as the name suggests, as-yet inoperative. We might imagine a scenario in which the rock in Kick is rolled down a hill by

---

[44] It is worth mentioning that I am making the assumption that moral sequencing is not concerned with *actual* operative (or dispositional) barriers—just knowing that a barrier is *ceteris paribus* a barrier is sufficient. (I make this assumption since knowing whether a barrier is actually operative is not wholly relevant for practical purposes, although epistemological rigour might demand it.)

an agent in order to roll into and settle on the path down which the vehicle is rolling, however, if the rock does not reach the vehicle in time the rock cannot be said to have acted as a barrier, and if the rock does reach the vehicle in time then the rock has acted as a barrier; but, importantly, in such a case the discussion has been redirected from talk of *removing* a barrier to harm to *installing* a barrier to prevent harm, the latter of which constitutes an intervention rather than removing a barrier (and thus bringing about a NPET and initiating a moral sequence); this case of installing a barrier, *viz.* intervening in a moral sequence, will be examined in chapter 4. Returning to the issue of removing a barrier to harm as constituting initiating a NPET and thus a moral sequence, we can conceive of a problem case that challenges the assumption that the removal of an as-yet inoperative barrier does not constitute initiating a NPET:

> *The Avalanche*
>
> Enemy lives at the foot of a snowy, avalanche-prone mountain. Sam notices that Enemy's house is below a ridge on which there is a barrier to hold-back the snow that accumulates over the winter months; without any meddling, the snow would accumulate in the winter, remain in place due to the avalanche barrier, and melt in the summer without causing any harm. Seeing an opportunity to harm Enemy, the following summer Sam weakens the avalanche barrier to ensure that once snow accumulates there the barrier will fail and that, as a result, the ensuing avalanche will kill Enemy who resides below.

Since the avalanche barrier was as-yet inoperative at the moment that Sam weakened it (during the summer), Sam cannot be said to have initiated a NPET and thus cannot be said

to have initiated a moral sequence—even though, six months later when the snow breaks through the weakened barrier, we would likely feel inclined to say that the avalanche was Sam's doing and that Sam should be held morally accountable. After all, if it were not for Sam's meddling, Enemy would (*ceteris paribus*) still be alive. However, this is only a *prima facie* problem case if we stick to a McMahanian understanding of relevant terminology. To recap §1.2.8., McMahan provides some cases to clarify the concept of operativeness. In Respirator the barrier of the switched-on life-support machine is operative since it is preventing harm from coming to the person attached to it; in The Pipe Sealer the barrier of the seal on the pipe is operative since the seal is preventing harm from coming to those who drink the water; and in The Dutch Boy the barrier of the boy's finger in the crack is operative since his finger is preventing the flood. The case McMahan provides to shed light on what it means to talk of an as-yet inoperative barrier is The Impoverished Village. Here, the barrier of the person's money is as-yet inoperative since the money has not yet been transferred and is therefore not yet preventing the death of the impoverished villagers. So, for McMahan, an as-yet inoperative barrier is a barrier that is currently inoperative. But this interpretation causes philosophical problems and is the root of the issue we face in The Avalanche.

The problem is that McMahan seems to assume that an operative barrier is a barrier that is currently holding back a harm (like in the cases of Respirator, The Pipe Sealer, and The Dutch Boy); in other words, for McMahan, an operative barrier is a barrier that is *currently in-operation*. However, I will argue that a barrier can be considered a type of operative barrier if it *would* play the role of an operative barrier under circumstances in which the barrier prevented harm from occurring. A barrier that is not currently in-operation, *viz.* is not currently holding back a harm, would on McMahan's account be considered an as-yet

inoperative barrier; if we are to look to The Impoverished Village for guidance, these as-yet inoperative barriers are simply barriers that are not yet in-place. So, for McMahan, an operative barrier is a barrier that is in-place and is currently in-operation (it is holding back a harm), and an as-yet inoperative barrier is a barrier that is not-yet in-place and is therefore not currently in-operation (it is not holding back a harm). But what about barriers that are in-place and that would function as an appropriate barrier should the situation arise, but that are currently not-yet in-operation? McMahan's example and explanation of barriers does not account for this type of *dispositionality*. To illustrate my point, consider an example of a dispositional barrier, the central reservation of a motorway: the central reservation is not an as-yet inoperative barrier since the barrier is in-place, and it is not (on McMahan's understanding) an operative barrier since it is not currently in-operation (it is not preventing a car that has collided with it from careering into the path of oncoming traffic); but the central reservation is a *dispositional barrier* since the barrier is in-place and, although not currently in-operation, would become operational and would (*ceteris* paribus) prevent harm to vehicles on the opposite side of the reservation should a vehicle collide with it. In other words, the barrier has the disposition to act as a barrier under the right conditions (in the same way that a vase has the disposition to break under the right conditions, e.g. being dropped).

If we now return to the case of The Avalanche, the problem arising from the case quickly dissipates. The concern was that, since the avalanche barrier was as-yet inoperative at the moment that Sam weakened it (during the summer), Sam cannot be said to have initiated a NPET and thus cannot be said to have initiated a moral sequence; but we nonetheless feel inclined to say that Sam should be held morally accountable for weakening the barrier. However, we can now see that the issue is merely terminological. The avalanche barrier is

not as-yet inoperative; during the summer the barrier is dispositionally operative (since the barrier is in-place but is not-yet in-operation) and in the winter the barrier is operative (since the barrier is in-place and is in-operation). Our instinct to hold Sam morally accountable for weakening the avalanche barrier during the summer is vindicated due to the fact that Sam's actions are preventing the barrier from ever being in-operation; Sam has removed a dispositional barrier that would otherwise have become operational when a situation that would have resulted in harm were it not for the barrier arose. It is for this reason that I propose that in moral sequencing both the removal of an operative barrier and the removal of a dispositional barrier constitutes initiating a NPET and thus constitutes initiating a moral sequence.

As a result of this discussion, 'enabling' a sequence to continue and removing a (dispositional or operative) barrier to harm (which for current purposes are synonymous) are (a) cases of initiating, and (b) not strictly speaking 'independent' of the person as Woollard believes; the removal of either an operative or dispositional barrier to harm brings about a NPET and thus initiates a moral sequence, thus contrarily making the moral sequence dependent on the agent removing the barrier.

However, there is a problem case of Woollardian enabling. In the example of Jones glassing in a local bar (introduced in the Introduction), we can imagine the following case.

Strangers

> Jones initiates a moral sequence by throwing a glass towards Smith. Between Jones and Smith are three strangers: Stranger 1, Stranger 2, and Stranger 3 (all strangers have recently entered the bar and are still wearing their motorbike helmets and so would not sustain harm if the glass hit them). If Stranger 1 decides not to duck out of the way of the glass, they can be said to be a barrier to harm. If Stranger 1 decides to duck, then they can be said to have removed that barrier to harm, and if Stranger 2 decides not to duck then they are now the barrier to harm. The same applies if Stranger 2 decides to duck; Stranger 3 now becomes the barrier to harm. If all three strangers decide to duck, then three barriers to harm have been removed and the glass will reach its target and will hit and harm Smith.

Strangers is a *prima facie* case of enabling in which Stranger 1, Stranger 2, and Stranger 3 do not initiate a NPET and therefore do not initiate a moral sequence; this seems to contradict my earlier claim that all cases of enabling are actually cases of initiating (since I argued that all cases of enabling involve initiating a NPET). However, this problem quickly dissipates if the nature of this problem case of enabling is properly examined. This case of enabling is actually a case of sustaining (properly understood). Even though, by ducking, each stranger removes a dispositional barrier to harm (their own body) (and they do not initiate a NPET since they do not, for example, catch the glass and throw it towards Smith), they have in effect sustained the sequence since they have provided the means for the sequence to continue; by ducking they have removed a dispositional barrier to harm (a seeming case of enabling), but by doing so the strangers sustain the moral sequence. Their

decision to duck sustains the moral sequence *but without directly acting on the current NPET*; on this understanding of sustaining (let us call this *sustaining-2* to differentiate it from Woollard's understanding of the term, which we will call *sustaining-1*), the strangers are passive with respect to the object of harm (i.e. the glass flying through the air). If one or more strangers had caught the glass and thrown it to continue its trajectory towards Smith (and the other strangers in-between), then the stranger(s) can be said to have directly acted on the NPET and, by doing so, initiated a new moral sequence. On a Woollardian understanding, the stranger(s) would be said to have sustained-1 the moral sequence; however, as we have seen, sustaining-1 is either a case of initiating (c.f. Rock) or cannot constitute being a case of sustaining at all (c.f. Rock 2); likewise, the alternative, Rock 3, is actually a case of snowballing. Instead, the concept of sustaining has to feature something similar to the case presented in Strangers; the concept of sustaining must actually be sustaining-2. Strangers therefore provides a case that *prima facie* seems to be a case of enabling but, after consideration, sheds light on the concept of sustaining a moral sequence and proves to be a case of *sustaining-2*.

> *Sustaining-2*
>
> An agent sustains-2 a sequence when they knowingly remove a dispositional barrier to harm but without directly acting on the current NPET.

However, the Strangers case provides a deeper problem, namely that it is possible for a case of sustaining-2 to simultaneously be a case of forbearing. The issue can be presented thusly:

Aggrieved Smith

> Following the case of Strangers, Smith is struck by the glass. Injured, he walks over to Stranger 1, Stranger 2, and Stranger 3. Smith demands to know why they decided to duck when they could have caught the glass. All three could have easily caught the glass and doing so would not have harmed them. If even one of them had caught the glass, they could have prevented Smith from being harmed.

The case of Aggrieved Smith highlights how, by deciding to duck out of the way of the glass, all three strangers sustained-2 the sequence, but additionally, *because they were able to catch the glass but did not do so*, they simultaneously forbore to prevent harm to Smith—they failed to install a barrier to harm (catching the glass) that would have averted the harm to Smith. If, however, the strangers protest that they were not able to catch the glass (perhaps it was traveling too quickly or they were not able to react quickly enough), then we can say that the strangers only sustained-2 the sequence (since they were not able to install a barrier to harm and thus did not forbear to prevent the harm to Smith).

Therefore, moral sequencing can maintain that enabling is in fact a case of initiating (via the discussion of The Avalanche), that the fringe case of Strangers is a case of sustaining-2, and that the further fringe case of Aggrieved Smith is a case of *both* sustaining-2 and forbearing to prevent. It can also simultaneously hold the view that sustaining-1 (as presented by Woollard) is philosophically incoherent and should not form part of moral sequencing. Sustaining-1 is: in Rock, a case of initiating; in Rock 2, a case of being an accessory to or complicit in the NPET; and in Rock 3, a case of snowballing. The concept

of sustaining, if anything, is *sustaining-2*—and, under certain conditions (c.f. Aggrieved Smith), can simultaneously be a case of forbearing to prevent[45].

2.2.2.4. FORBEARING TO PREVENT HARM IN A MORAL SEQUENCE

To recap the drive of the current discussion, I claimed that any talk of forbearing to prevent in a moral sequence being 'in some relevant sense *independent* of the agent' (Woollard, 2008: 266) requires qualification. Woollard thinks that both 'enabling' and 'forbearing to prevent' fall into this category, the former of which I discussed above and the latter of which I will discuss now.

Woollard describes a case of forbearing to prevent harm, Stayback (see Figure 5 in §1.2.10.), in which 'Agent does not interpose the rock, so the vehicle does not hit the rock and the sequence continues' (Woollard, 2008: 269). If we change Woollard's use of 'interpose' to 'intervene' (to bring the terminology in line with that of moral sequencing, c.f. §2.2.1.2.)[46], then it is plain to see that cases of forbearing to prevent harm are simply cases in which an agent does not intervene in a moral sequence. Forbearing to prevent harm is therefore the converse of intervening (c.f. §2.2.1.1.). We can frame an initial definition of forbearing thus:

---

[45] With this said, the moral sequencing presented in this thesis will not discuss sustaining-2 a moral sequence since this thesis is primarily concerned with ascertaining if and when an agent should intervene. However, whether responsibility can be attributed to an agent for sustaining-2 a moral sequence will be revisited in chapter 5.

[46] Foot uses "intervene" to refer to cases of forbearing to prevent: 'a sequence […] as already in train, and something the agent could do to intervene. (The agent must be able to intervene but does not do so.)' (Foot, 1994: 273). I highlight this to reinforce the fact that there is currently no standard use of the terms "intervene" and "interpose" and that they are often used synonymously. This provides another justification for dedicating time to identifying and defining the types of interferences (c.f. §2.2.1.2.) in moral sequencing.

Forbearing to prevent harm

An agent forbears to prevent harm when that agent is aware of the threat of harm to a Victim but fails to intervene in the moral sequence. To be a Forbearer, that agent has to previously have been a Deliberator in that moral sequence—the Forbearer must have been aware of the moral sequence and the threat of harm to Victim yet did not intervene.

It therefore simply cannot be the case that forbearing to prevent is 'in some relevant sense *independent* of the agent' (Woollard, 2008: 266) as Woollard believes. Failing to intervene, failing to install a barrier to harm, is dependent on that agent acting (acting by remaining inert or acting by doing otherwise).

Like the terminological clarifications of the term "interfering" yielding three distinct related terms ("intervening", "interposing", and "interrupting") due to the philosophical issues that arise in relation to their application in moral sequencing, similar clarifications can be made, and further related terminology can be offered, for forbearing to prevent. Using the above definition, it is unclear whether forbearing to prevent can account for a number of potential problem cases. The rest of this section will present a number of problem cases (namely possible objections to the way in which I have presented the concept of forbearing) in order to clarify the concept of forbearing and introduce sharper parameters and relevant terminology to clarify this sequence parameter.

Forbearing problem case: Installing an inappropriate barrier

An agent might fail to install an appropriate barrier (and thus fail to appropriately intervene) by installing an inappropriate barrier (and thus intervene inappropriately)—this inappropriate intervention might be akin to forbearing to prevent since that agent has failed to prevent harm; I might identify a moral sequence, I might further identify when to intervene, but I might intervene in an inappropriate way.

This case introduces the concept of the "appropriateness" of a barrier, and relatedly the concept of an "ineffective" barrier. This case asks whether an agent must install an appropriate barrier in order to appropriately intervene in a moral sequence, and whether an agent that installs an inappropriate barrier can be said to have forborne to prevent harm in that moral sequence. Issues arise when trying to pin-point what constitutes an appropriate or inappropriate intervention. What, exactly, is an appropriate/inappropriate barrier to install in a given moral sequence? And does an agent forbear to prevent harm in a moral sequence if they install an inappropriate barrier? One might think that the following constitute installing an inappropriate barrier:

(i) installing a barrier that is clearly inappropriate for the moral sequence, that is bound to fail to prevent harm from occurring to a Victim, and that any reasonable person would know would fail (e.g. shouting "Stop!" at a boulder that is rolling towards a Victim);

(ii) installing a barrier that is ineffective (e.g. pushing a rock in the way of a boulder that is rolling towards a Victim, but which subsequently breaks-up on impact with the boulder which permits the boulder to continue on its course);

(iii) installing a barrier that is only partially effective (e.g. digging a ditch along the road down which a boulder is rolling towards a Victim, but whose presence only serves to reduce the speed—but not change the course—of the boulder); and

(iv) installing a barrier that inadvertently increases the harm to a Victim (e.g. pushing a rock in the way of a boulder that is rolling towards a Victim, but which, on collision, increases the speed of the boulder and the harm done to Victim).

Let us take each in turn to assess whether each can help explicate the concept of the "appropriateness" of a barrier and can thus help to understand the concept of forbearing to prevent harm. Point (i) would certainly count as a case of forbearing to prevent harm; it epitomises the concept of an inappropriate barrier. Although shouting "Stop!" at an assailant might be an appropriate barrier to scare-off the assailant and prevent further harm to the Victim, the same barrier, shouting "Stop!", would clearly not cause a rolling boulder to halt in its path. In such a case, what we might call "blatantly inappropriate barriers"—the inappropriate installation of which would be blatantly obvious to any reasonable person—would count as an instance of forbearing to prevent, *but only so long as an appropriate barrier was available but was not installed*. If no appropriate barrier was available, I do not think we would begrudge an agent any attempt to prevent harm, even though any such action would almost certainly fail to prevent the harm—desperation, helplessness, or any other emotion or drive might make an agent attempt to intervene, even though the barrier they install is blatantly inappropriate to prevent the harm in question. But in cases in which an appropriate barrier was available, yet the agent (unwisely) chose to install an inappropriate barrier, the agent has clearly forborne to prevent harm. An "appropriate" barrier in this context can therefore be broadly construed as any barrier that any person could reasonably

expect to be (at least partially) effective and whose installation would not blatantly fail to prevent the harm in question.

Points (ii) and (iii) can be discussed together since they both involve the concept of installing an ineffective barrier (whether that be a fully or partially ineffective barrier).

> *Forbearing problem case: Installing an ineffective barrier*
>
> An agent might install an ineffective barrier by installing a barrier that is fully or partially ineffective since it does not prevent the threat from occurring or merely reduces but does not prevent the harm; I might identify a moral sequence, I might further identify when to intervene, but I might intervene in an ineffective way.

Although we might wish to state that installing an ineffective barrier is an instance of installing an inappropriate barrier, the two must be separated and differentiated. Indeed, the issue of the inappropriateness of a barrier is separate to the issue of the ineffectiveness of a barrier. There is, for instance, a huge difference between placing a small toy car and placing a real car in the path of a boulder rolling towards a Victim. We might say that both are ineffective—the toy car is crushed and the real car is knocked-aside by the force of the rolling boulder—but the former, the toy car, is an inappropriate barrier whilst the latter, the real car, is an appropriate but ineffective barrier. This is because, to draw on the discussion of (i) above, the toy car is a blatantly inappropriate barrier; no reasonable person would think that the installation of this barrier would halt the rolling boulder. Conversely, one could make a reasonable case for why the boulder should have been stopped by the real car—its weight, position, shape, and so on could lead any reasonable person to think that its

installation would halt the boulder. The concept of the "appropriateness" of a barrier therefore needs to be sharply separated from the concept of the "effectiveness" of a barrier. Just because an agent has installed a (fully or partially) ineffective barrier does not mean that the barrier was inappropriate. But does installing an ineffective barrier constitute forbearing to prevent harm?

Since forbearing to prevent harm simply means not installing a barrier to harm (*viz.* not intervening), the installation of a (fully or partially) ineffective barrier cannot count as an instance of forbearing since it involves installing a barrier to harm (albeit an ineffective one). Points (ii) and (iii) therefore cannot count as an instance of forbearing to prevent; instead, we might say that this is an instance of a *failed intervention* (in the case of a fully ineffective barrier) or a *sub-optimal intervention* (in the case of a partially ineffective barrier). There is, however, one exception to this: if the installation of a (fully or partially) ineffective barrier was installed at the expense of another agent, a would-be Intervener, installing a (more or fully) effective barrier. In such a case, we should say that the agent who installed a (fully or partially) ineffective barrier at the expense of another agent installing a (more or fully) effective barrier has forborne to prevent harm (this will be discussed in more detail in §5.1.2.).

Lastly, point (iv) is a clear case of snowballing (see §2.2.2.2.) and so cannot constitute forbearing to prevent harm. The fact that the agent installed (or rather attempted to install) a barrier to harm—even though that action increased the severity of harm—explains why this is not a case of forbearing to prevent; it is a case of snowballing.

Forbearing problem case: Failing to identify an available barrier

An agent might fail to identify an available barrier to install even though a barrier is available for them to install, and, as a result of not identifying an intervention, that agent forbears to prevent; I might identify the moral sequence, and I might also identify when I should intervene[47], but I might not be able to identify a barrier that is available to me—even though such a barrier is available to me to install.

This case introduces the concept of the "availability" of a barrier, and whether an agent must be able to identify an available barrier. I use the concept of a barrier being "available" to an agent in a broad way to refer to the availability of that agent to implement the intervention in a way that takes into account the agent's spatio-temporal proximity to the appropriate intervention, the agent's physical and cognitive capacities and disabilities, the agent's knowledge pertaining to the barrier, and so on. So, for instance, if an agent can see an appropriate intervention but is unable to reach or implement it in time, we cannot say that that intervention was available to that agent. Likewise, if an agent's physical disability prevents them from pushing a boulder, we can say that pushing the boulder is an appropriate intervention, but it is not one that is available to that agent. Moreover, if an agent cannot work or manage the barrier, such as not knowing how to drive the car required to prevent harm, then that barrier cannot be said to be available to that agent either. The concept of forbearing to prevent must therefore include the stipulation that a barrier is available to an agent and is available in a way that accounts for the issues outlined above.

---

[47] The most appropriate, or rather justifiable, moment to intervene in a moral sequence is discussed in chapter 4.

However, if an agent fails to implement an available barrier ("available" in the sense outlined above), then whether or not that agent has forborne to prevent harm is an issue that requires discussion. Since it is a matter of fact that barrier *b* was available to agent *A* to prevent harm to victim *V*, yet *A* did not install *b*, we might say that *A* forbore to prevent harm to *V*. Indeed, it is the case that, epistemically and in the strictest sense (*sensu stricto*), *b* was available to *A,* yet *A* did not install *b*. However, establishing whether *b* is or is not available to *A* can be questioned and is open to interpretation. For instance, could *A* be expected to have seen the vacant parked car with the keys still in the ignition, ran to that car, driven the car in front of the speeding boulder, and parked the car ahead of the boulder's path to prevent the boulder from killing *V* (with *A* then exiting the car and removing themselves from danger)? There are two issues that arise from this scenario/question. First, an element of hindsight is required to declare with epistemic certainty that *b* was available to *A*; whether or not *A* could have, for instance, reached the car (say, in the face of *A*'s protests that they could not have reached the car) would require *post hoc* verification. Second, it seems unreasonable to expect *A* to be totally aware of their surroundings and all possible available and appropriate barriers, and then, within the confines of their environment and apparatus available to them, install an available and appropriate barrier. Such epistemic perfection seems too strict. It does not seem feasible, pragmatic, or reasonable to require such strictness of availability in order to ascertain whether an agent has forborne to prevent harm in a moral sequence. Such strict cases of availability do not do justice to the fact that *A* was simply not aware that the barrier was available. However, at the same time, we might feel compelled to state that, in cases in which it transpires that *A* could have installed a barrier, *A* did *sensu strico* forbear to prevent harm. There is certainly a debate to be had regarding whether forbearing to prevent harm should adopt the "strict criterion of availability"—where if *b* is strictly available to *A*, *viz.* available *sensu stricto*,

then *A* can be said to have forborne to prevent harm—or whether it should adopt the "weak criterion of availability"—where if *A* is not aware of the availability of *b*, even though *b* is available to *A*, then, because it is unreasonable to expect *A* to be completely aware of all barriers that are spatio-temporally, physically, cognitively, and epistemically available to them, *A* should not be said to have forborne to prevent harm. Although I expect there to be much debate concerning which criterion should be adopted, for the reasons outlined above, I think that moral sequencing should adopt the weak criterion of availability, and with this brings the *ceteris paribus* stipulation: the concept of forbearing to prevent harm must include the stipulation that a barrier must be *ceteris paribus* available to an agent[48]. This ensures that a Forbearer is an intra-sequence agent (an agent involved in the moral sequence) rather than an extra-sequence agent—in other words, that immediately prior to being a Forbearer and agent is a Deliberator, necessitating that the agent is aware of the moral sequence.

[48] Importantly, adopting this weak criterion of availability, and employing the *ceteris paribus* principle, does not necessarily rule out accusations of negligence on *A*'s behalf. If it can be demonstrated, perhaps in a court of law, that *A* failed to install an available barrier because *A* was not observant or did not demonstrate due diligence, then we might still accuse *A* of forbearing to prevent harm—and doing so does not conflict with adopting the weak criterion.

Forbearing problem case: Ignorance of the moral sequence

> An agent might be ignorant of the moral sequence in which they could install an appropriate barrier and so fails to intervene; there might be a moral sequence happening right in front of me but, for any number of reasons (including ignorance, absent-mindedness, pre-occupation, cognitive deficiency, etc.), I might fail to identify the moral sequence and thus, when the opportunity to intervene presents itself, I might forbear to prevent harm as a result.

This case introduces the concept of "ignorance", namely being ignorant of a moral sequence in which a barrier could be installed. Can an agent forbear to prevent harm in a moral sequence of which they are ignorant? One might of course say that, in the strictest sense, it is the case that an agent has forborne to prevent harm in a moral sequence of which they are ignorant, since the fact that an agent is ignorant of the moral sequence does not mean that the moral sequence is not taking place, nor does it relinquish that agent from any culpability associated to their ignorance of the moral sequence; we might, for instance, wish to attach moral blame to that agent's failure to identify the moral sequence, as it is this ignorance that led that agent to forbear to prevent the harm. In a strict sense, then, moral sequencing might permit ignorance of a moral sequence to count as a case of forbearing to prevent, so long as the agent is able to intervene (c.f. *Forbearing problem case: Failing to identify an available barrier*, above). This will indeed have implications for attributing responsibility (discussed in chapter 5), but, to echo comments made in the preceding problem case, and in order to focus on the most salient features of intervention (namely, identifying if and when an agent should intervene in a moral sequence to prevent harm befalling a Victim), I think that moral sequencing should employ a weaker understanding of forbearing to limit Forbearers to those

agents that are not ignorant of a moral sequence. In other words, Forbearers must be intra-sequence agents who are aware of the moral sequence (in so far as they are aware of the threat of harm); extra-sequence agents (those who are not aware of the moral sequence) cannot be Forbearers. This is in line with the definition of Forbearers that I presented in §2.2.1.1., in which I argued that an agent is a Deliberator until such time that they do intervene (thus becoming an Intervener) or do not intervene (thus becoming a Forbearer), and is in line with the comment I made in footnote 40 in which I stated that those who are spatio-temporally and physically available to install a barrier but who are ignorant of the moral sequence or the threat of harm to the Victim are extra-sequence agents and what we might call *Potential Interveners*.

> *Forbearing problem case: Waiting too long*
>
> An agent might fail to identify the appropriate time to intervene and, by waiting too long or "for the right time" to intervene, forbears to prevent; I might identify a moral sequence that requires an intervention but, as a result of waiting "for the right time" to intervene, harm can occur to a Victim and I therefore fail to install a barrier and thus forbear to prevent harm to that Victim.

This problem case introduces the concept of "tardiness" and relatedly the concept of "intention", namely waiting too long to intervene and an agent's intentions behind acting or lack thereof. A Deliberator could wait too long to install a barrier and either (a) miss the opportunity to install a barrier at all or (b) install a barrier that, because of its tardiness, is (fully or partially) ineffective. But does (a) or (b) constitute forbearing to prevent? In other words, does (a) or (b) turn a Deliberator into a Forbearer? Since a Deliberator has failed to

install a barrier at all in (a), (a) is a clear case of forbearing. However, one might disagree on the grounds that forbearing to prevent harm includes the supressed premise that a Forbearer intends to forbear, whereas here we have a case in which the Deliberator does intend to intervene but simply fails to do so (as a result of waiting too long). Such a proponent might therefore say that we need to differentiate between, and provide different terminology for, cases in which a Deliberator intends to forbear and cases in which a Deliberator does not intend to forbear (but where, in both cases, the Deliberator does not install a barrier and therefore becomes a Forbearer). A Forbearer might *intentionally forebear* in cases in which they are motivated to ensure the occurrence of harm, or a Forbearer might *unintentionally forbear* in cases in which they simply fail to intervene but devoid of any motivation to ensure the occurrence of harm; however forbearing does not require the condition that actions of forbearing must be intentional—all that is required is for that agent to be an intra-sequence agent rather than an extra-sequence agent, and a Forbearer's intra-sequence agency is guaranteed by virtue of them being a Forbearer (since a Forbearer, by definition, is an intra-sequence agent). A Deliberator who fails to intervene is a Forbearer—they have forborne to prevent harm.

We can consider a related problem case (a variation of The Dutch Boy, discussed in McMahan (1993: 257) and in §1.2.8.) to clarify this claim.

Forbearing problem case: The Dutch Man

> A Dutch man, seeing that the dike is beginning to crack, considers sticking his finger in the crack to prevent the dike from breaking and flooding the town. He is already late for a date with his spouse. Confident that the crack is only small, he decides to defer attending to the crack until the next day. Within minutes of leaving to attend his date, the dike bursts and a flood engulfs the town, killing many.

In this related problem case, the dike is an operative barrier (since it is holding back the threat of a flood) whose crack undermines the integrity of the dike as a barrier. Since the Dutch man is aware of the crack yet chooses to ignore (or rather postpone addressing) the issue, has he forborne to prevent harm? Had he plugged the crack with his finger and then called for or sought help from the relevant authorities, the thought experiment invites us to believe that the dike would not have burst and that those who died from the ensuing flood would still be alive. The Dutch man therefore did forbear to prevent harm—he was an intra-sequence agent (a Deliberator) and an appropriate barrier was available to him to install. However, the fact that the Dutch man forbore to prevent harm does not *ipso facto* mean that he is *responsible* for the flood and ensuing death of the townspeople. Just because an agent has forborne to prevent harm does not necessitate that they are responsible for any resultant harm. Determining responsibility is the task of the Moral Assessor and is separate from determining whether an agent is a Forbearer. This will be discussed further in chapter 5.

Turning attention to (b)—namely whether installing a barrier that, because of its tardiness, is (fully or partially) ineffective constitutes forbearing—based on previous discussions in this section (in *Forbearing problem case: Installing an ineffective barrier*) it is plain to see

that this is not a case of forbearing since a barrier has been installed (albeit an ineffective one); this is therefore a case of a failed or sub-optimal intervention.

> *Forbearing problem case: Too many possibilities*
>
> An agent might take the concept of forbearing to prevent to the extreme and may identify a number of possible (and concurrent) moral sequences that one is forbearing to prevent; one might, for instance, say that I am forbearing to prevent harm to those in extreme poverty since I have the means to (at least temporarily) alleviate their suffering—I have both the funds and access to appropriate charities, etc.—to make sure financial aid reaches them.

This problem case introduces the concept of "sheer number" and "over-demandingness", namely there being numerous and often perpetual and concurrent moral sequences in which an agent could intervene but does not do so, and that requiring an intervention in all moral sequences would be over-demanding. The solution to this problem rests on ascertaining whether an agent can be said to be an intra-sequence agent. If, for instance, I'm sitting at home and a murder takes place on my doorstep, even if there is an available and appropriate barrier that I could have installed, I cannot be said to have forborne to prevent harm since I was not aware of the threat that formed part of the moral sequence. In this instance, I am an extra-sequence agent and not an intra-sequence agent (e.g. Deliberator) since I was not aware of the knife (NPET) being held by one of the arguing agents. However, this does not necessitate that I am devoid of responsibility for the harm that befalls the murder victim; I might have failed in my due diligence or ignored warning signs such as shouting, and I may therefore be guilty of acting negligently. But, importantly, such a decision would have to be

made by an appropriate agent, such a Judge—although this would be outside of the responsibilities of a Moral Assessor since a Moral Assessor is, in moral sequencing, concerned only with intra-sequence agents. Such an agent is therefore outside of the scope of moral sequencing. But what about cases involving moral sequences that I am aware of, yet still choose not to intervene? Consider the case of not giving to charity. I might be aware of the harm befalling a number of people, such as those suffering from or dying of malnutrition in Less Economically Developed Countries (LEDCs), in conflict areas, or in areas that have seen a natural disaster. Donating money or my time would, let us assume, at least partially alleviate such suffering—and if I am extremely wealthy, I might have the means to single-handedly alleviate or even eradicate poverty in a certain region[49]. And with the effective communication of worldwide issues via television adverts, news broadcasts and articles, and social media, we very often become aware of a number of issues—and most are aware of the issue of poverty in LEDCs. The question is now: if I do not provide money or my time (what I'm assuming are available and appropriate barriers), then have I forborne to prevent harm (to those suffering from poverty)? My awareness of the issue of poverty means that I am an intra-sequence agent and it is likely that I have the means to install at least a partially effective barrier (I can probably donate some money or volunteer some time). It is for this reason that I think that it is unavoidable that we call such instances cases of forbearing. However, as I've mentioned elsewhere in this chapter, it is important to recognise that a case of forbearing does not necessarily bring with it responsibility; it is through individual assessment, on a case-by-case basis, that a Moral Assessor would seek to determine whether a Forbearer—and in this case someone who has chosen to not give to

---

[49] Jeffrey Sachs (2006) estimated that an annual spend of around $175 billion for 20 years would end world poverty. The cost of eradicating poverty in a single country or region, especially a small one, could therefore be estimated to be a fraction of this, and well within the means of the wealthiest members of society.

a charity—is responsible for the harm that occurs. There therefore may be numerous and often perpetual and concurrent moral sequences of which we are intra-sequence agents, and we may think that there are too many to attend to and that to attend to all of them would be over-demanding, but this is an unavoidable part of what it means to be a member of society. Admittedly, it is impractical to intervene in all moral sequences in which we are an inter-sequence agent—and doing so might even initiate a new moral sequence in which we cause harm to others (giving all our money to charity might, for instance, cause harm to our children if we then do not have the means to attend to their needs). However, to which moral sequences we should attend (over others) is an issue that would require a discussion beyond the scope of this thesis. I therefore ask the reader to focus on what I will call *personal moral* sequences—namely those moral sequences whose harm could be alleviated or prevented only by those who are spatio-temporally proximate to the moral sequence—rather than *global moral* sequences—namely those moral sequences (e.g. poverty) whose harm could be alleviated or prevented by any member of society equipped with an available and appropriate barrier.

After analysing the potential problem cases discussed above, a more robust definition of forbearing to prevent can be offered to account for the discussions of each.

Forbearing to prevent

> An agent forbears to prevent harm when that agent, who is aware of the threat of harm to a Victim, knowingly remains inert with respect to intervening, and fails to install an available and appropriate barrier to the harm. A Forbearer was previously a Deliberator in the relevant moral sequence and must have been aware of the moral sequence and the threat of harm to Victim, despite the fact that they did not intervene.

It is quite possible that one forbears to prevent harm in other ways not included in the problem cases discussed above, however these cases help to identify possible issues and objections that might arise as a result of defining the terminology and understanding the concept of forbearing to prevent. These cases help to ground the concept of forbearing to prevent and have allowed us to refine the concept presented in the literature to make it more philosophically robust and viable for use in moral sequencing and more general moral issues.

It is now also plain to see that any talk of forbearing to prevent in a moral sequence being 'in some relevant sense *independent* of the agent' (Woollard, 2008: 266) is misguided. Woollard cannot maintain that forbearing falls into this category since forbearing requires that an agent, a Forbearer, knowingly remains inert with respect to intervening, and fails to install an available and appropriate barrier to the harm. Forbearing is therefore in some relevant sense *dependent on*, not independent of, the relevant agent (the Forbearer).

2.2.2.5. THE CONCLUSION OF A MORAL SEQUENCE: INTERVENTION AND HARM

I have proposed that the parameters of a moral sequence should be as follows: A sequence starts when an agent initiates a NPET and a sequence ends when this NPET occurs (harm occurs) or when another agent intervenes to prevent the NPET from occurring (harm does not occur). In this way, a sequence unravels as a series of sequence-events rather than as a series of causes and effects. But when does a moral sequence end?

A moral sequence ends either when the NPET occurs, *viz.* the threat causes harm to the Victim, or an agent (Intervener) intervenes in the moral sequence to prevent the occurrence of harm, *viz.* stops the NPET (both of these individually constitute what I will call a *sequence-outcome*). This is because the NPET, whose occurrence initiated the moral sequence, has either occurred or been prevented.

2.2.3. SEQUENCE-EVENTS

A sequence-event should be understood simply as a discrete spatio-temporal event, for example “Rock starts rolling”, “Jane stands up”, etc. Sequence-events can either be agential events (i.e. the action of an agent, such as throwing an object) or non-agential events (i.e. a strong gust of wind hurling a branch at an agent). Both agential events and non-agential sequence-events constitute sequence-events since we need to be able to account for the following sort of hybrid (and somewhat fringe) case. Suppose Initiator pushes a boulder down a hill towards Victim. Factors out of anyone’s control make the boulder go off course, say, because of a strong gust of wind blowing it off course. Then a falling tree knocks the boulder back on course. If agential events were not considered sequence-events then a

discussion of intervention could not occur (since interventions require agential input); and if non-agential events were not considered sequence-events then we would not be able to say anything meaningful about how harm was temporarily averted (by the boulder being blown off course) before being put back on its original route to cause harm (by the falling tree knocking the boulder back on course). Moral sequencing therefore must be able to account for cases in which an Initiator or anther agent acts (agential sequence-events), such as Jones throwing a glass at Smith, and cases in which non-agential events lead to harm, such as a falling tree knocking a boulder down a path towards Victim.

That said, accounting for and differentiating agential and non-agential sequence-events are philosophically and pragmatically useful for different reasons. Both agential and non-agential sequence-events can be considered when determining if and when to intervene in a moral sequence, but only agential sequence-events can be referred to when determining whether an agent (Initiator, Intervener, of Forbearer) can be held responsible for their actions (be that initiating, intervening, or forbearing).

However, an agential event must be a *positive* agential event. That is to say that an agential sequence-event must be framed as an action rather than an inaction. This is not to say that agential events in which an agent refrains from acting cannot be a part of a moral sequence, but rather that the sequence-event must be framed as an action of refraining from acting rather than a failure to act. For instance, a sequence-event could be 'Gunman stands still' but not 'Gunman does not move'. This may seem arbitrary and an unnecessary distinction and stipulation to make, but it is important in so far as agential sequence-events are understood as denoting the actions of an agent. This is similar to what Woollard (2015: 29–35) calls a 'substantial fact', namely a fact (or rather sequence-event) that is 'suitable to be

part of a sequence', gauged by the fact that a sequence-event 'tell[s] us about some change or addition to the world'. However, non-agential events do not require the stipulation that they be positive since non-agential events can describe states that are relevant to the moral sequence, be it relevant to the occurrence or prevention of harm. For instance, (non-agential) sequence-event 'Boulder rolls' (positive) may be just as important to note as 'Boulder is inert' (negative) depending on the context of the moral sequence—they both might be relevant to the decision to intervene in the moral sequence.

It must also be emphasised that, in moral sequences, we are not concerned with preventing harm in general but rather with ensuring that an agent (Initiator) does not harm another agent (Victim) in a particular moral sequence. And, in moral sequences, we are not concerned with whether an agent has acted morally or immorally by allowing a non-agential event to occur or preventing a non-agential event from occurring, since initiating a moral sequence relies on an agent (Initiator) bringing about a NPET. For these sorts of discussions and assessments, I direct the reader back to Woollard's sequencing.

2.2.4. SEQUENCE PRODUCTS

Representing the parameters of moral sequencing in the way detailed in §2.2.2. and understanding sequence-events in the way presented in §2.2.3. lends moral sequencing to being considered as a series of discrete sequence-events that, when viewed together, permit a Deliberator to determine the point at which an intervention could be justified (to prevent the NPET from harming Victim) and a Moral Assessor to determine the legitimacy of attributing responsibility to intra-sequence agents for the harm that did befall or would have befallen Victim. This section will outline these two products of moral sequencing.

2.2.4.1. REAL-TIME ASSESSMENT

In her sequencing Woollard assumes that interventions occur *between* sequence-events (think back to Stayback where intervention is only considered once the rock is already rolling). However, this sort of reasoning is problematic for moral sequencing and cannot hope to make sense of the normative evaluations we seek from moral sequencing (deciding if and when to intervene and who, if anyone, is responsible for the resultant or expectant harm). What moral sequencing attempts to capture is the missing practical element concerning the unfolding of a moral sequence in real-time, where moral sequences remain incomplete (to the extent that neither harm nor an intervention has occurred) and are unfolding sequence-event by sequence-event. If we were to keep the structure of sequence as presented by Woollard, where sequences can be viewed only in their entirety (so post-sequence conclusions can be made), agents (Deliberators) would be unable to make intra-sequence decisions (*viz.* deciding if and when to intervene to prevent harm). As well as ensuring that post-sequence conclusions can be drawn, (for Woollard, determining the moral difference between doing and allowing based on a specific sequence retrospectively), one needs to make intra-sequence *continual assessments* of a sequence as it evolves. By doing so, one will be able to determine the sequence-event at which intervention becomes justifiable (see the discussion of the "threshold of harm" in chapter 4), but early enough to ensure that there is minimal or preferably no risk of harm (this will form the basis for discussions of decision-making in chapter 3 and intervention in chapter 4). So, if we want to be able to say anything philosophically meaningful about whether an agent could have or should have intervened to prevent harm, it is essential that the system we employ is viable in real-time, where it can be updated based on actual in-the-moment occurrences. The system of moral sequencing that I will outline later in this chapter (and that will be developed throughout this thesis) can account for real-time events and so lends itself to

making the sorts of normative (and real-time) evaluations (i.e. deciding if and when to intervene) that are precluded in the sequences in the literature to date.

2.2.4.2. RESPONSIBILITY

Building on the discussion in the last section, although intra-sequence evaluations are essential for ensuring that the system can account for a continual real-time assessment of the moral sequence as it unfolds, post-sequence evaluations are also important so that the responsibility of intra-sequence agents can be ascertained for their intra-sequence actions. After the sequence has concluded, a Moral Assessor can make post-sequence evaluations based on the whole moral sequence—and can, based on the sequence-events and the actions of intra-sequence agents, attribute responsibility for the sequence-outcome (including any harm that occurs). A Moral Assessor tracks a sequence back to the Initiator and then, after considering the unfolding of the moral sequence, sequence-event by sequence-event, they can determine whether any intra-sequence agents are responsible for the sequence-outcome. This will be discussed in more detail in chapter 5.

2.2.5. TIME-FRAMES AND PERSPECTIVES

I anticipate a challenge to the system of moral sequencing that I am proposing, particularly relating to a certain aspect of its parameters. So far, I have remained silent on time-frame parameters; that is to say that I have not specified whether moral sequences can span or persist through an hour, a day, a week, a lifetime, and so on. Indeed, time-scale parameters for moral sequencing properly construed must permit such sequences to persist through any appropriate time-scale—from the moment a moral sequence is initiated to the moment it

ends (when harm occurs to a Victim or it is intervened). One could conceive of a potential (or actual) moral sequence that persists only for a few minutes, but equally one could conceive of a potential (or actual) moral sequence that persists for decades—and, indeed, everything in-between. The moral sequence involving Jones and Smith that I introduced in the Introduction would be a fairly short moral sequence lasting anywhere from a few minutes to an hour (perhaps depending on how long it took Jones to notice Smith, or how long it took Jones to become inebriated). These are the sorts of short (temporally-speaking) moral sequences that permeate this thesis and that form the basis for the majority of the ensuing discussion. However, there is much to be said for longer (temporally-speaking) moral sequences. One could for instance, conceive of the following possible case (in which there is a clear moral sequence):

The Inherited Revolver

> Andy's grandfather, Roger, has recently died. In his Will, he bequeathed a box of belongings. Upon obtaining this box, Andy looks at the contents and finds a revolver. Andy is quite young and, being fond of John Wayne films, plays with the gun in his room. Decades later, when Andy is in adulthood, he falls into financial difficulty and takes the revolver to be pawned. On his way to the pawnbrokers, Andy's bag splits and the gun falls onto the street. Andy does not notice. Vicky finds the revolver lying on the street and takes it to her friend Clara's house. There, they decide to loot Clara's father's ammunition stash and, after loading the revolver, go into a nearby forest to shoot some tin cans. Vicky hits all of her cans. She passes the revolver to Clara who, on her first shot, misses the tin can—the bullet finds its way into the head of a lone hiker, killing them instantly.

I think it is fairly uncontentious to claim that Clara shooting the revolver was the cause of death of the lone hiker—and that the Police would likely hold Clara responsible for the death. But when was the moral sequence initiated—was it when Roger purchased the revolver, when Andy was bequeathed the revolver, when the revolver fell out of Andy's bag, when Vicky picked the revolver up off the street, when Vicky and Clara went searching for ammunition, when Vicky handed the revolver to Clara, or when Clara fired (or at some other point)? Who was the Initiator—Roger, Andy, Vicky, or Clara? The way to provide an answer to these questions would be to determine who initiated the NPET. But The Inherited Revolver case shows how such a determination is problematic. It highlights a potential issue with moral sequencing, namely—because initiating a moral sequence relies on initiating a

NPET—determining when a NPET was/is initiated (and relatedly who the Initiator was/is). In other words, one could provide a compelling case for each of the possibilities above—for instance, why the moral sequence was initiated when Vicky handed Clara the revolver (making Vicky the Initiator) or why the moral sequence was only initiated when Clara fired the revolver (making Clara the Initiator). This case therefore highlights a potential pragmatic concern regarding identifying the initiation of a NPET. Indeed, identifying when the NPET was initiated is open to interpretation and debate, and it is this indeterminacy objection that could cause issues for how I've presented the concept of initiating a moral sequence above.

This objection can be addressed by demonstrating how the potential indeterminacy of the initiation of a NPET is only a *prima facie* concern. The problem originates since a proponent of this objection seeks to identity *the* point at which the NPET was initiated—they seek to pinpoint the objective moment at which a threat to a Victim was initiated by an Initiator. There are some issues with this. First, such reasoning could lead to an infinite regress. Second, such reasoning assumes there is one fixed point at which a NPET is initiated. *The solution is to assess the initiation of a NPET from the perspective of a Deliberator and/or the Moral Assessor*. The purpose of identifying when a moral sequence has been initiated is so that one can (a) assess if and when an Intervener should intervene to prevent harm to a Victim and (b) determine whether a Moral Assessor should find an intra-sequence agent (e.g. the Initiator and/or Forbearer) responsible for the harm that did befall or would have befallen a Victim. To this end, one can rely on *perspectival relevance* to determine when a NPET has been initiated. Since the decision to intervene to prevent a threat from occurring does not always rely on ascertaining who initiated the threat or how the threat came to be, a Deliberator may only be concerned with identifying if and when to intervene to prevent harm—and this can in many cases be achieved without needing to know who the Initiator

was. However, in some cases, knowledge of the Initiator is advantageous and can aid in the intra-sequence decision-making process concerning if/when to intervene (c.f. the case of Jones in the Introduction, where knowledge of Jones as a known glasser may help the Deliberator decide to intervene earlier than they would have without knowing who the Initiator was). In such cases, assessing the origin of the NPET and attributing this to an Initiator is achieved via the perspective of the Deliberator. In other words, it is from the perspective of the Deliberator whose task is to make intra-sequence evaluations concerning the threat of harm to the Victim that the origin/identity of the NPET/Initiator can be discerned. Likewise, since it is the task of the Moral Assessor to determine the responsibility of the Initiator (and/or Forbearer), the origin/identity of the NPET/Initiator can be discerned from their perspective—and, importantly, because the Moral Assessor is working in a post-sequence position to make *post hoc* evaluations, their perspective may be more historic than the perspective of the Deliberator. Accounting for perspectival relevance, particularly with reference to the perspective of the Deliberator and Moral Assessor, helps to ground claims concerning the origin/identity of the NPET/Initiator—and these perspectives may conflict, but that does not mean that either is wrong. It is entirely plausible that each have different perspectives and this might mean that each identify a different origin/identity of the NPET/Initiator.

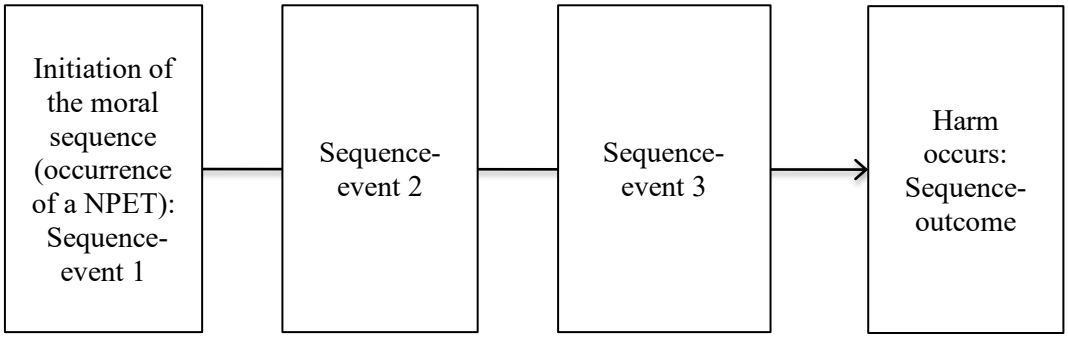
2.3. TWO ARCHETYPES OF MORAL SEQUENCING

Now that moral sequencing has been outlined, this section will synthesise this discussion by presenting two sequences archetypes: a sequence without intervention (§2.3.1.) and a sequence with intervention (§2.3.2.). Jointly, these two archetypes embody my proposed system of moral sequencing (although this will be developed in chapter 6 when considering a problem case to moral sequencing).

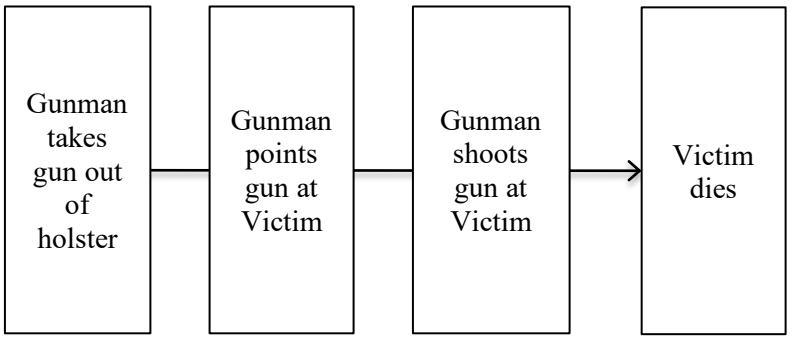
As one will notice, the structure of these sequences (presented in §2.3.1. and §2.3.2.) drastically deviates from that of Woollard's sequences (Figures 3 to 6). By providing this new system of moral sequencing via the two sequence archetypes, it is now possible to distinguish between sequences that involve and do not involve an intervention. In the same way that Woollard considered the contraries to Stayback and Kick ('Non-Stayback' and 'Non-Kick') (Woollard, 2008: 268–269) to diagrammatically illustrate how each sequence would differ with agent intervention, MSI indicates how an intervention prevents the threat from becoming actualised.

2.3.1. MORAL SEQUENCE WITHOUT INTERVENTION

Moral Sequence without Intervention (MS)



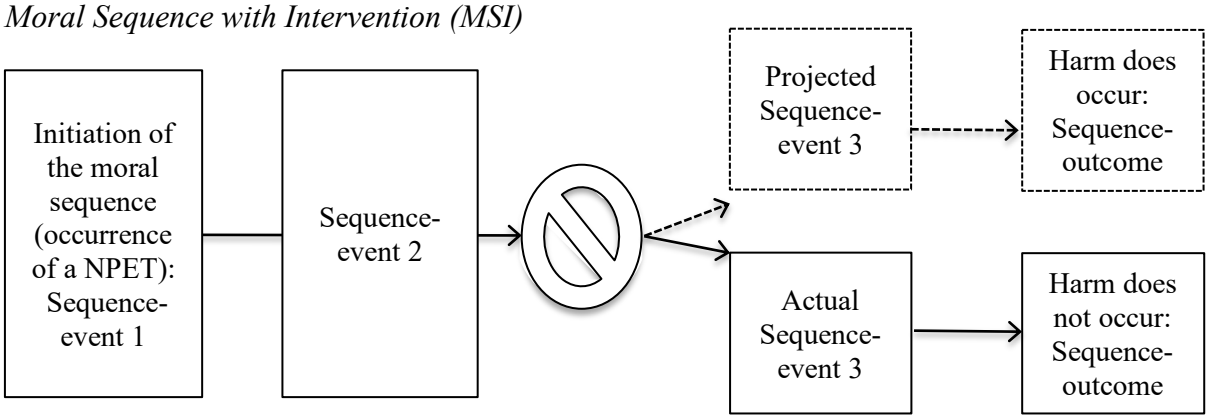
This general MS can be applied to a particular moral sequence and framed thus:



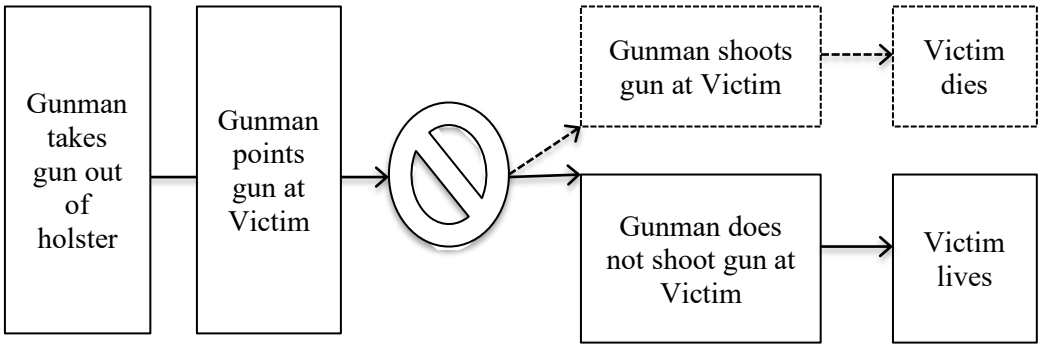
This is an archetype of a moral sequence without an intervention in which a distinct series of sequence-events occur (the progression of which is represented by each solid box, and the succession of which is indicated by the direction of the solid arrow) but where there is no intervention. Sequence-events are assigned an ordinal number according to the order in which the sequence-event occurs until the final sequence-event, which I will call the sequence-outcome. The initiation of the moral sequence is sequence-event 1, where 'Gunman takes gun out of holster', sequence-event 2 is 'Gunman points gun at Victim', sequence-event 3 is 'Gunman shoots gun at Victim', and sequence-outcome is 'Victim dies'.

2.3.2. MORAL SEQUENCE WITH INTERVENTION

Moral Sequence with Intervention (MSI)



This general MSI can be applied to a particular moral sequence and framed thus:



This is an archetype of a moral sequence with an intervention in which a distinct series of sequence-events occur and where there is an intervention. The sequence involves an intervention (represented by the "no entry sign") in which the series of Gunman's sequence-events is not permitted to continue past sequence-event 2. After the intervention, the solid arrows represent the continuation of actual sequence-events (sequence-event 3 is "Gunman does not shoot gun at Victim" and the sequence-outcome is "Victim lives"); the dotted arrows represent the projected continuation of the series of sequence-events if the intervention had not occurred ("Gunman shoots gun at Victim" is therefore projected sequence-event 3 and "Victim dies" is projected sequence-outcome).

2.4. CONCLUDING REMARKS

This chapter has dismantled the architecture of current sequencing from the literature, focussing in particular on Woollardian sequencing, and from the remains constructed a more stable and holistic model of moral sequencing. Understanding moral sequencing in this way and abiding by the sequence parameters presented and discussed provides the framework for understanding moral sequences and, as a whole, forms the system of moral sequencing that will be discussed in the following chapters.

The next chapter will kick-start this process by providing a method of decision-making in moral sequences so that we can isolate the most normatively favourable kind of moral decision so that, in chapter 4, we can discuss this kind of moral decision.

CHAPTER 3:

DECISION-MAKING IN MORAL SEQUENCING

"Should I take an umbrella to work today?", "Should I smoke a cigarette?", "Should I carry a knife?", "Should I intervene to stop those drunken men from harassing that woman?", "Should I kill the person that killed my spouse?" These are all questions that require a decision to be made; the answer to which is either "Yes" or "No"[50]. But what sort of decisions are they?

The only commonality between these questions is that they are all an appeal to oneself; they involve asking oneself: "What should *I* do in situation *x*?" But this question is, in itself, problematic. In asking oneself these questions, is one asking "What should I *generally* do in situation *x*?" or "What should I do in this *particular* situation *x*?" This reflects the two types of questions that one can ask oneself: *general* questions about what one should do in all circumstances of *x*, and *particular* questions about what one should do in a particular situation *x* in which one finds oneself. Both of these questions can be decided upon in two ways: one either makes a *reactive* decision to do *x*, that is, one can decide in the spur of the moment to act in a certain way, or one makes a *calculated* decision, that is, one can make a decision, either partially or fully, in advance of the situation presenting itself. I will argue that calculated decisions that are made in part rely on an added *reactive* element, whilst those made in their entirety are *pre-planned*. Another concept in play is a distinction

---

[50] One could, of course, answer "Maybe", but this response is arguably a placeholder until one decides to answer "Yes" or "No", whether verbally or indicated by his chosen course of action.

between the *types of decisions* required by these questions. "Should I take an umbrella to work today", for example, is a clear instance of a *non-moral decision*, whilst "Should I kill the person that killed my spouse?" can be regarded as a *moral decision*. These examples of non-moral and moral decisions, I think it is safe to assume, would be met with little resistance; they are clear examples of their respective types of decisions. But what about those in-between? Under which category does "Should I intervene to stop those drunken men from harassing that woman?" fall? This question will be the focus of §3.1., after which I will discuss, in asking oneself any of these questions, the *kind of question* one is asking oneself. For instance, in asking "Should I intervene to stop those drunken men from harassing that woman?" is one asking oneself "What *ought* I do" as a rational human being, or is one in fact asking oneself "What *will* I do?" if I witness such a situation? These questions reflect the difference between *normative* questions (these can also be prescriptive in nature) and *descriptive* questions that will be discussed in §3.2.

This chapter will therefore consider the following four lines of enquiry, which I take to be necessary in an investigation into the nature of decision-making in moral sequences:

(1) The distinction between *non-moral* and *moral* decision-making;

(2) The distinction between *general* and *particular* moral decisions;

(3) The distinction between *reactive* and *calculated* decisions; and

(4) The distinction between *normative* and *descriptive* models of decision theory.

This chapter will not be concerned with understanding the motivations or causes for deciding to act; rather, after the above topics have been discussed, I will be in a position to (a) highlight the problems with formal systems of decision-making (relevant for understanding how decisions can be made in real-time), and (b) provide a case for why what

I will call reactive-calculated moral decisions yield a normatively reliable process on which to base a decision to intervene in a moral sequence. This will pave the way to the next chapter discussing a particular reactive-calculated moral decision, namely deciding if and when to intervene to prevent harm, to establish when a decision to intervene is justifiable.

3.1. MORAL DECISION-MAKING

This section will consider three contrasting cases of types of decisions and will, based on this discussion, present five kinds of moral decisions and will isolate the most relevant kind of moral decision for moral sequencing, namely reactive-calculated moral decisions. Chapter 4 will then discuss a particular reactive-calculated moral decision, namely deciding if and when to intervene to prevent harm, to establish when a decision to intervene is justifiable.

3.1.1. NON-MORAL *VS.* MORAL DECISIONS

All decisions can, broadly speaking, be divided into two categories. Decisions are either non-moral or moral in their nature[51]. These two types of decisions can be differentiated

---

[51] One might question whether the dual-categorisation of decisions being either moral or non-moral can capture all categories of decision-making; whether decision-making is moral or non-moral is not "black and white". On this I ask the objector to reconsider their objection in the light of the Law of Excluded Middle (LEM) (which states that for any statement p, either p is true or $\neg p$ is true, *viz.*, $p \vee \neg p$). Applying this to moral decision-making, either a decision is moral (p) or ($\vee$) a decision is non-moral ($\neg p$). I therefore ask the reader to hold that LEM extends to the current issue and demonstrates that a decision either is moral or it is not. It is, however, worth noting that Intuitionists deny LEM. For a definition and overview of Intuitionism, see Anthony Flew (1979: 178); for a succinct overview of why

based on a number of factors that, either singularly or when combined, demarcate non-moral and moral decisions. I will not list all possible differential factors here, for there are too many to note. Instead I will focus on one differential factor, namely that a decision involves a *principle consideration of harm*: if a decision *is* predominantly based on a consideration of harm, then it is a moral decision; if a decision *is not* predominantly based on a consideration of harm, then it is a non-moral decision.

Some clarificatory comments are required in order to demarcate this factor. First, the scope of "harm" must be clarified. In its ordinary use, "harm" often refers to and encompasses both harm to oneself (*self-regarding harm*) and harm to other agents (*other-regarding harm*). I propose that, for the purpose of this thesis, this scope is reduced to other-regarding harm only. This is not to say that self-regarding harm is not itself a harm worth considering or the sort of harm that constitutes a moral decision; I wish to limit the scope of harm to other-regarding harm in order to (a) circumnavigate issues associated with self-regarding harm (such as determining whether to prevent an agent from harming himself, e.g. self-harm, smoking, suicide, etc.), and (b) to reflect the notion that other-regarding harms are rights-violations. It would be odd for one to say that one can violate one's own rights—for instance, it would be incongruous with our ordinary understanding of rights to say that, in harming oneself, one has violated one's (negative) right to non-interference (c.f. §1.2.5.). However, other-regarding harm is compatible with rights-violation—if an agent harmed another agent, one could say, without fear of transgressing our ordinary understandings of

---

Intuitionists deny LEM, see Frances Howard-Snyder (2009: 373–374). However some logicians, including Howard-Snyder et al. (2009: 374), think that the Intuitionist's reasoning is unsound since 'it is *false* that the truth of a statement consists in its provability. The truth of a statement consists in its describing things as they are'. I therefore ask the reader to reject the Intuitionist's denial of LEM.

rights, that the agent has committed a rights-violation, *viz.* the agent has violated the other agent's (negative) right to non-interference. So, if an agent is deciding whether to ski down an insipid but safe route down a mountain or to ski down a fun but riskier route, and if one assumes that no other agent is in the vicinity or at risk, then I ask the reader to consider this agent to be making a non-moral decision—even though the agent's decision, if the agent chooses the fun but risky route, could lead to the agent being harmed. There are of course other rights-violations besides those associated with harm, and one could easily put forward a case for holding that such alternative rights-violations should be a factor in deciding whether a decision is moral or non-moral. However, all that is important for this chapter and for this thesis is that the scope of the term "harm" is defined in order to understand one of potentially many differential factors that divide moral and non-moral decisions.

With these remarks and clarifications in mind, non-moral decisions, moral decisions, and the difference between them can be summarised and exemplified by considering the following example.

Betty's decision

> Betty is about to leave her house to walk to a nearby friend's house for evening drinks. She's got her purse, her coat, and her keys. She opens the door to leave. At that moment, she remembers the news broadcast from earlier that day. There has been a spout of local muggings that has left their victims hospitalised. She considered cancelling her evening plans, but quickly dismissed the idea; why should she let her fear dictate what she does? But she decides not to leave the house unprepared. She grabs hold of a nearby letter opener, puts it in her coat pocket, and briskly exits her house.

Is Betty's decision to carry the letter opener a moral or non-moral decision? The object itself, the letter opener, is first and foremost a tool—the blade is used for opening letters. So when Betty usually decides to carry the letter opener over to a pile of unopened post, she usually has in mind the task of opening a letter. We might say that, when deciding whether to carry the letter opener, Betty's reason for doing so is to open a letter (and not to, say, harm the Postman). Betty's decision to carry the letter opener over to her pile of post therefore seems to be a clear case of a non-moral decision. But what exactly grounds this claim that Betty's decision here is a non-moral one? We can agree that the letter opener is a tool used to open letters, but this does not mean that the object cannot be used as a weapon. Indeed, although the letter opener's blade is blunt, it could inflict serious and even life-threatening injuries on a person. So an object whose apparent primary purpose (or we might say *telos*) is not to cause harm but to be used as a tool (to open letters) does not *ipso facto* mean that the object cannot be used to cause harm, and so an object's *telos* cannot be used to ground claims of the moral or non-moral nature of a decision to, for example, carry that

object. Whether a decision (here whether to carry a letter opener) is moral or non-moral therefore cannot be determined simply by reference to the object itself. Let us therefore go back a step and appeal to Betty's reason for carrying the letter opener in her coat pocket. When Betty has previously decided to carry the letter opener (over to her pile of unopened letters), her decision to carry it was based on a consideration of wanting to open her letters. Here, Betty's *principal consideration* when carrying the letter opener is "I want to open those letters"—and it is for this reason that we can consider Betty to have made a non-moral decision. (Betty might have other reasons for carrying the letter opener, for example to ensure that she can put it somewhere safe, but these are *subsidiary considerations*). However, Betty's decision to carry the letter opener to her friend's house is different. Betty's *principal consideration* when carrying the letter opener is "I want to protect myself". Betty's decision to carry the letter opener to her friend's house carries with it a principle consideration of harm—whether that be a deterrence due to its potential to inflict harm or the fact that she would use it or attempt to use it to inflict harm on any assailants; in other words, Betty principally decided to carry the letter opener due to its potential harm-causing attributes. (Again, Betty might have other reasons for carrying the letter opener, for example to show her friend her new and expensive letter opener, but these are *subsidiary considerations*). Betty could brandish the letter opener as a deterrent to would-be muggers or those who attempt to mug her, or she could use the letter opener as a weapon and attempt to harm or actually harm any assailants. It is for this reason that we can consider Betty to have made a moral decision.

That said, one might object to my categorisation and framing of what constitutes a 'moral decision'. As I mentioned at the fore of this section, I offer and ask the reader to consider only one kind of moral decision, namely those that predominantly involve a consideration

of harm (or, more specifically, those decisions in which the principal consideration is that of harm). There are a number of other ways in which one might frame a moral decision, and which may or may not involve harm as a factor. Some authors refer to how an agent must have 'moral awareness' to make a moral decision. Rest (1986) says that moral awareness relies on 'identifying what we can in a particular situation, figuring out what the consequences to all parties would be for each line of action, and identifying and trying to understand our own gut feelings on the matter' (Rest, 1986: 3), and that for an agent to be morally aware 'the person must have been able to make some sort of interpretation of the particular situation in terms of what actions were possible, who (including oneself) would be affected by each course of action, and how the interested parties would regard such effects on their welfare' (Rest, 1986: 7). This view is nicely summarised by Jones (1991: 380) who argues that '[f]or the moral decision-making process to begin, a person must recognize the moral issues'. Importantly, a moral decision is not moral because the outcome of the decision tallies with a particular moral theory, rather it is moral because, in Tenbrunsel and Smith-Crowe's (2008: 565) words, 'the ethical relevance of the issue at hand has been recognized and this recognition prompts a consideration of moral implications, but it does not necessarily lead to ethical decisions'. Indeed, it is interesting to note that current research cannot demonstrate that moral awareness leads to actually making what many would consider to be a moral decision: although, for instance, Singhapakdi and colleagues' (1996) research demonstrates a link between moral awareness and moral decision-making, the literature is full of contradictory findings. Valentine and Fleischman's (2003) research, for example, did not find a link between moral awareness and making moral decisions[52]. It is

---

[52] However, I do concede, as Tenbrunsel and Smith-Crowe (2008: 552) have argued, that '[h]ow decision makers construe the dilemmas before them is critical to whether decision makers achieve moral awareness or not'. I think that these 'decision frames' ('the type of

this disparity in the literature that has prompted Tenbrunsel and Smith-Crowe (2008: 553–555; 2008: 565–566) to create a new typology that separates 'intentionality' from 'ethicality'; decisions are, according to this typology, divided into four categories based on the process by which the decision was made (moral or amoral decision making, *viz.* moral and non-moral decision making) and the resulting decision (ethical or unethical). These four categories can be described as follows:

(1) 'intended ethicality': an agent makes an 'intentionally ethical' decision if that agent makes a moral decision;
(2) 'intended unethicality': an agent makes an 'intentionally unethical' decision if that agent makes an amoral (a non-moral) decision;
(3) 'unintended ethicality': an agent makes an 'unintentionally ethical' decision if that agent makes a decision that that agent believes to be amoral (non-moral), but where the decision is in fact moral; and

---

decision that individuals believe that they are making' (Tenbrunsel and Smith-Crowe, 2008: 561; c.f. Tenbrunsel and Messick, 2004) offer an important insight into what separates moral decision-making from non-moral decision-making, but they can simultaneously impact on the extent to which an agent is morally aware. If a decision is framed as a moral decision, then an agent can bring to it moral awareness. If, however, a moral decision is framed as a non-moral decision, then an agent might think that he is mistaken in being morally aware during his decision-making process, and he might therefore disregard his moral awareness based on the (framing of the) situation. This might, I submit, explain some agents' troubles in dealing with conflicts—external sources could provoke a non-moral decision frame even though the agent is morally aware, and this could confuse an agent and even make the agent disregard his moral awareness in favour of non-moral awareness (whatever that might consist of) that more properly align with the situation (see Tenbrunsel and Messick (1999) for a discussion of how a sanctioning system can influence how an agent perceives a situation).

(4) ‘unintended unethicality’: an agent makes an ‘unintentionally unethical’ decision if that agent makes a decision that that agent believes to be moral, but where the decision is in fact amoral (non-moral).

Tenbrunsel and Smith-Crowe (2008: 555) argue that ‘ethical [moral] decision making is predicated on whether decision makers are morally aware’; in other words, if an agent is morally aware then that agent is engaged in a moral decision-making process, and if the agent is not morally aware then that agent is part of a non-moral decision-making process[53]. Other authors have noted the importance of moral awareness too: Hunt and Vitell (1986: 7) claim that ‘[i]f the individual does not perceive some ethical content in a problem situation, subsequent elements of the model do not come into play’; Clarkeburn (2002) found that, similar to Yetmar and Eastman’s (2000) findings, it was possible to determine whether a participant had moral awareness based on the participant’s answers to various scenarios (i.e. in order to produce milk that is used to treat cystic fibrosis, should genetically modified cows be created?)[54]. Tenbrunsel and Smith-Crowe’s (2008) insistence on aligning moral decision making with intentionality is not a topic that will be discussed here. This is because moral decision-making can be discussed meaningfully without reference to intentionality. To achieve this, I ask the reader to take only a small step in the direction of common-sense morality. Many philosophers, social scientists, and other researchers have relied on, and continue to employ, a shared understanding of the stake of harm in moral decision making and moral judgement: Cushman, Young, and Hauser’s (2006: 1083) study involves three

---

[53] It is worth noting that Tenbrunsel and Smith-Crowe (2008: 555) call the converse of moral decision making ‘amoral decision making’, although the two terms can be used interchangeably for current purposes.

[54] For a good overview of moral awareness and the literature on this topic, see Tenbrunsel and Smith-Crowe (2008: 555–564).

principles comprising different sorts of harm; Appiah (2007) ask participants to consider a case involving harm to the environment; Butterfield, Treviño, and Weaver's (2000) study involves a consideration of the probability of harm in a clinical setting; May and Pauli (2002) discuss how the probable magnitude of harm shape moral expectations; and so on. I therefore posit that harm is subsumed in moral awareness, *viz.* we are morally aware of the ethicality of harm in the moral decision-making process. Harm is therefore at least one differential factor between moral and non-moral decision-making. Thus, this thesis' focus on harm as one differential factor is not as short-sighted and needlessly narrow as one might initially consider. This thesis will henceforth assume that agents have a moral awareness that is sensitive to the importance of harm as one (but not necessarily the only) differential factor between moral and non-moral decision-making.

3.1.2. GENERAL *VS.* PARTICULAR MORAL DECISIONS

Moral decisions can either be decisions about general moral situations, such as what one should or would do in a generic situation, or a decision about a particular moral situation, such as what one should or would do in a specific situation. On the one hand, one might make a *general moral decision* that "I will not kill anyone". This decision could be based on the belief that "Killing is wrong", which itself is based on a number of religious, personal, and social motivations, which consequently informs their general decision that any questions in the form of "Should I kill *P*?" should be met with a negative decision. Conversely, one might make a *particular moral decision* that "I will kill my boss", based on a number of factors and motivations. The distinction between the two is therefore a difference in logical constant or variable; a general moral decision is of the form "I will/will

not do *x* in situation *y*", whereas a particular moral decision is of the form "I will/will not do *x* in situation *A*".

But what happens when there is a conflict between the two? The agent that made the general decision not to kill might simultaneously have made a particular decision that his spouse's life is more valuable to him the life of a stranger, and where any question in the form of "Should I kill a stranger if it guarantees that my spouse's life will be spared?" should be met with a positive decision[55]. This decision-addendum could either be pre-emptive, that is, made in advance of any moral situation presenting itself, or it could be made in the moment of a moral decision. Either way, the two decisions remain in conflict. Can this conflict be solved, or does it simply reveal an inconsistency in an agent's decision-making process? I do not think that this reveals an inconsistency; instead, such a conflict indicates that general decisions simply cannot be married with most particular decisions. There is no conflict that necessarily needs resolving, rather, in such a conflict, one will be left having to decide which decision to honour; one decision will prevail, and this will reflect the commitment one has to that sort of moral decision[56].

---

[55] I am assuming that if he decides to not kill the stranger then he is in effect permitting his spouse to die. By deciding to act in any way other than killing the stranger, he decides to act in a way that permits the death of his spouse; he decides to permit his spouse to die. I use the term 'permitting/permit' instead of 'killing/kill' or 'letting/let' so as to avoid any issues associated with the distinction between killing and allowing to die (discussed in chapter 1).

[56] One could, of course, abstain from deciding between a general and particular decision altogether. The agent left to choose between upholding his decision not to kill and having to choose between saving either his spouse or a stranger might abstain from making a decision. One might therefore argue that such an agent that abstains from deciding between his two conflicting decisions can be said to have maintained both his general and particular decision. I do not, however, agree that this is the case. By abstaining from deciding which decision to honour, the agent *is* inadvertently deciding; in this case, he is deciding to honour his general decision to not kill.

This said, it seems as though general decisions are usually *calculated*; that is, they are decided in advance of a moral situation. Likewise, particular decisions can be calculated, but they can also be, and arguably are usually, *reactive*; that is, they are decided during a particular moral situation.

3.1.3. CALCULATED *VS.* REACTIVE MORAL DECISIONS

"Dad, if you had to save either Mum or me, and you had to choose, who would you save?" Such a question might prompt Dad to consider what he would do if such a moral dilemma, for whatever reason, presented itself. He might consider a number of factors, including: appealing to whatever ethical principles he holds or follows, whether this involves following religious doctrine (e.g. the Ten Commandments), philosophical or ethical theories (e.g. Utilitarianism), or a personal attitude towards the matter (e.g. an idiosyncratic worldview or belief) (*following principles*); he may reflect on and consult his own feelings towards both his young son and his wife, or consider how he might feel knowing that he had chosen one over the other (*consulting emotions*); and he may even make some form of computational or probability-driven calculations (i.e. a hedonic calculus, Markov decision process, Bayesian probability) to help him make a decision (*formal decision-making*). Such a question might further prompt him to consider other moral dilemmas or ethical situations, both ordinary and bizarre. These sorts of "armchair decisions", where the moral deliberator is separated from any actual moral situation, are characteristic of calculated decisions. There are, however, two different types of calculated decisions: those that are *pre-considered* and those that are *pre-planned*.

In a calculated decision one has the luxury of time to (at least partially) consider the decision one is making; one can calculate one's decision based on a number of factors and influences, such as following one's principles, consulting one's emotions, or utilising a system of formal decision-making. There are a number of conceivable situations in which moral decisions can be made over a significant period of time that would allow enough time for a calculated decision to be made, although the nature of these calculated decisions vary according to whether a calculated decision is pre-considered or pre-planned. To illustrate the difference, consider the following example of a pious man and the Breivik case.

Consider an example of a pious man who, after a period of reflection and introspection in which he consulted religious scripture/doctrine (e.g. Exodus 20:15) and his own personal ethical views, concludes that he thinks stealing is morally wrong, and so concludes that under no circumstances will he steal. He therefore decides upon a *pre-considered moral decision* about a *general moral decision*: never steal. Compare this to the case of Anders Breivik who, on 22 July 2011, massacred 77 people in Oslo and Utøya, Norway. During his court trial, Breivik revealed that he had made a nine-year financial plan that would ensure he could fund his attacks (Taylor, 2011)—i.e. money to rent enough land to justify buying the quantities of fertiliser he would need for making explosives; and claims to have spent five years meticulously planning the logistics of the attacks—although the court could only find evidence of two years planning (Pidd, 2012). Breivik's moral decision can therefore be considered to be: (a) calculated in so far as he decided upon his actions in advance of him being in that situation; (b) pre-planned in so far as he decided upon a course of action for a particular, rather than general, moral situation, *viz.* "To do *x* in (a particular) situation *A*" rather than "To do *x* in (a general or generic) situation *y*". The Breivik case is thus paradigmatic of a *pre-planned moral decision* about a *particular moral decision*: Breivik's

decision to kill was not spontaneous but involved a long process of deciding who he should target, where his attacks should take place, how his attacks should be carried out, etc. (calculated); and he decided to kill as many as he could in Oslo and Utøya (pre-planned).

However, this raises the question: is it possible to make calculations in advance, *but not necessarily commit oneself to a course of* action, so in a situation that demands quick decision-making in an unusual, dynamic, or reactive situation, one can make a decision *based on* a calculated decision? I will refer to these types of moral decisions, where a decision about how to act in a particular moral situation is guided by a calculated decision as *calculated-reactive moral decisions*. Consider the use of triage during a military operation, natural disaster, or a large-scale incident such as a terrorist attack. The UK Armed Forces use the SMART Incident Command System to prioritise casualties. This is especially used when there is a mass-casualty incident (MCI), where medical personnel need to determine how to prioritise their limited resources/medical care, which allows them to decide which casualties to treat and in what order. Casualties in a MCI are prioritised according to 4 categories: Dead or Non-Salvageable (casualties that are either dead at the scene, require such advanced treatment that death is expected, or where even with medical attention or medical evacuation (MEDIVAC) the casualty will likely die); Priority 1 (casualties that require immediate medical attention or MEDIVAC); Priority 2 (casualties that require medical attention, but whose injuries are such that medical attention or MEDIVAC can be delayed until after Priority 1 casualties have been treated); and Priority 3 (casualties that are walking wounded, but whose injuries are such that treatment can either be received at home, or whose treatment can be postponed until both Priority 1 and Priority 2 casualties have received medical attention or MEDIVAC). During a MCI, medical personnel have to decide which priority label to assign each casualty. The decision about

what types of injuries would determine a casualty as being Priority 1 rather than Priority 2, for instance, was *calculated*, *viz.* planned, prior to the particular MCI in which the medical personnel find themselves involved. The SMART Incident Command System that determined the priority level benchmarks was determined in advance of the situation, and so can be considered as a calculated procedure. However, the medical personnel's decisions that are based on this calculation are *reactive* in as much as they have to make a decision *in the situation* about how to apply the calculated system to this particular MCI, and thus decide which casualties are assigned the different priority levels. This triage case is therefore paradigmatic of a *calculated-reactive* moral decision: the decisions of the medical personnel are firstly calculated, that is, they decide the what sort of injuries and casualties warrant certain priority levels, after which a secondary reactive element comes into play, that is, they decide which casualties should be prioritised during a particular MCI. However, it is not always possible for one to appeal to a calculated decision on which a decision about a particular moral situation can be made. For starters, it is hard to imagine that one could make calculations for *every* moral sequence imaginable; even someone who has too much time on their hands will be unlikely to have enough time to conceive of, and then calculate, what one would do in certain types of moral sequences. (Not to mention the epistemic problems of this claim.) But these epistemological and time-limited objections do not discount the possibility that such calculations could be made for a finite number of moral decisions. Indeed, the case of triage appears to be a paradigmatic example of how these sorts of calculated moral decisions can (and often do) incorporate a reactive element.

However, whilst the formal decision-making process is done *in advance* of a moral situation presenting itself in calculated-reactive moral decisions, there is another closely linked, but wholly different, kind of moral decision. *Reactive-calculated moral decisions* are decisions

in which the formal decision-making process occurs not prior to, but rather *during*, a particular moral situation. Firstly, however, there is a reactive element: one reacts to a moral situation, be it an emotional reaction or otherwise and, whether one is knowingly or sub-consciously informed by such reactive elements, this reaction then plays a role in what courses of action are to be considered during the formal decision-making process that follows. This will be the topic of the later sections of this chapter, specifically regarding its application to intervening in a moral sequence, but to illustrate this type of decision for comparison with the others this sort of decision can be likened to a decision of whether or not to cheat on one's spouse. If one has cheated before, and as a result has experienced a negative emotion, say a feeling of guilt, then this emotion, whether knowingly or not, can affect one's decision-making process. This negative emotion linked to the outcome of cheating can narrow the options one considers during the decision-making process, and ultimately impacts on the decision one makes.

It is also worth noting that there is a significant difference between calculated moral decisions that involve either a commitment to a pre-considered (general) course of action or a pre-planned (particular) course of action, calculated-reactive moral decisions that involve a reactive moral decision to be made based on a (pre-considered or pre-planned) calculated decision that has been decided on in advance, reactive-calculated moral decisions that involve reactive elements narrowing the options one considers during the formal decision-making process, and *purely* reactive moral decisions. In these sorts of decisions, time is usually of the essence and the situation demands that a moral decision must be made in a situation where no previous commitment to a course of action has been decided. Consider the legal case of *E7 v Holland* (2014). E7 was a Firearms Police Officer. E7 and two other police vehicles stopped a vehicle that was thought to be carrying guns. E7 claimed that one

of the suspects in the vehicle, AR, had reached down in the car. E7 thought that AR was reaching for or had picked up a gun, and so E7 fired at AR, six shots of which hit AR. Three firearms were later found in the suspect vehicle, but a public inquiry found that E7 had fired 0.06 of a second after his car had stopped. This meant that E7 could not have seen AR reach down in the car, nor could he have rationally believed that AR had picked up a gun. It was determined that E7 firing on AR was not proportionate and was not lawful. E7 therefore seems to have made a reactive decision to shoot AR in as much as E7 decided during that particular situation upon a course of action. The fact that the decision to shoot took only 0.06 seconds indicates that it is unlikely that E7 could have made a calculated decision; what is more likely is that E7 simply responded to the situation and momentarily decided to shoot[57]. It is this sort of purely reactive moral decision that many people seem to make when confronted with an everyday moral situation, or, at the very least, involve a reactive element. Even the pious man might be forced to reassess his pre-considered moral decision in light of a particular situation. If his children are starving, and he has no means of buying food, even the most pious man would be hard-pressed to maintain his calculated decision to never steal.

[57] One might argue that E7 could have, in fact, made either a calculated moral decision instead of a purely reactive one. Indeed, E7 might have made the pre-considered moral decision to always shoot a potential gunman regardless of whether the nature of the situation casts doubt over whether the supposed gunman is actually carrying a gun. However, for the purposes of this example, and because it seems more likely that E7 simply reacted to the situation, I consider this an example of a purely reactive moral decision.

3.1.4. FIVE KINDS OF MORAL DECISIONS

Based on the discussions in §3.1.1. to §3.1.3., these five different kinds of moral decisions can therefore be summarised as follows:

(1) *Pre-considered (calculated) moral decisions*: Decisions about general moral situations, where one decides upon *and commits to* a course of action for a generic situation *x,* where *x* represents a generic type of moral situation. (E.g. "If confronted by a gunman, I will always shoot them" or "If I had to choose between saving my wife or young son, I will always save my son".) In these situations, there is no reactive moral decision to be made, as the course of action for that generic moral situation has been pre-considered.

(2) *Pre-planned (calculated) moral decisions*: Decisions about particular moral situations, where one decides upon *and commits to* a course of action in advance. (E.g. Breivik deciding upon killing as many as he could in Norway on 22 July 2011, and then committing himself to act on that decision.) Again, there is no reactive moral decision to be made, as the course of action for that particular moral situation is pre-planned and was decided on in advance of the actual situation.

(3) *Calculated-reactive moral decisions*: Decisions that involve both a *prior* calculated decision and a reactive element. These decisions involve one contemplating and deciding upon a course of action for either a general or particular moral situation (this calculated decision can therefore be either pre-planned or pre-considered) *in advance* of a moral situation; one decides, in advance, what to do in either a general or particular moral situation *x*. However, there is also a *reactive element* in play in

so far as one either has not decided to commit to a course of action in advance (and so, strictly speaking, cannot be considered to be pre-considered or pre-planned moral decisions) or, when in that situation, one decides to amend one's prior moral conviction, and so cannot be said to be following the original pre-considered or pre-planned moral decision. (E.g. the use of triage in a MCI, where the decision on how to prioritise casualties is calculated and decided on in advance, but where the moral decision of which casualties are assigned particular priority levels is a reactive decision based on a number of factors relevant to the situation of a particular MCI.)

(4) *Reactive-calculated moral decisions*: Decisions made about a particular moral situation and involves both a calculated decision and a reactive element, but where both elements are decided *during*, not in advance of, a moral situation. (E.g. deciding to cheat on one's spouse, or deciding if and when to intervene in a moral sequence.)

(5) *Purely reactive moral decisions*: Decisions made in a particular moral situation, where one has not made a prior calculated decision on how to act, and therefore does not commit oneself to act in a way that is either pre-considered or pre-planned. (E.g. finding oneself in a situation in which one has to make a decision about which sibling to push out of the way of a speeding car, or whether to save one's spouse or young child from a gunman.)

To clearly see what each moral decision involves, see the table below.

Table 1: Kinds of Moral Decisions

	What elements are involved?		When is the decision made?		What type of situation does it apply to?	
	Calculated	Reactive	In advance	During	General	Particular
Pre-considered	x		x		x	
Pre-planned	x		x			x
Calculated-reactive	x	x	x	x	x	x
Reactive-calculated	x	x		x		x
Purely reactive		x		x		x

But how are these moral decisions actually made? Is there a difference between how decisions are (descriptively) made, and how decisions should be (normatively) made? After all, even purely reactive moral decisions, where no calculations are (or seem to be) made can be, and sometimes are, explained retrospectively. And what does this say about the practicality of the formal decision-making processes that underpin calculated moral decisions (and perhaps the purely reactive ones that are analysed and justified in retrospect)? The rest of this chapter will consider decision theory in an attempt to oust formal systems of decision-making as impractical for moral decision-making (normatively speaking) and support the idea that reactive-calculated moral decisions are (normatively) the best method for moral decision-making and can be used to justify a decision to intervene in a moral sequence.

3.2. DECISION-MAKING AND ITS PROBLEMS

Broadly speaking, there are two schools of thought that discuss the process of decision-making. *Normative* theories discuss how decisions *should be* made; this can also incorporate a *prescriptive* element, where the theory aims to describe how people *ought to* make decisions. Both normative and prescriptive theories aim to provide models or systems that describe how a rational human being should or ought to decide upon a course of action. Normative theories employ formal logic, probability theory, or decision theory to decide how decisions should or ought to be made. For instance, decision theory uses formal 'decision analysis' (see Howard, 1988), heuristics, or computer-based computational software called a 'decision support system' (e.g. Payne, 2000) that aim to help users make decisions. Probability theories tend to use Bayes' theorem to decide upon which course of action has the best probabilistic outcome. These probability-driven theories aim to determine the best course of action by employing, for instance, Bayes' theorem (see Bayes and Price, 1763), and can take the form of Bayesian networks or Bayesian inference to determine the probability of relationships between *A* and *B*, or determine the probability of *A* given *B*, *C*, and *D*. Normative theories use a number of axiomatic rules to aid optimal decision-making—where "optimal" is usually used synonymously with "rational" decision-making—because the mathematical systems that employ axiomatic rules are understood to be able to provide the most reliable methods of optimal or rational decision-making.

However, as most people know, many decisions are irrational or contravene the outcomes that the mathematical or computational models would recommend. In other words, the optimal decisions that a normative decision theory would advocate as the optimal or most rational course or action is rarely the decision that people make. Many decisions are irrational or illogical, and so some philosophers, psychologists, and mathematicians focus

on *descriptive* decision theory which examines how decisions *are* made. These descriptive theories look at empirical evidence and analyse particular cases of decision-making to determine how people actually make decisions. These descriptive theories are then sometimes compared to normative theories to determine the extent to which the way people make a decision deviates from what a particular formal or computational model would propose as the optimal decision. This is sometimes referred to in terms of the extent to which a normative theory can be *falsified* in relation to the evidence of a descriptive theory. Hansson (1994: 7–8), for instance, describes three different types of falsification:

(1) 'A decision theory is *falsified as a descriptive theory* if a decision problem can be found in which most human subjects perform in contradiction to the theory.'

(2) 'A decision theory is *weakly falsified as a normative theory* if a decision problem can be found in which an agent can perform in contradiction with the theory without being irrational.'

(3) 'A decision theory is *strictly falsified as a normative theory* if a decision problem can be found in which an agent who performs in accordance with the theory cannot be a rational agent.'

There are a number of formal systems that could be used in decision-making, both Bayesian and non-Bayesian. One might, for example, base a decision to intervene on decision or utility matrices (see Bradley, 2017), info-gap decision theory (see Ben-Haim, 2006), fuzzy logic (see Klir, 1995), Markov decision processes (see White, 1993), partially observable Markov decision processes (see Littman, 2009), etc. What I will focus on in this section is decision-making that uses probability. This is because discussions of "probability" or "likelihood" permeate dialogues in which people descriptively discuss decision-making procedures ("Judging by his nasty cough, John probably has an infection"; "Considering

that Billy is carrying a gun, it is likely that he will harm someone"), and also because probability is, in many ways, indicative of the most rational basis for acting ("Since John has a cough and probably has an infection, he should see a Doctor"; "Since Billy is carrying a gun and will likely harm someone, we should call the Police").

Probability theory is used in many formal decision-making processes, an early form of which was popularised by the correspondence between Blaise Pascal and Pierre de Fermat who discussed games of chance (specifically in relation to parlour games). What is sometimes referred to as the "classical theory of probability" uses the formula $P(s) = f/n$ to determine the probability (P) of an event occurring or a statement being true (s), where f is the number of favourable outcomes and n is the number of possible outcomes. For instance, the probability of rolling an even number using a fair six-sided die is 1/2:

$$P(s) = f/n$$

$$P(\text{even number}) = \frac{\text{the even numbers on the dice}}{\text{the six possible numbers on the dice}}$$

$$P(\text{even number}) = 3/6 = 1/2$$

This theory is extremely useful when determining the probability of an event or a statement being true *if one is indifferent to the possible outcomes*. The classical theory relies on the *principle of indifference*, which states that the possibilities are equally probable if, and only if, one does not favour one possibility over another. In other words, the classical theory can only calculate the probability of rolling an even number if the die being used is a fair die and cannot calculate the probability of rolling an even number if the die has been weighted or tampered with[58]. There is a bigger problem, however, and that is that the classical theory

[58] For a good criticism of the principle of indifference as a method of critiquing the classical view, see Bertrand's Paradox discussed in van Fraassen (1998: 303).

relies on all the possible outcomes being identifiable; yet in most practical examples, this is both impracticable and unachievable. Howard-Snyder and colleagues (2009: 546) give the example of a man with lung cancer: a 50-year-old man with lung cancer wants to know the probability that he will die in the next 10 years. The problem comes when trying to identify all the possible ways the man might die. He might die from lung cancer, or he might also die from an 'automobile accident, heart attack, murder, shark attack, nuclear war, and so on'. The problems with trying to use the classical theory in such situations are: (1) that it is impossible to identify every possible outcome (required by the formula); (2) this case is not a simple "numbers game", rather the context of the case needs considering (i.e. is the man in a position to die from shark attack? Probably not if he does not live near or visit the coast); and (3) it is not the case that the possible outcomes are equally probable (the probability that the man will die from a shark attack is less than the probability that he will die from an automobile accident). Howard-Snyder and colleagues' example also highlights broader problems with using classical theory in moral situations or moral sequences. Wanting to know the probability of an agent harming another agent is another case in which the problems outlined in (1), (2), and (3) also stand. For instance, if a Deliberator were trying to ascertain whether to intervene to prevent harm befalling a Victim, this Deliberator would not be able to identify every possible outcome; there is a vast and likely uncountable number of actions that could lead to a Victim either being harmed or not harmed. Is there an alternative method that employs probability that can be used in moral sequences?

Another theory of probability, the relative frequency theory, usually associated with John Venn, runs into similar problems. This theory uses the formula $P(s) = f_\circ/n_\circ$ where $f_\circ$ represents the total favourable outcomes and $n_\circ$ represents the total observable outcomes. If one observed a die being rolled 1000 times, and the number 3 was rolled 167 times, then,

using the relative frequency theory, the probability of 3 being rolled would be 167/1000 (roughly 1/6). The two problems with this theory are (1) *the problem of the single case*, and (2) *the reference-class problem*. To explain the problem of the single case, Howard-Snyder and colleagues (2009: 548) use a case of a thousand-sided die. Using the classical theory, the probability of rolling 9 is 1/1000, but using the relative frequency theory one encounters two problems. First, because there is, in all likelihood, no actual thousand-sided die, there would be no probability of rolling 9; but this just seems wrong because 'there is *some* chance that a thousand-sided die would come up 9 when rolled, even if no rolls actually take place' (Howard-Snyder et al., 2009: 548). The point here is that even though one might never construct a thousand-sided die one would, rightly I think (although without wanting to enter into a debate on Platonic Forms), feel compelled to argue that there *is* still a probability that a 9 could be rolled using a thousand-sided die. This case therefore emphasises the gap between using *pure* probability (classical theory) and *veridical* or *empirical-based* probability (relative frequency theory).

This leads on to the second problem outlined by Howard-Snyder and colleagues, namely that *even if* 'a deranged gambler' could and were to construct a thousand-sided die, which he then rolled once and only once, would this provide a reliable probability of 9 being rolled? In this scenario, the deranged gambler could either roll a 9 or not roll a 9 during his singular roll of the die. If a 9 were rolled then, using the relative frequency theory, the probability of rolling a 9 would be 1/1000, and so using the relative frequency formula the probability would be 1/1; and if a 9 were not rolled then the probability of rolling a 9 would be 0/1000 or 0/1 when input into the relative frequency formula. However, both of these outcomes are unfavourable, as this does not account for (a) the possibility of another number besides 9 being rolled even though 9 was rolled in the singular instance, and (b) the

possibility that 9 could be rolled even though it was not during the singular instance. Therefore, neither (a) or (b) are accounted for in this "all-or-nothing" conception of singular-case probability.

The reference-class problem, on the other hand, is a general worry also present in the classical theory, namely a problem of context identification. To elaborate, the problem is that in making calculations of probability, both the classical theory and relative frequency theory are used in *general* rather than *individual* cases. The example of the man with lung cancer, for example, asks for the probability of *any* 50-year-old man with lung cancer dying within 10 years, rather than an individual case of John who has been suffering from lung cancer for 5 years, has already undergone extensive chemotherapy which has not helped, suffers from depression as a result of his wife dying, and also suffers from osteoporosis as a result of the chemotherapy. The problem here is that unlike the case of rolling a die 1000 times to see how many times a 3 is rolled, it is problematic to say that one must observe a John-like man a certain number of times before a decision can be made about the probability of John dying in 10 years. In other words, the problem here is that, as well as John suffering from lung cancer, he also suffers from depression and osteoporosis. This means that John belongs to a number of classes that: (a) need to be exactly duplicated in other observed cases that are used to determine the probability of John dying in 10 years if calculations can be made using the relative frequency theory; and (b) pose problems for the question being asked, namely, what effect, if any, does the fact John suffers from depression have on the probability of him dying in 10 years? The duplication problem of (a) and co-class problem of (b) make the reference-class problem one of the biggest obstacles to being able to use relative frequency theory in moral sequences. In the case of determining the probability that harm will occur, a Deliberator must safeguard against generalising in as much as it is

imperative that he focus on the individual case, which itself involves accounting for both an uncountable number of possible outcomes (i.e. Gunman shooting the ground instead of shooting the would-be Victim) and a number of co-class factors (i.e. Gunman is holding a gun, but he is also a known pacifist and has no prior violent criminal convictions).

Both the classical theory and relative frequency theory discussed above are *objectivist* theories about probability, in as much as both theories believe that by assigning a probability value to an event (*s* in the formulas), one is saying something objective about the event itself. In the classical theory, the probability of a six-sided die rolling an even number is 1/2, whereas in the relative frequency theory, depending on the number of favourable outcomes relative to the observed outcomes (say, rolling a six-sided die 1,000 times, 167 times of which results in rolling a 3), the probability of rolling a 3 would be roughly 1/6. The only difference is that in the classical theory the probability of *s* is determined by the relation of *s* to other possibilities, whereas in the relative frequency theory the probability of *s* is determined by its relation to empirical studies in which *s* is tested. But what about *subjectivist* theories of probability?

Subjectivist theories of probability, unlike objectivist theories of probability, hold that probability is a difference in, or rather a degree of, belief[59]. That is, if one were to proclaim

---

[59] It is worth noting that just because the subjectivist believes that probability is a degree of individual belief and is thus subjective, this does not mean that one can or should hold a degree of belief that is irrational. One is always subject to the rules of probability calculus, i.e. the restricted disjunction rule, general disjunction rule, restricted conjunction rule, general conjunction rule, negation rule, etc.) For a good overview and explanation of the rules of probability calculus, see Howard-Snyder et al. (2009: 552–561); and for a good explanation of why it is not the case that "anything goes" in the subjectivist theory, and why even the subjectivist must safeguard against irrational beliefs, see the "Dutch book argument" discussed by F. P. Ramsey (1926) and Bruno de Finetti (1937).

that the probability of rolling an even number using a fair six-sided die was 1/2, or that the probability of rolling a 3 was 1/6, then one would be expressing something, a belief, about oneself. One *believes* that one has a fifty per cent chance of rolling evens using a six-sided die, and one *believes* that one has a one-in-six chance of rolling a 3. This idea can be expressed by subjectivists about probability using the formula $P_n(s) = x/(x + y)$ where $P_n(s)$ is the probability (*P*) that person *n* assigns to an event or the truth of a statement *s*, and where *n* gives *x*-to-*y* odds on *s* being true. For example, Howard-Snyder and colleagues (2009: 549–550) give the example of one's friend giving 3-to-1 odds that the next card drawn from a fair 52-card deck will not be a spade. One's friend thinks that it is 3 times more likely that one will draw a non-spade, given that 3 out of the 4 suits (seventy-five per cent) are non-spade and only 1 out of the 4 suits (twenty-five per cent) is a spade. Thus, one's friend believes that it is 3 times more likely that a non-spade will be drawn than a spade and so offers 3-to-1 odds that a spade will be drawn, knowing that he will likely receive £1 rather than having to fork-out £3. Put another way, $P_n(spade\ will\ not\ be\ drawn) = 75/(75 + 25) = 3/4 = 0.75$ indicates that it is probable that the next card will be a non-spade, hence why one's friend can be confident in offering 3-to-1 odds that the next card will be a spade (he knows that there is only a twenty-five per cent chance that a spade will be drawn). There are, however, numerous problems with the subjectivist theory, two of which concern the objectively probabilistic nature of science, and the need to assign numerical values to the variables in the formula. First, the laws of nature are objectively probabilistic in so far as, to consider the example discussed by Howard-Snyder and colleagues (2009: 550–551), the half-life of carbon-14 is 5730 years. Although the laws of radioactive decay are indeterministic, that is, the amount of time it takes for specific samples of carbon-14 to decay vary, it takes *on average* 5730 years for carbon-14 to half-decay. This is unlike deterministic laws of nature, such as gravity, where

it is one hundred per cent certain that an apple falling from a tree will, *ceteris paribus*, hit the ground. As far as science is concerned, then, laws of nature, whether deterministic or indeterministic, are objective rather than subjective. A larger problem, however, is that even a subjectivist theory of probability cannot help account for decisions in moral sequences because the formula to be used relies on assigning numerical values to *x* and *y*. But how can one assign a numerical value to, or calculate the odds of, an agent harming another agent? And what information should one use to calculate the probability of harm? In order to circumvent the problems outlined in the two objectivist theories discussed (classical theory and relative frequency theory) and the general subjectivist theory, many philosophers turn to Bayesianism in an attempt to discover if and how probability can be employed in decision-making.

At this juncture, it is important to make a distinction between decisions that are made in certainty and those that are made in non-certainty. Decision-making under certainty involves the types of decisions where my decision to do *x* is based on the certainty of a 'state of nature' (*viz.* a factor for consideration). For example, suppose I am a meteorologist and I *know* that it will rain. In this circumstance, my decision to take an umbrella to work is simple and is an example of decision-making under certainty because I (claim to) know for *certain* that it will rain. But what if I am uncertain as to whether or not it will rain? How confident can a rookie meteorologist or layperson be that he should take an umbrella to work? This problem is exacerbated in moral decisions, which arguably always involve a level of uncertainty. Duncan Luce and Howard Raiffa (1957: 13) explain certainty and the two types of non-certainty (risk and uncertainty):

‘We shall say that we are in the realm of decision making under:

(a) Certainty if each action is known to lead invariably to a specific outcome (the words prospect, stimulus, alternative, etc., are also used).

(b) Risk if each action leads to one of a set of possible specific outcomes, each outcome occurring with a known probability. The probabilities are assumed to be known to the decision maker. For example, an action might lead to this risky outcome: a reward of $10 if a 'fair' coin comes up heads, and a loss of $5 if it comes up tails. Of course, certainty is a degenerate case of risk where the probabilities are 0 and 1.

(c) Uncertainty if either action or both has as its consequence a set of possible specific outcomes, but where the probabilities of these outcomes are completely unknown or are not even meaningful’[60].

Although many decision-makers are faced with a situation somewhere in-between risk and uncertainty, these categories provide accurate examples of the types of situations in which a decision-maker has to make a decision. The discussion so far has been focussed on decision-making under certainty, where the priors are known, can be observed, or can be calculated. But what of decision-making under risk or uncertainty? This is where Bayesian

---

[60] A stricter version of “uncertainty” is sometimes identified as “ignorance”. Hansson (1994: 28) describes decision-making under uncertainty as those made with ‘partial probabilistic knowledge’, and decision-making under ignorance as those made with ‘no probabilistic knowledge’. (He further differentiates these from the ‘deterministic knowledge’ that is indicative of decision-making under certainty, and the ‘complete probabilistic knowledge’ that comprises decision-making under risk.)

probability and Bayesian decision theories (BDT) enter the picture; here, probability is used to gauge the trade-off between different decisions and the expected value or expected utility[61] of a certain action, where the relevant information (be it a numerical input or otherwise) can be unknown, and can further be used to measure the risk or cost of deciding *x* over *y*.

Richard Bradley (2007: 233) describes BDT as 'formal theories of rational agency: they aim to tell us both what the properties of a rational state of mind are [...] and what action it is rational for an agent to perform, given the state of mind [...]'. One might therefore decide to use BDT for decision-making, which involves a complex formal process based on Bayes' theorem, and is usually employed because of its rigorous computational procedures. BDT are predominantly used as a statistical approach to pattern classification that determine probability, but which also helps one calculate the cost of error, such as a doctor determining whether it is likely that his patient has a common cold or a viral infection. BDT considers, for example, different 'states of nature' and determines their *a priori* (prior) probability, from which 'Bayes' formula' can be used to determine *a posteriori* probability (i.e. probability given the evidence) using prior and conditional information; it also uses a 'Bayes Decision Rule' (i.e. a probabilistic decision rule) which can itself be used to calculate risk (i.e. 'expected loss' or 'conditional risk') also known as 'Bayes Risk'. The mechanics of

---

[61] Paul Schoemaker (1982: 529) describes expected utility as 'the major paradigm in decision making since the Second World War. It has been used prescriptively in management science (especially decision analysis), predictively in finance and economics, descriptively by psychologists, and has played a central role in theories of measurable utility'. For a good history and explanation of expected utility see Schoemaker (1982), and see Peterson (2009) for an overview of how this 'more precise version' compares to the closely related concept of expected *value*. A more technical introduction can be found in Resnik (1987: chapter 4).

BDT is not relevant to this discussion, all that is important is that one understands that BDT are formal decision-making processes that can be used to determine the probability of *x* based on one's beliefs about the prior probability of the 'states of nature' (of which *x* is an instance), which also involves, amongst others considerations, a calculation of cost, error, or risk[62]. For simplicity, and to avoid unnecessary calculations that are only relevant to pattern classifications, I will focus on Bayes' theorem instead of the (much more complex) BDT.

Bayes' theorem can be expressed as[63]:

$$P(h/e) = \frac{P(h) \times P(e/h)}{[P(h) \times P(e/h)] + [P(\neg h) \times P(e/\neg h)]}$$

Here, *P* stands for probability, *h* is the hypothesis, and *e* is the evidence for that hypothesis[64]. (Upper-case instances of the lower-case *e* and *h* should be taken to denote specific statements.) To determine the extent to which the given evidence supports a particular hypothesis using Bayes' theorem, one must know: $P(h)$, the prior probability of the hypothesis before any supporting evidence *e* is taken into account (from which, once established, $P(\neg h)$ can be determined using the negation rule of probability calculus (see

---

[62] For a good introduction to Bayesianism, see Bernardo and Smith (2000); for a comprehensive account of Bayesian analysis, including BDT, see Berger (1985: chapter 4); and for a thorough explanation of BDT, its procedures, and application see Aksoy (2014).

[63] For the proof of Bayes' theorem, see Howard-Snyder et al. (2009: 564). This proof, and the theorem provided, is adapted by Howard-Snyder and colleagues from Skyrms (1986: 153). Other, more traditional, versions of the theorem (i.e. Peterson, 2009) are expressed as $p(B|A) = \frac{p(B) \bullet p(A|B)}{[p(B) \bullet p(A|B)] + [p(\neg B) \bullet p(A|\neg B)]}$ given that $p(A) \neq 0$.

[64] It is worth noting that one could further distinguish between what Howard-Snyder and colleagues (2009: 565) call 'background evidence' (*b*) from 'the "new" evidence or phenomenon to be explained' (*e*) and, if one wished to make this distinction, Bayes' theorem could be amended to

$$P[h/(e \bullet b)] = \frac{P(h/b) \times P[e/(h \bullet b)]}{\{P(h/b) \times P[e/(h \bullet b)]\} + \{P(\neg h/b) \times P[e/(\neg h \bullet b)]\}}$$

For simplicity, I consider both *b* and *e* to constitute the *e* used in the in-text theorem.

Howard-Snyder et al. (2009: 555)); $P(e/h)$, the probability that the evidence for *e* exists, assuming that h is true; and $P(e/\neg h)$, the probability that the evidence for e exists, assuming that h is false. Assuming that one decides upon a course of action based on the probability of the hypothesis being true, how useful is Bayes' theorem in decision-making? Consider a moral decision that I have based on the example given by Howard-Snyder and colleagues (2009: 565–566). A Deliberator, a Police Officer (*PO*), thinks that an agent (*A*) will harm another agent. This agent can either harm another agent (h), or not harm them ($\neg h$)—but, due to the law of non-contradiction, he cannot do both (he cannot do both $h \bullet \neg h$). Let us assume that *PO* is learned and has many years of experience in law enforcement, and that because of this he knows that about thirty per cent of people in *A*'s position harm, and the other seventy per cent do not harm. Let us further assume that *PO* has looked at other cases similar to *A*'s situation and has found that ninety per cent of cases in which harm occurs involve a gun, but only ten per cent of cases where harm does not occur involves a gun. What is the probability that *A* will harm given that he is carrying a gun? We therefore want to find $P(H/E)$ which is the probability that the hypothesis "*A* will harm" given the evidence that "*A* is carrying a gun". (Here, *H* stands for "*A* will harm" and *E* stands for "*A* is carrying a gun"). From the information above, one can work out:

$$P(H) = 30/100 = 3/10$$

$$P(\neg H) = 1 - 3/10 = 7/10$$

$$P(E/H) = 90/100 = 9/10$$

$$P(E/\neg H) = 10/100 = 1/10$$

One can then use this information and input the relevant data into Bayes' theorem:

$$P(H/E) = \frac{3/10 \times 9/10}{[3/10 \times 9/10] + [7/10 \times 1/10]} = \frac{27/100}{27/100 + 7/100}$$

$$= \frac{27}{34}$$

As 27/34 is roughly equal to 0.79, this makes the probability that *A* will harm given that he is carrying a gun about 79%. So, based on the fact that there is a seventy-nine per cent chance that *A* will harm another agent, it seems reasonable to say that *PO* would be justified in intervening to prevent *A* from harming. But just how reliable is this method of decision-making? Even if one assumes that the statistics used to determine the probability of harm given that a gun is present are correct (one might assume that *PO* has used a database of national statistics of offences over the past five decades, such as the Police National Computer (PNC), for example), there remains a crucially subjective element, namely *PO*'s knowledge that about thirty per cent of criminals in *A*'s situation harm. This is, however, completely subjective and would differ according to whether the Deliberator was privy to the relevant knowledge and experience. How might a Rookie Police Office (*RPO*) make such a decision, if he could not assign a numerical value to the probability of *A* harming? Or what if there were no national database that could be used to determine the probability of *A* harming given the evidence-based information? In other words, does a lack of assignable numerical value(s) render Bayes' theorem irrelevant or impractical for decision-making?

There is a method by which one could determine the probability of *A* harming without assigning the numerical values needed for $P(h)$, $P(\neg h)$, $P(e/h)$, and $P(e/\neg h)$, all which are required for Bayes' theorem, by assigning *relative values* instead of *numerical values*. Despite lacking the relevant belief, knowledge, or experience (*priors*), *RPO* might believe

that $P(h) \geq P(\neg h)$, that is, the probability that harm will occur is equal to or greater than the probability that harm will not occur, and could further argue that $P(e/h) > P(e/\neg h)$, that is, the probability that the evidence that *A* is holding a gun supports the hypothesis that harm will occur is greater that the probability that the evidence that *A* is holding a gun does not support the hypothesis that harm will occur. This would then allow *RPO* to conclude that $P(h/e) > P(\neg h/e)$. To emphasise the idea that Bayes' theorem can be applied without assigning numerical values, suppose that $P(h) = P(\neg h) = 1/2$, $P(e/h) = 4/5$, and $P(e/\neg h) = 1/5$, then

$$P(h/e) = \frac{1/2 \times 4/5}{[1/2 \times 4/5] + [1/2 \times 1/5]} = \frac{4/10}{4/10 + 1/10} = 4/5$$

$$= 0.8$$

$$P(\neg h/e) = 1/5 = 0.2$$

One could therefore legitimately say that: if $P(h) \geq P(\neg h)$ and $P(e/h) > P(e/\neg h)$ then $P(h/e) > P(\neg h/e)$, assuming that one has good reason to suppose that $P(h) \geq P(\neg h)$ and $P(e/h) > P(e/\neg h)$. In other words, so long as one has good reason to believe that the probability that harm will occur is equal to or greater than the probability that harm will not occur, and has good reason to believe that the probability that the evidence that *A* is holding a gun supports the hypothesis that harm will occur is greater that the probability that the evidence that *A* is holding a gun does not support the hypothesis that harm will occur, then one can legitimately generalise that the probability that the evidence supports the hypothesis that harm will occur is greater than the probability that, given the same evidence, harm will not occur. But how useful is the assigning of relative values for practical cases? That is, what problems are there with probability-orientated decision-making processes, especially subjectivist theories, such as those that use Bayes' theorem, which may involve unknown priors? Martin Peterson (2009: chapter 6.4) suggests that one might be able to either form

or update one's beliefs (on which the priors are based) using relevant evidence; but what if no evidence were available?

It is at this point that a Markovian understanding of probability could prove helpful. Andrey Markov has come to be known in many research areas for his work in stochastic processes[65] and for what have been called Markov chains (see Norris, 1998). A Markov chain is one type of stochastic process used in systems in which there are a number of states linked in a chain of probabilistic transitions, and where the current state determines only the next state, and where the states are completely observable (*viz.* the states are known). Markov chains are usually discrete-time Markov chains (see Everitt, 2002)—although there are other types, such as continuous-time Markov chains (Parzen, 1962)—and has a Markov property (*viz.* the stochastic process has what is called a 'memoryless property' (see Feller, 1971). For a good overview of discrete-time Markov chains, see Tijms (2003: chapter 3); and for an example of such a chain, see Tijms (2003: 82–84) example of 'the drunkard's random walk'. A closely related model, the Hidden Markov Model (see Fraser, 2009), shares the same features as a Markov chain but with one essential difference: the states are not completely observable (*viz.* the states are not known). Essentially, both Markov chains and Hidden Markov Models are not under the control of a decision maker. What is pertinent to the moral

[65] The term "stochtastic", derived from the Greek word στόχος meaning "aim", is used to refer to processes or systems that contain random (or unknown) variables and are thus difficult to predict.

decision-making process are Markov Decision Processes (MDPs) [66] and Partially Observable Markov Decision Processes (POMDPs)[67] (Kaelbling, Littman, and Cassandra, 1998), both of which are under the control of a decision maker. Although both MDPs and POMDPs rely on decisions being controllable by a decision-maker (unlike Markov chains and Hidden Markov Models), what separates a MDP from a POMDP is that a decision-maker in the latter can only partially observe, and is unable to completely observe, the states. POMDPs therefore incorporate a key element to moral decision-making, namely decision-making in situations where the states are only partially observable. Coelho, da Rocha Costa, and Trigo (2014), for instance, have written an exceptional paper on how 'the POMDP seems to exhibit some well suited structure to describe a (flavour of) morality' and how the POMDP can be employed in the moral decision-making process. The intricacies of POMDP are too vast to do justice to here and would be a tangential discussion to the aims of this chapter and thesis[68]; so, for current purposes, I ask that the reader understand only that Markovian systems, particularly POMDPs, are unique probabilistic systems that can be employed in a wide range of situations (depending on whether states are completely

---

[66] A MDP is a 5-tuple $(S, A, P.(\cdot,\cdot), R.(\cdot,\cdot), \gamma)$ where: S is a set of states; A is a set of actions (and where A_s is a finite set of actions from state s); $P_a(s, s') = \Pr(s_{t+1} = s' \mid s_t = s, a_t = a)$ is the probability that at time t action a will, in state s, result in state s' at time $t + 1$; $R_a(s, s')$ is the reward once the transition from s to s' has occurred; and $\gamma \in [0,1]$ is the discount factor (*viz.* the disparity between current and future rewards). There are a number of variants, but I offer only the more general version presented above. To elaborate, in a MDP the process is always in some state (s) at a certain time (t) and a decision-maker must act (a) in s, after which the process transitions to a new state (s') depending on a and the state transition function $(P_a(s, s'))$, which then presents the decision-maker with a reward $(R_a(s, s'))$.

[67] A POMDP is a 7-tuple $(S, A, T, R, \Omega, O, \gamma)$ where: S is a set of states; A is a set of actions; T is a set of conditional transitional probabilities between states; $R: S \times A \to \mathbb{R}$ is the reward function; Ω is a set of observations; O is a set of conditional observation probabilities; and $\gamma \in [0,1]$ is the discount factor. To elaborate, the environment (world) is in a particular state $(s \in S)$ and when the decision-maker acts $(a \in A)$ he triggers a transition to state s' with probability $T(s' \mid s, a)$ during which he makes an observation $(o \in \Omega)$ depending on $O(o \mid s', a)$, and after which he receives a reward $(R(s, a))$.

[68] For a technical overview of MDPs, including POMDPs, see White (1993).

observable, partially observable, unobservable, and so on), including moral decision-making.

3.2.1. THE IMPRACTICALITY OF FORMAL MODELS

Formal decision-making systems that are based on the theory of probability, and even other formal non-probabilistic systems, are indeed useful for *certain* practical applications (e.g. medical decisions, economic strategies, etc.), and are generally useful for establishing and producing demonstrable proofs that can be used to give credence to the idea that a decision either *would* be the best or most rational course of action for a projected/possible situation, or *was* the most rational course of action as decided in a *post-hoc* justification of a particular action. However, formal systems of decision-making have few—if any—practical *moral* applications, in as much as: (1) they cannot inform decision-making *during* the moral sequence, in so far as (1a) formal systems (both probabilistic and non-probabilistic) are too complex and time-consuming to be considered a realistic moral decision-making process, and (1b) require an element of reflection that is impractical in many moral situations; (2) they require some form of numerical or relative value (in the case of probabilistic theories) that boils down to arbitrary or uninformed values; and (3) they do not permit the inclusion of emotions in decision-making.

What started out as a strong investigation into formal systems of decision-making at the start of the 20th Century has arguably, in recent years, started to trail-off. With the advancement of robotics and artificial intelligence has come an overwhelming hurdle in the comprehension of computational decision-making. At present, morally-relevant decision-making has not found its way into robotics. But why is this? If decision-making were just a

matter of formal computation of a formal system of decision-making, it seems strange that research in robotics has not been able to find a way for robots to use these formal procedures to make decisions. Even some of the best examples of modern artificially intelligent systems are incapable of, for example, recognising the written word 'chair' and then pointing to the corresponding object in the room, let alone deciding whether to steal food, save one person's life over another, or intervene to prevent a potential gunman shooting another person. And *even if* such mechanisms or computational procedures could be put in place, it is still unclear whether a robot could make the "right" choice. Perhaps the best anecdotal example of this is a scene in Alex Proyas' film *I, Robot* (2004) where an artificially intelligent robot (an NS4), when faced with the dilemma of saving Detective Del Spooner or a twelve-year-old girl Sarah, both whom are trapped in separate cars that have crashed and are sinking in a river, 'calculated that I [Del] had a 45% chance of survival. Sarah only had an 11% chance […]' and so opts to save Del. But, as Del exclaims, 'That was somebody's baby. 11% is more than enough. A human being would've known that'. The NS4 robot appears to have used a probability-driven decision-making procedure to decide to save Del over Sarah, but, as Del explains, many people would usually have opted to save Sarah. What is it, then, that divides what is arguably the most optimal or rational decision to save Del from the practically-relevant and intuitive decision to save Sarah? More than anything, it boils down to an emotional reaction to a particular situation that was not, and arguably might not be able to be, programmed into robotic decision-makers. There therefore appears to be a missing element in decision-making, *the role of emotion*, that needs to be investigated; a discussion of this will take place in the next section.

This is not to say practical decisions that are not based on formal decision-making processes are in some way based solely on intuitive or emotionally-influenced actions. On the

contrary, there does seem to be a basic logical, or rather rational, process to decision-making. Similar calculations might underpin the way we cognise and process decisions, but we certainly do not conceive of decision-making processes in such ways. These models imply, and their advocates appear to be of the opinion that, 'Most decisions are not [purely] reactive. They take time, and it is therefore natural to divide them into phases or stages' (Hansson, 1994). This is perhaps true of deciding one's life goals, or—to use two of Hansson's examples—calculating whether one should buy this or that house, or whether one should take an umbrella that day, but I do not think (a) that the same is necessarily true of moral decisions, and (b) this fits with how moral decisions are descriptively cognised.

(a). Think back to the last moral decision you made. However you decided on your course of action, I doubt that you thought it 'natural' to divide this decision into phases or steps. I also doubt that your decision was not, in some way, a reaction to the situation at hand; that is, I doubt that your decision was solely based on a rational or reasoned course of action. What is more likely is that your decision was either momentary and devoid of any recognisable decision-making process (purely reactive) or, if calculated, was at least partly decided by an appeal to factors other than a rational, reasoned, or cost-benefit analysis, and likely consisted of an emotional response to the situation (reactive-calculated). It is hard to discover a first-personal moral decision-making situation where one simply appeals to a formal decision-making system or process. (Arguably, third-person moral decision-making might be different, as either spatio-temporal and/or emotional distance from the moral situation might permit the decision-maker to reign-in any non-formal or emotional factors.) Take, for instance, the *I, Robot* scenario described above. The reason, it seems, that Del thought it unimaginable that a person would have chosen to save him over Sarah is because a person would have had an emotional response to the situation, whatever that might consist

of (perhaps stemming from empathy for the child, or imagining if Sarah was one's own child); and this emotional reaction to the situation was simply unavailable to the formal process- and probability-driven NS4 robot. Emotional reactions therefore seem to be a natural part of a moral decision-making process. But does this necessarily discount formal processes being employed in such decisions?

(b). There is another side to the coin, namely that calculating decisions using formal systems or processes is highly impractical in many moral situations, such as those where a purely reactive moral decision must be made (i.e. situations where time-constraints require a momentary rather than calculated decision to be made). Consider, again, the *I, Robot* scenario. The probability-driven formal decision made by the NS4 to save Del rather than Sarah must have involved at least a basic utilisation of probability theory. For the NS4 which, I assume, had advanced computational resources at its disposal, even the most complex calculations of probability, risk, etc. would have been easy given its processing power; but this processing capacity is not necessarily something that is available to human decision-makers. Granted, humans have a complex neural system and the ability to process data in a relatively computational manner, and there are some who can calculate extremely complex mathematical calculations both on paper and "in their head", so to speak; but even these mathematical geniuses would be hard-pushed to make the relevant calculations in a situation where a momentary and time-sensitive decision must be made. Therefore, even if the normative decision theorists argue that the most reliable methods of decision-making are those that are based on formal systems or processes, this seems to be in direct conflict with the empirical evidence that some—if not most—moral decisions cannot involve any—or, at a push, involve minimal—formal processing.

However, as I have previously mentioned, there does seem to be a basic logical or rational process to decision-making; that is, the mechanics of formal systems might be used in the sub-conscious processing of decisions unbeknown to the decision-maker. Jean Lave, Michael Murtaugh, and Olivia de la Rochga (1984) investigated the use of mathematics whilst shopping for groceries and found that situational context is important for recognising one's cognitive processes. The participants of her study were able to do the maths, i.e. simultaneous equations, needed for comparing prices of products whilst shopping, but could not necessarily do the same mathematical equations when asked to do so[69]. This, Lave claims, illustrates the problems of employing mathematical processes in practical situations. Lave found that 'rather than learning by replicating the performance of others or by acquiring knowledge transmitted in instruction, […] learning occurs through centripetal participation in the learning curriculum of the ambient community' (Lave and Wenger, 1991: 100). In other words, the reason the participants could employ the maths needed in practical systems but not necessarily do the same formal mathematical equations is because we learn how to *use* the necessary processes, rather than learning the mathematical processes or formulation of the equations *per se*. Therefore, regardless of the method of processing data that might go on sub-consciously or experientially, it is certainly not the way we necessarily perceive our decision-making process. This study demonstrates how calculations can be made sub-consciously; if this is possible in the case of simultaneous equations, it is not entirely implausible that the same sub-conscious calculations can be Bayesian in nature. This would of course require separate studies that are beyond the scope of this thesis. But even if one would not concede that this study demonstrates that complex formal calculations can be made sub-consciously, we can take something else from this

[69] For an overview of the study, findings, and its implications, see Lave (1988: chapter 3).

study: we might make sub-conscious calculations, but the way we make decisions seems, at least empirically, more intuitive than that.

What we are left with, then, is the idea that emotional reactions are a natural part of a moral decision-making process, but this does not necessarily discount formal processes being employed in such decisions. However, these formal processes are, to a large extent, impractical in many moral situations. Arguably, the reason that formal systems or processes are (normatively) favoured over emotional reactions to a situation, or are (descriptively) seen to function in this way, are because we tend to look for *reasons* for acting, and we believe the best way to track and ensure decisions are rational or reasonable is to lay them out formally. This, however, is not needed. The next section will argue that reactive-calculated moral decision-making (a) is common place, (b) can provide accurate results, (c) is just as effective at ensuring decisions are rational (the idea that non-formal theories would open the decision-maker up to importing their own beliefs, motives, etc. will be addressed at this point), and (d) is both reliable and the most practical method for everyday moral decision-making.

3.2.2. THE ROLE OF EMOTIONS IN DECISION-MAKING

> 'The traditional view of rational decision making is that agents should keep emotions out and let reason guide them towards the best available response by engaging in a cost-benefit analysis for each possible response. Deliberation takes considerable time and effort (in terms of attention, working memory, and computation), but allows agents to choose the best option in a rational way' (Bortolotti, 2015: 87).

The traditional view is that decision-making is largely governed by reason, which is tied-up with formal decision-making processes, such as using Bayesian probability, BDT, etc. However, in moral situations and moral decision-making there is, I will argue, also a reactive emotional element that informs one's decisions. If this is true, and emotion does play a role in moral decision-making, does this undermine how rational, reasoned, or "right" a moral decision can be?

There are a number of examples where emotions can be seen to inform, bias, and even govern decision-making; this is illustrated in the work of Alice Isen and Robert Patrick (1983), and Rajagopal Raghunathan and Michel Tuan Pham (1999). Isen and Patrick (1983: 194) found that 'subjects who had reason to be feeling elated bet more than control subjects on a low-risk bet, but wagered less than controls on a high-risk bet', and that 'elated subjects were more daring than controls on a "long shot"'; Raghunathan and Tuan Pham (1999: 56) found that 'sad individuals are biased in favor of high-risk/high-reward options, whereas anxious individuals are biased in favor of low-risk/low-reward options'. The same can be said of one's attitude towards a moral decision, as illustrated by Jonathan Haidt's (2001) description of how 'most people' react to the story of two siblings, Julie and Mark, who,

taking the appropriate precautions and birth-control, 'decide that it would be interesting and fun if they tried making love'. Haidt argues that people maintain that it is wrong for Julie and Mark to be sexually involved with each other even though it is clear that no inbreeding will occur, that no harm will befall either sibling, etc. When further pressed for reasons for judging the sexual relationship to be wrong, Haidt (2001: 814) argues, 'one becomes a lawyer trying to build a case rather than a judge searching for the truth. One puts forth argument after argument, never wavering in the conviction that Julie and Mark were wrong'; in the end, it comes down to a person saying '"I don't know, I can't explain it, I just know it's wrong"'. What this indicates, according to Haidt, is that moral judgements can be governed by an emotional reaction to the moral situation one is considering, and this emotional reaction cannot only be descriptively inaccessible to that person, but it is often directed by the social institution of which that person is a part. This indicates that emotions have the ability to influence decision-making on some level and provides the foundations for a deeper investigation into emotion-based, or rather emotionally-informed, decision-making.

Antonio Damasio (2006) discusses the role of emotions in decision-making; he describes the case of Elliot who underwent surgery to remove a meningioma (a brain tumour), along with frontal lobe tissue that had been damaged by the tumour. After the surgery it became clear that 'Elliot was no longer Elliot' (Damasio, 2006: 36). Even though 'he was still physically capable and most of his mental capacities were intact', his personality changed dramatically and, perhaps most devastatingly of all, 'his ability to reach decisions was impaired', which itself impacted on most aspects of his life (Damasio, 2006: 37). His inability to concentrate, failure to manage his time, and constantly 'losing sight of his main goal' saw Elliot unable to be trusted by his colleagues, family, and friends; 'flawed business

and financial decisions' saw Elliot lose his job; and his 'foolish' decisions saw his first marriage break down, then his second (Damasio, 2006: 36–37). This personal destruction was identified as being caused by damage to his prefrontal lobe. Damasio (2006: 38) summarises:

> 'The tragedy of this otherwise healthy and intelligent man was that he was neither stupid nor ignorant, and yet he acted often as if he were. The machinery for his decision making was so flawed that he could no longer be an effective social being. In spite of being confronted with the disastrous results of his decisions, he did not learn from his mistakes. He seemed beyond redemption […]'.

As time progressed, Damasio (2006: 51) realised that Elliot's decision-making problems were linked to the 'reduction in emotional reactivity and feeling' demonstrated by Elliot. Damasio (2006: 45) describes how he 'never saw a tinge of emotion […] no sadness, no impatience, no frustration […] He tended not to display anger, and on the rare occasions when he did, the outburst was swift; in no time he would be his usual new self, calm and without grudges'. When asked to recount negative life events, Elliot reflected without any seeming emotional response to these events:

> ‘[T]he magnitude of his distance was unusual. Elliot was exerting no restraint whatsoever on his affect. He was calm. He was relaxed. […] He was not inhibiting the expression of internal emotional resonance or hushing inner turmoil. He simply did not have any turmoil to hush. […] In some curious, unwittingly protective way, he was not pained by his tragedy’ (Damasio, 2006: 44).

Astounded, Damasio and colleagues asked Elliot to perform five tasks, each centred on imagined moral or financial dilemmas/questions, which were designed to determine the connection between his emotional disengagement and the problems of his decision-making (Damasio, 2006: 46–48). Now, according to the traditional view that “good” decisions are made by rational deliberation, the fact that Elliot, and those with similar brain injuries had an emotional disengagement with past, current, and imagined situations implies that they would be excellent decision-makers. However, Elliot seemed to struggle to make even the most basic of real-life decisions yet performed well in the laboratory tasks. Damasio therefore hypothesised that ‘Not all actions commanded by a brain are caused by deliberation. On the contrary, it is a fair assumption that most so-called brain-caused actions being taken at this very moment in the world are not deliberated at all’ (Damasio, 2006: 89).

Instead, we place *somatic markers* governed by an emotional reaction that aids decision-making[70].

Damasio's (2006: 165–222) *somatic-marker hypothesis* provides an explanation of the role emotions play in decision-making, and can explain why Elliot was a bad decision-maker despite being able to provide reasoned and justifiable responses to all five laboratory tasks that required him to make decisions on a variety of imagined situations. Damasio and colleagues soon realised that the reason Elliot could make reasoned laboratory decisions, but failed miserably at making real-life decisions, was because 'the ongoing, open-ended, uncertain evolution of real-life situations was missing from the laboratory tasks' (Damasio, 2006: 50); this led Damasio to the conclusion that Elliot's defective decision-making ability was due to Elliot's inability to settle on a decision, in as much as he was perpetually stuck in the motion of a probabilistic or cost-benefit analysis of decision-making. Somatic markers are placed before performing any probabilistic calculation or cost-benefit analysis and are placed based on a feeling one has—itself based on a particular emotion[71]—towards a situation. More specifically, when a 'bad outcome connected with a given response comes into mind, however fleetingly, you experience an unpleasant gut feeling', which 'forces

---

[70] Dunn, Dalgleish, and Lawrence (2006) succinctly explain the concept of somatic markers; they state that 'emotion is the representation and regulation of the complex array of homeostatic changes that occur in different levels of the brain and body in given situations. When making decisions, a crude biasing signal (a somatic marker) arising from the periphery or the central representation of the periphery indicates our emotional reaction to a response option. For every response option contemplated, a somatic state is generated, including sensations from the viscera, internal milieu, and the skeletal and smooth muscles [...] These somatic markers serve as an indicator of the value of what is represented and also as a booster signal for continued working memory and attention [...] Particularly in situations of complexity and uncertainty, these marker signals help to reduce the problem space to a tractable size by marking response options with an 'emotional' signal. Only those options that are marked as promising are processed in a full, cognitive fashion'.

[71] See Damasio (2006: 127–164) for an explanation of the difference and link between feeling and emotion.

attention on the negative outcome to which a given action may lead' (Damasio, 2006: 173). What this does, in effect, is provide that person with a warning, a "gut feeling", that this course of action will likely result in a negative outcome. This warning signal then compels that person to immediately reject that course of action, which then allows that person to focus on other factors when deciding on a course of action. Any formal systems of decision-making therefore come *after*, *and in light of*, reacting to these feelings-based somatic markers. However, Elliot's emotional dysfunction rendered him incapable of placing somatic markers. This explains his indecisiveness; he was continually overwhelmed by formal decision-making processes due to a lack of "gut feelings" narrowing the options or courses of action open for consideration.

Damasio also argues that somatic markers *assist* decision-making, but they 'do not deliberate for us' (Damasio, 2006: 174). In other words, decisions cannot be made on feeling alone; one's feelings instruct a somatic marker to be placed, this is then used to reduce the number of options open for consideration (somatic markers for negative outcomes are withdrawn from consideration), and this underpins the course of action open to the formal decision-making process. But where does this leave purely reactive moral decisions that do not involve, or at least do not consciously involve, post-somatic marker formal processing or deliberation?

I think Damasio's somatic-marker hypothesis is correct in how it describes the "gut feelings" many of us seem to have and act on, but I do not agree that somatic markers cannot be used without formal decision-making processes. I concede that the twinning of somatic markers and formal decision-making can yield reasonable and justifiable decisions. But it seems strange that once the somatic markers are in place, and in situations where there is a

short time-frame for decision-making, these somatic markers cannot solely be used to inform one's decision. Decisions that are purely reactive are arguably made based on these somatic markers alone, although I think that such decisions are open to being irrational, ill-considered, or unjustified. It seems reasonable to suggest that the only way to ensure that the rationality of decisions—characteristic of formal decision-making processes—is accounted for and upheld in decision-making is to ensure that formal processes are used in some capacity during the decision-making procedure. For this reason, I think that purely reactive moral decisions are prone to erroneous reasoning, can result in irrational or unjustified decisions being made, and should be avoided.

Turning attention to moral decisions in which the formal decision-making process is done *in advance* of a situation presenting itself, one might imagine somatic markers being used to decide pre-considered, pre-planned, and calculated-reactive moral decisions. In deciding how to act in a general or particular moral situation the factors for consideration are narrowed by the use of somatic markers. Thus, it seems reasonable to claim that decisions that are either calculated (pre-considered or pre-planned) or involve a calculated element (calculated-reactive) implement somatic markers during the decision-making process, and this is entirely consistent with the requirement of such decisions to use formal decision-making processes. In pre-considered and pre-planned moral decisions, somatic markers are used to narrow the number of options available for consideration during the formal decision-making process. In calculated-reactive moral decisions, the calculated element works in the same way but with an added reactive emotional element post-calculated decision which, as an addendum to the formal decision-making process that has been informed by somatic markers, is *updated* by the feelings one has towards the situation. Importantly, this can then be used to form new somatic markers.

But what about moral decisions in which the formal decision-making process is done *during* a particular moral situation? The somatic-marker hypothesis appears to best-fit reactive-calculated moral decisions, where the somatic markers aid a formal decision-making process that happens *during* the situation. I believe this is the best method of moral decision-making. This kind of moral decision involves one having a "gut feeling" about a particular moral situation that is based on, whether knowingly or sub-consciously, having an emotional reaction to that situation. This consequently narrows the options that one considers during the formal decision-making process.

Emotions, or rather feelings, are therefore not, as the traditional view of decision-making would have us believe, detrimental to good decision-making; if anything, their use in forming somatic markers positively impacts on the decision-making process. Decisions are tailored to that person in as much as the somatic markers that govern the narrowing and selection of options for consideration during the formal decision-making process reflect something about the decision-maker. These, combined with somatic markers speeding-up the whole decision-making procedure, therefore render reactive-calculated moral decisions normatively favourable above the other four kinds of moral decisions.

3.3. CONCLUDING REMARKS

In this chapter, I identified five different kinds of moral decisions and found that reactive-calculated moral decisions yield a normatively reliable process on which to base a decision to intervene in a moral sequence. I argued that formal systems/theories of decision-making are impractical for moral decision-making, in so far as they require too much time to calculate, which does not adhere to how we descriptively make decisions. It became clear

that a lack of formal decision-making does not render a decision unreasonable or irrational; emotions, properly applied to the decision-making process in the form of somatic markers, can not only provide equally reliable decisions, but also accounts for a Deliberator's own feelings towards a particular situation and, perhaps more importantly, enables quick decisions to be made. From this, I claimed that the most practical way of making moral decisions is to employ Damasio's somatic-marker hypothesis in the broader framework of formal decision-making processes.

Building on my claim that reactive-calculated moral decisions are normatively favourable above the other four kinds of moral decisions and provides the best moral decision-making process, the next chapter will discuss a particular reactive-calculated moral decision, namely deciding if and when to intervene to prevent harm, to establish when a decision to intervene is justifiable. During this analysis, I will argue for a basic (reactive-calculated) probability-driven formal decision-making process that is underpinned by a reactive element linked, first and foremost, to the progression of the *primary narrative* of the sequence (*viz.* how the sequence-events unfold) and, if available, an assessment of the *secondary narrative* of the same sequence (*viz.* one's prior knowledge of the beliefs, dispositions, etc. of the Initiator that the Deliberator is evaluating)[72]. These primary and secondary narratives underpin the *tertiary narrative*, which holds that there are three philosophical considerations relevant to deciding when it is justifiable to intervene that must be satisfied for an intervention to be justifiable. When all three narratives are combined and the *threshold of harm* is passed, a Deliberator is, I will argue, justified in becoming an Intervener.

---

[72] A detailed explanation of the distinction between primary and secondary narratives will be featured in §4.2. to §4.5.

CHAPTER 4:

INTERVENING IN A MORAL SEQUENCE

We all, at some point in our lives, have to come to terms with deciding if we should intervene in a situation, and if we should, when it is most appropriate or justifiable. This may take the form of deciding if and when to take an alcoholic or drug addict to rehabilitation in an attempt to save them from their addiction, or might involve deciding if and when to tell your friend's spouse that your friend has been having an affair to prevent that spouse from enduring any more betrayal.

This chapter will add to the system of moral sequencing introduced in chapter 2 by assessing if, when, and in what ways one can justifiably intervene in a moral sequence. I will argue that an intervention is only justifiable when a sequence has passed the threshold of (physical) harm, gauged by assessing the primary, secondary, and tertiary narratives of a moral sequence.

4.1. THE TRADE-OFF: RESPECTING RIGHTS AND AUTONOMY

Before continuing, let us consider why a Deliberator should bother intervening[73]. What is at stake—both in terms of intervening and in terms of not intervening (forbearing)? In other words, what are the potential moral costs of intervening and what are the potential moral costs of forbearing to prevent harm?

---

[73] I thank Jonathan Parry and Patrick Tomlin for pressing me to elaborate on this issue.

The concept of intervention relies on understanding the importance of negative rights, namely a Victim's right to not be harmed. Warren Quinn's (2003) view of negative rights can be used to better understand the importance of self-governance for morally proper actions. Quinn (2003: 370) describes negative rights (in contrast to positive rights which 'are claim rights to aid or support') as 'claim rights against harmful intervention, interference, assault, aggression, etc.'. Quinn (2003: 372) argues that there has to be a 'precedence of negative rights', otherwise someone's body—or 'one might say his person'—'is not in any interesting moral sense *his*'. If one's negative rights were not enshrined in morality, then it would, under morality's own permissions, be morally acceptable for one to use another person's body as one sees fit. The problem, in Fiona Woollard's terms, is that '[i]f morality does not include such a constraint [against violating another's negative rights], then it treats the victim's body and mind as common property' (Woollard, 2015: 106). In order to prevent another from harvesting my organs, for example, it is essential that negative rights precede any other rights, including positive rights (e.g. the right to interfere). Without the enshrinement of negative rights, one simply cannot be said to be an agent that is in control of one's own life. So if, I think it can be agreed, bodily and mental self-governance should be enshrined, then negative rights *must* precede other rights, including the right to choose. Therefore, if an agent's (Initiator's) actions will lead to the violation of another agent's (Victim's) negative rights, then it *prima facie* seems that another agent (Intervener) can justifiably intervene to prevent Initiator's actions impacting on Victim's right to non-interference. The issue is that there can sometimes be a conflict between respecting an agent's negative right to not be harmed and respecting another kind of *prima facie* right, namely the right to autonomous action—that is, an agent's right to act in the way they see fit. So what happens in a case where an agent *A* wants to and does act in a way that threatens harm to another agent *B*? *A*'s seeming right to act in the way they

see fit (here threatening harm to *B*) conflicts with *B*'s negative right to not be harmed. The solution is simple—whilst an agent has an unrestricted negative right to not be harmed, the right to autonomous action (if indeed it is a right at all and not some other less binding, perhaps personal, prerogative) must be restricted. Without wanting to enter a debate on whether and the extent to which an agent should have the right, ability, or what have you to act in the way they wish without restriction, I think it is fairly uncontentious to claim that rights surrounding autonomy are restricted (after all, the government can restrict a prisoner's autonomy, but (at least in the UK) cannot restrict their negative right by harming them). It therefore seems sensible to assert that an agent has the right to decide how to act, but on the proviso that others' negative rights should outweigh that agent's right to choose. In other words, although intervening might cost an Initiator part of their autonomy, this is a small price to pay for protecting an agent's negative rights. Indeed, in cases where an Initiator decides to initiate a moral sequence that will cause death to a Victim, the cost of limiting Initiator's actions is even more justified. This claim can be supported by looking to the Precautionary Principle, which represents the notion that *in dubio pro natura* ("if in doubt, decide in favour of the environment"). Although the principle is often discussed in relation to environmental law and policy, the central idea is that of favouring precaution or demonstrating due caution in areas of irreversible damage, and so has been applied to other areas too. It can be interpreted for our purposes as the idea that in cases where there are irrevocable harms, one must err on the side of caution. As Colin Klein (2017: 1) puts it, 'Nuclear war, catastrophic climate change, or species extinction are not merely very bad things: they can't be wound back even at great cost. That's why it seems right to forgo goods, even widely acknowledged goods, in order to avoid a reasonable chance that they occur'. The same applies to death. Death is irreversible (clinical intervention

notwithstanding), and so we can partly jettison an Initiator's right or otherwise to act how they wish to ensure that irrevocable harm does not occur to a Victim.

We can also look back to the philosophical literature to buttress the claim that negative rights should take precedence. Intervening is essentially a case of other-defence—an Intervener intervenes to avert, mitigate, or otherwise diminish a threat of harm to Victim. Intervener is therefore defending Victim from harm. It is a simple intuitive claim that defending a Victim in such a way is *prima facie* morally required. It is this intuition that might lead us to 'conceive of other-defence as a duty rather than a mere permission' (Frowe, 2011: 26), and such line of thought tracks early philosophers' (e.g. St Ambrose's) beliefs that defensive force can be 'actually demanded on moral grounds' (Swift, 1970: 535). We might even say that a Deliberator has a duty to intervene. Linking back to and reiterating what I said in §1.2.4., according to Foot (1984: 181), 'there are rights to non-interference [negative rights] […] and there are also rights to goods or services [positive rights]'. Doing harm typically involves violating another's right to non-interference, and this right is more morally binding than allowing harm, which typically only involves violating another's right to goods or services (i.e. the 'organ transplant procedures' (Harris, 1975: 81)). Simply put, agent *A* has a *negative duty* to not harm agent *B* and agent *C* has a *positive duty* to not allow agent *B* to be harmed.

Section 4.6.2. will develop the discussion of the trade-offs of intervening by discuss the concepts of costly and costless interventions, and harmful and harmless interventions, and from that a more nuanced and robust understanding of the various costs of intervening in various ways will emerge. Moreover, the discussion in chapter 5 will speak to the issue of

the moral costs of intervening and not intervening more directly. With that said, let us now focus on establishing the threshold of harm.

4.2. ESTABLISHING THE THRESHOLD OF HARM: THE THREE NARRATIVES

How should we define and understand the threshold of harm? If, as I propose, the threshold of harm defines the moment or point at which an intervention becomes justifiable, we must ensure that a philosophically rigorous discussion is sought. What follows is such a discussion. In order to present one such way of understanding this threshold, the remainder of this chapter will discuss (a) the importance of prioritising how a moral sequence unfolds, (b) supplementing this narrative with available or known information pertaining to intra-sequence agents, and (c) some philosophically relevant conceptual issues. Although the reader might disagree that this is the *best* way of conceptualising the threshold of harm, my aim here is not to enter into debate about various threshold models, nor is it my intention to compare, contrast, and evaluate these competing models to favour the victor; such an endeavour would likely still be met with philosophically rich criticism by some due to the contentious, interlinked, and morally difficult conceptual issues that must take centre stage in any discussions of such a threshold. I offer one model, a philosophical and reasoned contender, of such a threshold of harm that can make sense of some challenging cases/examples and can accommodate and make sense of the importance of narratives and somatic markers in making the reactive-calculated moral decision to decide if and when to intervene in a moral sequence to prevent an Initiator from harming a Victim.

We can conceive of the threshold of harm as lying somewhere on a continuum between the initiation of the moral sequence and the projected/anticipated harm to a Victim. The threshold of harm is lowered (or is lower) under certain conditions and if certain criteria are met, and, likewise, the threshold is higher (that is, the bar for the threshold is raised) if certain important criteria are not met.

In §4.3. I will outline the primary narrative and will explain how this tracks the unfolding of sequence-events in a way that enables a Deliberator to predict the likelihood of harm eventuating based on what is happening in the moral sequence (and further based on a prediction of how that moral sequence might further unfold). Section 4.4. will outline the secondary narrative and will explain that a Deliberator can use various epistemically available information to inform and update the probability of harm eventuating. These two sections culminate in a discussion in §4.5. on how the primary and secondary narratives can be joined to ascertain a probability that harm will eventuate to a Victim without an intervention and that this can be used in the threshold of harm. In §4.6. I will outline the tertiary narrative and argue that a number of philosophically relevant considerations are required to form a philosophically informed threshold of harm; these considerations include the necessity of an intervention (including the necessity to act), the proportionality of an intervention to the threatened harm, and the liability of certain agents to be harmed in an intervention.

However, while these criteria will help us ascertain whether we can consider the threshold of harm to have been met in a particular moral sequence, they do not help us conceptualise more functional considerations pertaining to our understanding of the nature of the threshold of harm. So how might we conceptualise it? We might consider whether the threshold of

harm is a fixed, objective, and agent-neutral threshold whereby there exists an optimal moment during a moral sequence where, beyond this point, a Deliberator would be justified in intervening to prevent harm to a Victim. However, it would be a mistake to conceive of the threshold of harm in this way. To say that there is an optimal, fixed, and/or objective point in every moral sequence at or beyond which an intervention would be justified would be to ignore the complexities of real-life moral sequences, discount the fact that each Deliberator (in addition to the primary narrative) brings with them certain (or perhaps no) secondary narrative which would affect their assessment, and, importantly, totally discount the intricacies of relevant philosophical topics, including: special/equal concern, claim to a resource, liability, necessity, and proportionality. The threshold of harm should therefore be understood as being subjective, variable, and flexible in that it is determined by and from the perspective of the Deliberator—my reasons for holding this view of the threshold of harm will become clearer throughout this chapter.

We must next clarify the role of the primary and secondary narratives in ascertaining the threshold of harm—or, more precisely, their role in establishing the probability threshold in the threshold of harm.

4.3. THE PRIMARY NARRATIVE: HOW THE MORAL SEQUENCE UNFOLDS

What is the primary narrative and why do we need it? Assessing the actual events of the sequence (primary narrative) provides vital evidence for establishing the threshold of harm; in other words, an assessment of the probability of harm eventuating must take into account the moral sequence itself. To explain the primary narrative and its importance, consider the legal case of *R v Moloney* (1985). After a night of drinking, the defendant was challenged

by his stepfather to see who could load, draw, and shoot the fastest. During the challenge, the defendant shot and killed his stepfather. The defence argued that the defendant did not intend to kill his stepfather, and his death was a horrible accident. In this case, the sequence of events (drinking alcohol, followed by an aggressive discussion about the defendant wanting to leave the army of which the stepfather disproved, followed by instigating an act that involved instruments that had the potential to seriously harm) proceeded in such a way that it would have been obvious to an external observer (Deliberator) that, after the guns were handled by intoxicated agents, harm would be the probable sequence-outcome; the moral sequence was progressing in such a way that it would have been clear to a Deliberator fairly early-on that it was probable that harm would eventuate. This case highlights the importance of the primary narrative, where this narrative simply tracks and explains the actual sequence-events in a moral sequence (see §2.2.3. for a definition of a sequence-event). Whether or not the defendant intended to harm[74], for example, is on its own irrelevant to the question of whether intervention is justified. The fact that the defendant engaged in a particular sequence of events that would have led (or rather did lead) to harm would have provided a partial justification to intervene in the sequence (i.e. take away the guns or put the two drunks to bed). The defendant drank alcohol, he aggravated his stepfather by discussing the army, and he partook in an activity that had the potential to

---

[74] To understand the relevance of intention to this case, a distinction needs to be drawn between the legal terms *direct intent* and *oblique intent*. The former is used to refer to when an agent acts such that he intends there to be a particular and foreseeable outcome to his actions. The latter is used to refer to when an agent acts such that the consequence is 'virtually certain' given his actions, and he knows that by acting in this way the harmful consequence was 'virtually certain'. It is for this reason that the judge believed it necessary to direct the jury on *oblique intent*, and the defendant was convicted by the jury of murder (based on the notion that the defendant obliquely intended the death or serious harming of his stepfather). The defendant took his case to the Court of Appeal, and after it was rejected he appealed to the House of Lords, who overturned the conviction of murder and amended his conviction to manslaughter. They determined that it was not a case of oblique intent.

cause serious harm under those circumstances. The defendant's actions provide vital evidence to form the basis of a decision to intervene (although, as we shall see in §4.4. and §4.5., this is not a sufficient condition for a justifiable intervention).

Even though assessing the actual sequence-events should be the primary factor for consideration, ascertaining intent—or specifically direct or oblique intent—can of course be highly useful in ascertaining a decision to intervene and forms part of the secondary narrative. In the case of Jones (introduced in the Introduction), for instance, intentions *can* be used in a Deliberator's arsenal—along with character, dispositions, etc.—to decide if and when to intervene, but crucially this is a *secondary* narrative (discussed in the next section) to that which the Deliberator can extract from the *primary* narrative of the sequence (i.e. how the sequence actually unfolds, and the actions of the agent within that moral sequence). Intention is therefore a red herring in as much as one should not assume that this is the information one must extract from, and use to primarily assess, a moral sequence. The primary narrative, *viz.* the sequence-events and projected sequence-outcome, provide the foundations for an assessment of intervention, which can be supplemented with the secondary narrative of the Initiator's intentions, personality, etc. To focus on the secondary narrative diverts the Deliberator's attention away from the actual sequence (primary narrative) and towards a psychological assessment of the Initiator. Although this secondary psychological assessment can indeed be used to assess the possibility of a harmful sequence-outcome, without principally assessing the actual sequence-events and the trajectory of these events towards a possible harmful sequence-outcome the Deliberator runs the risk of: (a) intervening on conjecture and an analysis of the mental state of the Initiator, such as evaluating his actions or supposed reasons for acting based on an assessment of his past actions, determining his cognitive abilities, or psychoanalysing his past and current mental

states (including sub-conscious material, etc.), rather than evaluating the most fundamental and crucial facts that can be ascertained from the primary narrative; and (b) assessing intentions *over* the actual sequence-events. Practically speaking, defending an intervention based on a secondary narrative ("I think he intended to kill, so I stopped him") is also extremely difficult to justify and prove. This is, arguably, the point of contention in many legal battles in as much as the prosecution asserts that the defendant intended to harm and so was detained and/or is prosecuted on that basis, whilst the defence argues that there was no intention to harm. For instance, in the case of *R v Grant* (2014), the defendant tried to appeal against his conviction for attempted murder and two counts of causing grievous bodily harm with intent. The defendant (G), along with two other members of his gang (M and K), shot at and chased a member of a rival gang (B) into a shop. B was unhurt, but one of G's bullets hit and paralysed a five-year-old boy, and another bullet became lodged in the head of a customer. This is a clear example of how, in decisions of intervention, intentions have no weight of their own and should not be relied on without reference to other considerations (especially the primary narrative). G was acting in such a way that harm was likely to (and did) result from his actions. In other words, the fact that G did not intend to harm those that were harmed would not itself curtail the justifiability of an intervention (had, for example, the Police been present).

Further consider the example of Gunman used in the sequence archetypes outlined in §2.3.1. and §2.3.2. Gunman initiates a sequence by initiating a NPET, namely taking the gun out of the holster. The threat of being shot did not exist until Gunman acted in such a way as to make the non-harmful, holstered gun potentially harmful. (There are, after all, very few ways, if any, a holstered gun can cause harm to another agent.) Follow this moral sequence one event further and the potential harm of Gunman increases. When Gunman points the

gun at Victim, the sequence-outcome seems to tip in the favour of the likelihood of harm being done to Victim. That is to say that there become fewer future sequence-events that can occur in which harm will not likely be the sequence-outcome, given the preceding sequence-events. In addition, there are few rational reasons for this action other than wanting to harm, or at least threaten, Victim (although Gunman might, for instance, want to merely show Victim his new weapon, but this would certainly be unusual behaviour). This is not to assume that Gunman must be acting rationally; he might be acting out of a variety of delusional beliefs. But to focus on and evaluate agent rationality or agent intentions would be to place emphasis in the wrong place. What one assesses is, in fact, the *primary narrative* of the sequence; Gunman's actions are assessed objectively and without reference to Gunman's rationality or delusions, and without reference to an overarching understanding of rationality. Importantly, the focus of the assessment is not on why one *would* act in such a way if harm were not their intention, but simply on how the moral sequence *is* unfolding. With this in mind, as the sequence-events progress from the first sequence-event, it becomes increasingly difficult to provide an explanation for how the gun could be unholstered without harm to Victim becoming the sequence-outcome, to the extent that when the gun is pointed at Victim, there are very few, *if any*, legitimate explanations for how the moral sequence could end without harm being done to another agent. This procedure reflects the gathering of evidence that informs the formal decision-making process that underpins reactive-calculated moral decisions. The process is probabilistic in nature in so far as a Deliberator calculates the probability of the harm occurring to a Victim by assessing how the moral sequence is progressing; in other words, a Deliberator gathers evidence from the primary narrative, from the how the moral sequence has unfolded and how the sequence might further unfold.

But where does this leave the reactive element employed in reactive-calculated moral decisions? Well, firstly, a Deliberator's emotional reaction to the primary narrative can cause the Deliberator to experience the relevant "gut feelings", but this can also be (and arguably is mostly) influenced and guided by other factors that form part of the *secondary narrative*.

4.4. THE SECONDARY NARRATIVE: THE EPISTEMIC IMPORT AND BOUNDARIES OF A DELIBERATOR

As the previous section has already mentioned, the secondary narrative supplements the primary narrative with epistemic factors that include, but are not limited to: knowledge of intra-sequence agents (including the desires, dispositions, personality, etc. of those agents[75]); situational knowledge, including what is known about the circumstance of the moral sequence and happenings/events prior to the moral sequence, etc.; contextual information, including the social, cultural, and political context of the moral sequence; and expert knowledge of the Deliberator, e.g. the level of expert knowledge pertinent to a particular moral sequence, including the Deliberator's past experiences and prior involvements. All of these factors, individual or jointly, enable the Deliberator to treat this information as available evidence (which might thus update any Bayesian-like prior or prompt "gut feelings").

---

[75] This can include a Deliberator's knowledge about their own desires, dispositions, etc. that might influence or impact on how they perceive, approach, or respond to the moral sequence and other intra-sequence agents.

The secondary narrative therefore has both epistemic import (in that a Deliberator's knowledge about various moral sequence relevant factors, including knowledge about intra-sequence agents, are utilised to inform and update the evidence available) and defines a Deliberator's epistemic boundaries (in that it helps to limit the evidence that the Deliberator can and will use in their decision-making process). But how, exactly, do the primary and secondary narratives work together?

4.5. COMPARING AND JOINING THE PRIMARY AND SECONDARY NARRATIVES: THE ROLE AND LIMITATIONS OF THE PROBABILITY OF HARM

There is plentiful literature on the role of the probability or likelihood of harm eventuating. For instance, some authors argue that for defensive killing (e.g. Intervener or Victim killing Initiator to defend Victim against Initiator's threat of harm) to be (subjectively) justifiable it must be more likely than not that Victim is liable to be killed (e.g. Haque, 2012), whereas other authors argue that an agent (e.g. Initiator) is (objectively) liable to be killed if the defensive killing of that agent by a defender (e.g. Intervener or Victim) is made on that agent's justified belief, gauged by the passing of a probability threshold, that the agent to whom defensive harm is inflicted is objectively liable to be harmed (e.g. Ferzan, 2005; Frowe, 2010; McMahan, 2011). Other authors take a less strict view of the role of probability, with some authors arguing that an agent forfeits their right to not be killed simply if it is likely that that agent is liable to be harmed (e.g. Zimmerman, 2008). Lazar (2018: 864) also draws attention to other instances of the use of a probability threshold in the philosophical literature—citing Michael Huemer (2010) and Frank Jackson and Michael Smith (2006; 2015) as attributing a threshold view in their discussions of absolutist moral

theories[76]—and in criminal law—citing Walen's (2015) defence of the concept of 'beyond reasonable doubt' and the USA's (under the Obama administration) admitting to killing in war in those cases where there is 'near certainty that the terrorist target is present' and 'near certainty that non-combatants will not be injured or killed' (Office, 2013). Although such authors highlight the importance of accounting for the probability of harm, they fall short of being able to account for the role of probability in a threshold of harm. So what role does the probability of harm eventuating play in the threshold of harm?

Importantly, calculating the probability of harm occurring to a Victim from the primary narrative is not achieved at the expense of assessing the motivations of the Initiator. These are the sorts of situation-specific considerations that prompt the utilisation of somatic markers in the use of the decision-making process. A Police Officer, for example, might know from past experience that he has previously not intervened in similar situations, but which later transpired to result in harm. However, by first and foremost assessing the agent rather than the sequence, one must be vigilant against getting weighed-down in an assessment of the agent's character, dispositions, beliefs, etc. (secondary narrative) rather than focussing on an assessment of the primary narrative of the sequence. If, however, this secondary information (i.e. the agent's dispositions) is available to the Deliberator of the sequence, then this information should certainly be taken into account, but not at the expense of the primary narrative of the sequence. To do so would be to de-contextualise the decision-making procedure—it would involve making a determination about intervening without any reference to the actual moral sequence itself.

---

[76] However, Hawley (2008) and Aboodi, Borer, and Enoch (2008) have argued against this.

The process by which the Deliberator makes a reactive-calculated moral decision to intervene is directed by the Deliberator's emotional reaction to the particular situation in as much as his somatic markers prompt him to have a "gut feeling" about a particular course of action, which then narrows down the number of options he considers in his formal assessment of the primary narrative. His "gut feelings" can be governed by both his reaction to the primary narrative (i.e. his feelings about the sequence itself) and his reaction to the secondary narrative if available (i.e. his feelings about what he knows about the agent thought to be on course to harm Victim). The speed at which this decision is made is constrained by the available evidence (whether both a primary and secondary narrative are available). The more information available to a Deliberator, the more somatic markers will be utilised (he will perhaps have a greater "gut feeling"), which instructs the narrowing of options to be considered in the processing of the information and licences a quick and reasoned decision. Building on the discussion in chapter 3, what should now be clear is that emotions are not only compatible with rational decision-making, but also positively impact on the speed of decision-making, a crucial component of moral decision-making.

The primary and secondary narratives therefore jointly help a Deliberator to include as part of their decision-making procedure factors relevant to both the situation as it objectively unfolds and an array of pertinent information related to various relevant intra-sequence agents and their knowledge broadly construed that will help to determine the likelihood of harm occurring to a Victim. These two narratives therefore aid a Deliberator in identifying and updating the evidence available in complex formal calculations that are made sub-consciously, for instance, as part of a Bayesian-like formal decision-making procedure; knowledge of how the moral sequence has unfolded and is unfolding, knowledge about various intra-sequence agents, and other situationally and contextually relevant information

known by the Deliberator can all help to update the evidence available (what are often called 'priors' in Bayesian calculations) which can then be used to update the probability of harm occurring which, in turn, informs, produces, and/or updates the Deliberator's somatic markers that guide their decision-making.

Importantly, whilst the probability of harm is undoubtedly an important factor for consideration in establishing a threshold of harm, it should not take centre stage—that is to say that a determination of the probability of harm to Victim eventuating does not solely ground the placement of the threshold of harm in a moral sequence and therefore cannot be solely used to determine whether an intervention is or was justifiable. Whilst we might agree that harm should *ceteris paribus* be probable (or some other expression) for an intervention to be deemed justifiable, such line of reasoning encounters numerous issues. For instance, although we might think that, *ceteris paribus*, it would be wrong to intervene to prevent a threat from occurring if there was a low probability (say, 10%) of harm eventuating, and although we might further say that, *ceteris paribus*, a higher probability (say, over 50% or perhaps certainly over 95%) of harm eventuating would generally warrant an intervention to prevent that harm, the following problematic examples arise[77]. These examples serve to highlight the issues inherent in solely looking to probability to determine the threshold of harm—we shall see that although the probability of harm occurring to a Victim (assessed via the primary and secondary narratives) is certainly an important threshold factor, we need

[77] I am thankful to Jonathan Parry and Patrick Tomlin for providing some of the counter examples that I will shortly outline and for pushing me to more clearly conceptualise and explain the threshold of harm that I outline in this chapter.

to further supplement this information with a tertiary narrative to avoid many of the inherent issues when the concept of probability of harm is applied to moral decision-making[78].

Boulder

Deliberator can enact a costless intervention to stop a boulder (which is rolling by natural events) that has a 10% chance of killing Victim.

The probability of harm to Victim eventuating is minimal and the intervention to prevent this harm is costless (*viz.* it is a non-agent-affecting intervention, see §4.6.2. for a delineation of the different kinds of interventions). Intuitively, we would likely want to say that Deliberator is justified in intervening since there is some threat to Victim that can be averted without any cost to any agent. Boulder highlights how it would seem counter-intuitive to say that there must be an optimal probability expressed as a percentage only beyond which an intervention would be justifiable. For instance, if we were to say that part of the threshold of harm relies on ascertaining whether it is more likely than not (*viz*. that there is more than a 50% chance) that harm to Victim will eventuate without an intervention, and this "more likely than not rule" sets the threshold beyond which an intervention is justifiable, then intervening in Boulder would be unjustifiable. However, this seems intuitively wrong, since we would likely not feel obliged in this circumstance to wait until harm to Victim is probable, more likely than not to occur, or what have you; in other words, it seems that we should not require that a set probability threshold be passed before intervening. Indeed, such a probability threshold, if we were to impose one, would likely require a greater than 10%

---

[78] Please note the discussion that follows will employ terminology that I do not explain until later sections. Understanding this terminology is not necessarily required at this stage to understand the drive of my claims; I have simply used this terminology here to ensure retrospective consistency.

probability of harm, which would therefore preclude an intervention in Boulder. In short, it is difficult to impose a philosophical requirement that a set probability threshold be met in this circumstance—Deliberator would, for reasons discussed later in this chapter, be justified in intervening in Boulder. We therefore need a threshold of harm that can account for justifiably intervening in such cases of costless intervention that will avert a threat that has at least some chance of causing harm to Victim and without solely relying on establishing or imposing a set probability threshold only beyond which an intervention would be justifiable. But we must also be careful not to throw the baby out with the bathwater—the threshold of harm should, it seems, at least incorporate some element of the probability of harm occurring. After all, if there was less than 1% chance of harm occurring to a Victim, it seems that an intervention that would cause harm to an intra-sequence agent would be difficult to justify[79,80], and, equally, if there was a greater than 99% chance that

---

[79] According to the United States of America's National Safety Council's (2019) data on the 'Lifetime odds of death for selected causes, United States, 2017', these are the odds/probabilities that an American will die by the following means: 1 in 243,765 chance (~0.0004% probability) of dying as a railway passenger; 1 in 188,364 chance (~0.0005% probability) of dying as a passenger on an aeroplane; 1 in 115,111 chance (~0.0009% probability) of dying in a dog attack; 1 in 46,562 (~0.002% probability) of dying by hornet, wasp, or bee sting; 1 in 4,047 chance (~0.02% probability) of dying as a bicyclist; 1 in 2,696 chance (~0.04% probability) of dying by choking on food; 1 in 556 chance (~0.2% probability) of dying in a pedestrian incident; and 1 in 103 chance (~1% probability) of dying in a motor vehicle crash. It is arguably these low probabilities of harm that make it highly unlikely that one would, for reasons related to probability alone, decide to stop another person from: travelling on a train; travelling on a plane; owning or interacting with a dog; utilising any outdoor space where hornets, wasps, or bees were in the vicinity; using a bicycle; eating food; walking near roads or otherwise being a pedestrian; or driving or being a passenger in a car.

[80] However, there are of course grey areas which will be discussed later in this section. For instance, according to Cancer Research UK (2018), tobacco smoking caused 19% of all deaths in the UK in 2015 (and caused 27% off all cancer deaths in the UK in 2010 (Peto et. al., 2018)). Tobacco (both active and passive smoking) also has a 3 in 20 chance (15% probability) of causing cancer (Brown et. al., 2018). The question now concerns whether a 19% probability of death as a result of tobacco use is high enough to justify an intervention (say, extinguishing a friend's cigarette). This question will be indirectly addressed in the rest of this section.

harm would occur to a Victim without intervention, it would be difficult to justify not considering intervening (although, as we shall see, in both circumstances there are other relevant factors that might make intervening justifiable/unjustifiable even though there is a low/high chance of harm occurring to a Victim). With this in mind, consider the following:

> *Russian Roulette*
>
> Aggressor has captured Victim and is playing Russian Roulette with a six-chambered gun with one bullet. Deliberator can enact a harmless (but autonomy-affecting) intervention to avert harm to Victim by removing the bullet.

Deliberator can intervene without doing any harm (they can enact a harmless intervention) although doing so will reduce Aggressor's autonomy (it is autonomy-affecting since Aggressor has chosen to initiate the moral sequence). Initially there is a one-in-six chance—roughly a 17% chance—that Victim will be killed when the trigger is first pulled, with the odds decreasing (and the chance of harm increasing) with each trigger-pull. However, although the intervention would be autonomy-affecting (at the expense of Aggressor), the intervention would be harmless (since no agent would be harmed as a result of the intervention). The issues here are: since there is only roughly a 17% chance that harm will occur to Victim at the first trigger-pull, it cannot be said that harm occurring to Victim is likely, more likely than not, beyond reasonable doubt, etc.—like in Boulder, all that can be said is that there is *a* chance (and a low chance at that) that harm will occur. Even if we were therefore to express the threshold of harm as a percentage probability that harm will eventuate to Victim, it is unlikely that we would set such a percentage threshold as low as 10% (as in Boulder) or 17% (as in Russian Roulette), yet we would, I think, want to

intuitively claim that Deliberator would be justified in intervening. What separates the two examples, however—apart from the percentage probability—is that Boulder is a costless intervention whereas Russian Roulette is not (see §4.6.2. for an overview of the different kinds of interventions); Russian Roulette involves infringing on Aggressor's autonomy (since intervening prevents Aggressor from acting in the way they wish). Continuing the discussion in Boulder, the threshold of harm must therefore avoid expressing the threshold of harm as a percentage probability of harm eventuating, and further it must be able to account for those interventions that are costly—that is, interventions that cost something of an intra-sequence agent (here, Aggressor's autonomy, but also other costs, such as harm to other agents).

> *Scratch 1*
>
> Aggressor is attacking Victim with a 40% chance of success. If successful, Victim will die. Deliberator can intervene by scratching Aggressor.

Here, the probability of harm occurring to Victim has increased whilst the threatened harm to Victim remains the same (Victim will die); Deliberator can intervene to avert (the 40% chance of) Victim's death by scratching Aggressor. What we have here is a case of a large threatened harm (of death) to Victim that can be averted by intervening in a way that sees a small harm inflicted on Aggressor. Let us assume that intervening here would be justified. Now, if we want to use probability as the sole threshold of harm, what is required is some sort of account of why a 40% probability chance here justifies an intervention. We could just say that 40% is a reasonable threshold beyond which an intervention would be

justifiable, regardless of the fact that it is clearly not the case that it is more likely than not, nor beyond reasonable doubt, that harm will occur to Victim[81]. However, this is problematic.

Scratch 2

There is a 40% chance that Aggressor will scratch Victim and will succeed, unless Deliberator intervenes by killing Aggressor.

In this case the magnitude of the harms is reversed, so that the magnitude of the threatened harm to Victim is small whereas the magnitude of the harm required by the intervention to avert that harm to Victim is high. The probability of the threatened harm occurring remains the same as the previous example. Intuitively, we would likely say that intervening here is not justified (we might cite reasons that the intervention is clearly disproportionate (liability aside) to the threat), and if this is the case then the 40% threshold cannot be objectively applied in all cases. This illustrates the problem of trying to set an objective, or rather universal, probability threshold beyond which an intervention would be justified. What separates the cases are other salient factors (such as the proportionality of the intervention) which need to be reconciled with an account of the probability of harm in a threshold of harm. Problems with focussing on probability continue:

Scratch 3

Aggressor is trying to scratch Victim and will succeed unless Deliberator intervenes by killing Aggressor.

---

[81] This example and the two that follow also introduce the concept of proportionality which is discussed in §4.6.4.2.

In this variation, there is a 100% chance (or as close as one can be to this certainty) that Victim will be scratched unless Deliberator kills Aggressor. If all we are concerned with is the probability that harm will occur to Victim, and if this solely grounds claims of the justifiability of an intervention to avert that harm, then Deliberator should intervene and kill Aggressor. But this just seems intuitively wrong (for the same reasons that killing Aggressor in Scratch 2 seems intuitively wrong). The unjustifiability of killing Aggressor in both cases is divorced from the probability of harm occurring to Victim[82]. All three Scratch cases highlight the issue that solely relying on the probability of harm to Victim to assess and determine a threshold of harm is philosophically problematic for the reasons outlined—the magnitude of the harms, amongst other factors, also need to be considered.

There are other issues related to focussing on probability, for instance where it is not clear why a certain probability threshold would have to be met for an intervention to be justified. Consider the following two cases:

[82] These other 'tertiary' factors relevant to the threshold of harm will be discussed in §4.6.4.

Wait and See 1

Aggressor is attacking Victim with a likely (say, over 50%) chance of success. Deliberator could kill Aggressor now (and thereby avert the harm to Victim), but there will later be an option to kill Aggressor once it becomes clear (say, with 95% certainty) that Victim will be harmed.

Wait and See 2

Aggressor is attacking Victim with a likely (say, over 50%) chance of success. Deliberator could avert harm to Victim by killing Aggressor now, or Deliberator could wait until later when he will have the chance to avert harm to Victim by breaking Aggressor's leg.

For the purpose of making a claim related to the Wait and See cases, let us assume that the threshold of harm is set at the point at which it becomes more likely than not (*viz.* greater than 50% chance) that harm will occur to Victim. Both Wait and See cases involve a situation in which waiting to see what happens *beyond* an established probability threshold would be beneficial. In Wait and See 1, it seems that waiting to see if the probability of harm to Victim increases before intervening would be beneficial, especially since the intervention would cause the death of Aggressor. If by waiting to see if the probability of harm will increase it becomes clear that Victim will not sustain any harm, then waiting to see and not simply acting at or shortly after a probability threshold is beneficial. This introduces another factor for consideration, namely the necessity of intervening, which will be discussed in §4.6.4.1. Similarly, in Wait and See 2, not acting at or shortly after a probability threshold would be beneficial; although the probability of harm remains constant, the magnitude of the harm required to avert the threatened harm to Victim (that is,

the magnitude of the harm that the intervention would cause) reduces. By waiting, the Deliberator can intervene by breaking Aggressor's leg rather than killing them. In such a case, it seems like waiting to see would be more beneficial than acting at or shortly after a probability threshold has been reached. What both cases illustrate is how the justifiability of an intervention seems to hinge on other factors apart from the probability of harm to Victim.

Indeed, what we want to avoid when establishing a threshold of harm is a situation in which (a) Deliberator is justified in averting harm by any means beyond a probability threshold and without due consideration for whether the probability of harm might subside, and (b) Deliberator is justified in acting now to kill Aggressor rather than waiting to see if the same threat of harm to Victim can be averted by less harmful other-defensive actions (such as breaking Aggressor's leg).

Where does this leave the role of the probability of harm in establishing the threshold of harm in a moral sequence? Should the requirement that the probability of harm be a factor in the threshold of harm be jettisoned? If not, how should it be restricted and in what way will it form part of the threshold? The cases discussed above highlight general concerns with asserting that the threshold of harm is simply a point at which the probability of harm to Victim (be that 25%, 50%, or any other percentage probability) justifies an intervention. As we have seen, there are a number of issues that arise from this line of thought and other seemingly relevant factors are ignored. But this does not mean that the probability of harm plays no role in establishing a threshold of harm. For if we were to remove any percentages provided in the cases just discussed, and if asked whether Deliberator should intervene, one would still likely want to know the probability (or some derivative) of harm occurring to

Victim. Probability does play an important role in the reactive-calculated moral decision of deciding if and when to intervene in a moral sequence, it is just that this consideration (a) needs clarifying and limiting and (b) is not the *only* consideration required. Indeed, as Lazar (2018: 877) puts it: 'Who brings a calculator to a gunfight?'; 'How can we possibly expect anybody facing an actual decision problem—especially when lives are at stake—to whip out their calculator and work out the expected moral utilities?' Although this is an amusing and somewhat facetious remark, Lazar does rightly highlight the issues surrounding (a) and (b), and that we need to look for a plausible explanation of how probability fits into the threshold of harm.

Issues associated to ascertaining a threshold of harm and intervening based on percentages (that is, the probability of harm to Victim eventuating) can be alleviated if we consider that the threshold of harm, and the probability of harm to Victim eventuating, can be ascertained without establishing an objective or universal probability threshold (objective in the sense that it remains the same for every Deliberator in a given moral sequence and universal in the sense that the same probability threshold, that is the same percentage, should remain set across different moral sequences); instead, we can take a relative view of probability. It is important to remember that numerical values do not necessarily have to be attributed in Bayesian calculations (the sort of calculations that, in chapter 3, I argued that we make as the formal part of the decision-making process). To reiterate what I said in §3.2., we can assign relative values instead of numerical values needed for $P(h)$, $P(\neg h)$, $P(e/h)$, and $P(e/\neg h)$ in Bayes' theorem. Based on the primary and secondary narratives, a Deliberator might believe that $P(h) \geq P(\neg h)$, that is, the probability that harm will occur is equal to or greater than the probability that harm will not occur, and could further believe based on these narratives that $P(e/h) > P(e/\neg h)$, that is, the probability that the evidence that Jones

is holding a glass supports the hypothesis that harm will occur is greater that the probability that the evidence that Jones is holding a glass does not support the hypothesis that harm will occur. This would then allow the Deliberator to conclude that $P(h/e) > P(\neg h/e)$. In other words, the probability that the evidence that Jones poses a threat supports the hypothesis that harm will occur is greater than the probability that the evidence that Jones poses a threat does not support the hypothesis that harm will occur. But how does this help us understand the role of probability (or risk) in establishing the threshold of harm? Simply put, Deliberator can use the primary and secondary narratives, cashing-in on their "gut feelings", to ascertain whether the evidence (gained via the primary and secondary narratives) supports the hypothesis (that is, their concern) that harm will occur to Victim (and indeed whether this is greater than the probability that, given the same evidence, harm to Victim will not occur).

This means we can (a) avoid objections related to the claim that the threshed of harm requires establishing an objective or universal probability that harm will occur (or, rather, avoid issues related to the objectivity/universality of the probabilistic element of the threshold of harm), and (b) avoid issues associated to assigning numerical values to evidence to inform probabilities that, in turn, drive what seems to be an arbitrary boundary (or threshold) beyond which one could say that harm is "likely", "more likely than not" or "probable", "beyond reasonable doubt", etc., all of which are imprecise terms that bring with them certain values (we might think, for instance, that something is likely if the probability of it occurring is greater than 25%, that something is more likely than not if the probability of it occurring is greater than 50%, and that it is beyond reasonable doubt that something will occur if its probability is greater than 95%—or, indeed, any other seemingly arbitrary percentage). We can do both (a) and (b) whilst retaining the idea that probability

does play *a* role in the threshold of harm—but with certain caveats and not at the expense of other factors. Deliberator uses the available evidence to assess whether there is a threat of harm to Victim, but importantly this does not determine if or when Deliberator should intervene. Indeed, Lazar (2018: 865–866)—speaking on the arbitrariness of deciding on a number for a probability threshold—highlights that '[w]e should take the implied precision of numbers with a pinch of salt […] We cannot pin them down to infinite decimal places'; and Dougherty (2013) shows how our moral reasons are epistemically vague even though they are not metaphysically indeterminate.

Primary and secondary narratives, although enabling a Deliberator to potentially intervene quickly and in a way that factors in a calculation of risk of harm to a Victim, do not say anything about nor do they ensure that the decision to intervene is *morally* justifiable. In other words, the "gut feeling" that a Deliberator has only gets us so far. What is further required by a threshold of harm, and what has been alluded to in this section, is a consideration and understanding of philosophically relevant factors to determining the threshold of harm. The next section will look beyond simple risk and towards philosophically relevant factors in establishing a threshold of harm.

4.6. THE TERTIARY NARRATIVE: PHILOSOPHICAL CONSIDERATIONS

This section will outline the tertiary narrative and argue that a number of philosophically relevant considerations are required to form a philosophically informed threshold of harm; these considerations include the necessity of an intervention (including the necessity to act), the proportionality of an intervention to the threatened harm, and the liability of certain agents to be harmed in an intervention.

4.6.1. A STARTING POINT: SHOULD THE NUMBERS COUNT?

John Taurek (1977) questions whether the number of people harmed should be a philosophical consideration in moral situations; for our purposes, the question relates to the extent to which the number of people affected should be taken into account by a Deliberator when deciding if/when to intervene in a moral sequence. Indeed, *prima facie* it seems that one should not inflict harm on more people than the number of people saved or to a greater extent than the harm threatened to Victim. In other words, it seems intuitive that it is not justifiable to kill three people to save one person, nor that one person should be maimed to prevent another from receiving a scratch.

Indeed, it is important to reflect on Taurek's argument in relation to moral sequencing and position the discussion that will follow in the overarching aim of this chapter. I will not offer a direct discourse on Taurek's argument, nor will I present an overview of the literature related to Taurek to date (though I will signpost the reader to some relevant responses), nor shall I attempt to offer a solution to the many issues arising from Taurek's account. To do so would place this discussion firmly in a debate concerning whether the number of Victims counts to a decision to intervene (although moral sequencing as presented in this thesis is concerned only with moral sequences involving one Victim). That said, multiple moral sequences might be initiated by the same Initiator (maybe even via the same threat), but their being directed at individual Victims results in each being its own moral sequence in virtue of the singularity of the Victim. (For example, Initiator *I* may roll a boulder towards a group of people *G* containing three people, their individual membership of the group represented by a unique number following *G*. *G1*, *G2*, and *G3* may therefore all share the shame threat (of the boulder) and may share the same Initiator (*I*), but all three are the Victim of their own moral sequence.) What this does, in effect, is to place moral sequencing outside

of such debates on whether the number of Victims should count, for there simply can ever be only *one* Victim of any given moral sequence (this is discussed further in §4.5.1.5).

However, moral sequencing does not preclude the possibility of other intra-sequence agents from being harmed as a result of or due to the intervention. In other words, it is possible that, by intervening, an actual or would-be Intervener can only prevent harm to Victim by causing or allowing harm to befall another intra-sequence agent or agents. This is important as this sort of consideration, namely whether and in what ways and to what extent harm might or would befall other intra-sequence agents, must be undertaken by a Deliberator when deciding if and when to intervene. There may also be the further consideration that intervening at t^1 would cause harm to Bystander[83] *B1*, intervening at t^2 would cause greater harm to *B1*, and intervening at t^3 would cause harm to *B1* and to another Bystander *B2*. There is therefore a *prima facie* issue confronting Deliberator concerning whether and the extent to which intervening to save Victim from death would be permissible at t^1 (where the act of intervening causes *B1* to suffer a broken arm), at t^2 (where intervening inflicts death on *B1*), or at t^3 (where intervening inflicts death on *B1* and *B2*). These questions will be addressed throughout this chapter, with each helping to construct an understanding of the threshold of harm.

Moreover, the relevance of discussing Taurek will become increasingly apparent; through a discussion of some of the most salient points in Taurek's argument we will see the importance of delineating different types and kinds of intervention, which will itself highlight a number of philosophically relevant and related issues—including the necessity

---

[83] The notion of a Bystander was introduced in §2.2.1.1.

of the intervention, the proportionality of the intervention, and the liability of the agent(s) harmed by the intervention—for establishing the threshold of harm. These issues will become more apparent and will sprout from the more general consideration of the number of people harmed, as discussed by Taurek. Some key philosophical literature relating to the above concepts, predominantly from the self-defence literature, will be discussed and applied to the case of other-defence (the Intervener's defence of Victim) as the intervening action of Intervener in a moral sequence. Before this, however, there will be a discussion about which of the interventions delineated are the most relevant to moral sequencing and will isolate those interventions on which an assessment of the threshold of harm will be based. This discussion will ultimately paint a picture of a philosophically relevant threshold of harm for moral sequencing.

4.6.1.1. PRIORITISING ONESELF AND THOSE OF SPECIAL CONCERN

Taurek (1977) offers some grounds for challenging the utilitarian position that, *ceteris paribus*, we ought to save the greatest number of people (group) rather than an individual. In making this claim, Taurek considers Foot's case of a person who requires the entirety of a drug to survive at the expense of the lives of five others who, with only one-fifth of the drug each, would all survive. There isn't enough of the drug to save all six. Taurek (1977) considers that, in a situation in which the individual owns the drug, utilitarian considerations—including the (potential) happiness of the group and the intrinsic value of the group (in comparison to the individual)—do not outweigh the individual's reasonable action to consume the entirety of the drug (thereby ensuring the demise of the group). This might be because, for instance, the individual values their life more than the others in the group. If this is so, and the individual has no special obligation to the group, then, according

to Taurek, the individual's choice to consume the drug (owing to them owning the drug) can extend to permitting another (third-party) individual to consume it in certain cases. If the drug owner does not need the drug, but their friend or spouse does, then, since their lives are 'more important' than the lives of the strangers in the group, the drug can be given to such individuals with whom the owner has a 'special concern' (Taurek, 1977: 303) or to whom the owner has a 'special obligation' (Taurek, 1977: 296). Other authors also think that a one can, and perhaps even must, give greater importance in their deliberations to those for whom they have special concern (Hurka, 2007; Lazar, 2013; for criticism, see Lefkowitz, 2009). The origin of this special concern/obligation can be due to a number of factors that include but are not limited to: personal obligation/concern (e.g. the concern that a parent has for their children), professional obligation (e.g. the obligation that a doctor has to their patients), or obligations arising from 'an explicit contract or promise' (Taurek, 1977: 296)[84].

What Taurek does, in effect, is offer a reason for why the owner *O* of drug *d* can offer *d* to a third-party *T*—the importance or value of a third-party individual to *O* can direct the distribution of *d* either in part or in totality. Importantly, then, Taurek explains why *O* is not obliged to offer *d* to a third-party group *Tg* over a third-party individual *Ti* for whom *O* has a special concern. *O*'s ownership of *d* permits *O* to offer *d* to *anyone*, regardless of the number of people that could otherwise have benefitted from *d*. In a case where *O* values *Ti* more than either the members of *Tg* or *Tg* as a whole (e.g. *Ti* is a close friend or spouse of *O*), *O* has reason (*viz.* is permitted but is not required) to give *d* to *Ti* over *Tg*. In short,

---

[84] There is a related issue concerning the degree of obligation/concern. I might, for instance, have special concern for both my brother and my spouse, but if I value or like my spouse more (say, because she is also the mother of my children), does this increased degree of concern for her mean that I am entitled to save her over my brother? This raises the issue of degree of concern but that, due to reasons of space, will not be discussed here.

according to Taurek, the fact that a greater number of people would benefit from receiving *d* does not in itself justify nor should it direct the distribution of *d*; the distribution of *d* is determined by the importance or value of a person or people to *O*. This line of reasoning can be extended, Taurek claims, to other moral situations. For example, if I am presented with a situation in which I have to choose between saving my arm or saving another person's life, I might choose to save the former—I might place greater value on myself and my arm than the life of another—and doing so would, according to Taurek (1977), be a reasonable course of action. Similarly, in a case in which I have to choose between saving my friend's/spouse's arm and saving the life of a stranger, it would be reasonable for me to recognise that I have a special concern for my friend/spouse that might drive me to save my friend's/spouse's arm over a stranger's life[85].

Taurek's claim that we can save our own arm/life over a stranger's life tracks the common intuition and evolutionary psychological claim that we prioritise our own wellbeing over those of strangers[86]. However, for moral sequencing, it not usually the case that a Deliberator would be presented with a case in which he must choose to save himself over the Victim—or, more specifically, decide to intervene in a way that forces the Deliberator to choose to save himself over the Victim. However, one might conceive of a case in which

---

[85] Gregory Kavka (1979) argues that Taurek's claim that I would be justified in saving my own arm over another agent's life and that I would be justified in giving the drug that I own to a friend over a stranger has 'some intuitive appeal', although he disagrees with Taurek's reasoning that it has any bearing on whether the numbers count—and further that, because Taurek 'violates a very plausible transitivity requirement' and 'ignores the significance of an important analogy between morality and prudential rationality', Taurek is simply not entitled to claim that the numbers do not count.

[86] Other authors are sympathetic to the claim that an agent has a personal (or rather agent-centred) prerogative to prioritise their own interests (including their body) over another agent; see Nancy Davis (1984), Helen Frowe (2008), and Jonathan Quong (2009).

the only way for the Deliberator to intervene to prevent harm befalling a Victim is to act in a way that is self-harming. This will be discussed in §4.6.2.6.

4.6.1.2. BEING OF NO SPECIAL CONCERN AND DEMONSTRATING EQUAL CONCERN

But what of cases in which *O* does not have any special concern for *Ti,* or indeed where *O* does not value *Ti* any more or less than members of *Tg*—who should receive *d*? In such cases, Taurek thinks that each individual has an equal claim to *d* and each should be offered an equal chance of survival. *O* should therefore express 'equal concern' for each individual; one way of expressing equal concern for all individuals would be for *O* to 'flip a coin' (Taurek, 1977: 303) to decide who receives *d*. To those who would exclaim "But surely the lives of, say, fifty people matter more or outweigh the life of a single person?", Taurek (1977: 306) would respond that even in cases where he could either save only one person (*Ti*) or save fifty people (*Tg*), he 'cannot see how or why the mere addition of numbers should change anything'. In a fire where there are six objects of equal value, five of which are collected together in Room 1 and one of which is in Room 2, but where I cannot save all six items due to time constraints, I would of course choose to save five objects rather than saving just one—saving five objects is more valuable to me than saving just the one. However, unlike objects whose value can be cumulative, Taurek (1977: 306) does not think we should 'think of [people] in just this way'; instead of considering a person's value, we should instead seek to 'empathize with them'. We should move away from thinking of people as having objective value and should instead consider what it would be like to be in an individual's position and, in doing so, we would be 'terribly concerned about what happens to him' (Taurek, 1977: 307). Indeed, both *Ti* and members of *Tg* would, *from their*

perspective, think that saving the opposite (group or individual) would be worse for them: for *Ti,* giving *d* to *Tg* would be worse for him; and for members of *Tg,* giving *d* to *Ti* would be worse for them. Therefore, in a situation where *O* has no special concern for any individuals, flipping a coin is a suitable method of expressing equal concern for each person involved. However, even some minor scratching of the surface of this argument reveals a number of issues for deciding the threshold of harm.

This might go so far to providing some justification for whether and in what ways a Deliberator mobilises their personal resources when intervening to save a Victim—if the Deliberator has special concern for Victim, he might mobilise his personal resources quicker or to greater effect than he would if Victim was not of any special concern to him. This might explain why we might intuitively act quicker in situations involving those for whom we have special concern—a mother would likely decide to run into the road to save her child from a speeding car in much less time than it would otherwise take her to choose to do the same for someone else's child in the same circumstance. This can be explained by considering 'special concern' as a type of secondary narrative than can aid in the decision-making process. In other words, a Deliberator would likely err on the side of caution in their decision-making in those cases where they have special concern for the Victim, thereby driving them to intervene earlier in the moral sequence than they would for someone for whom they have no special concern. So, although a Deliberator must first and foremost assess the primary narrative of a moral sequence to ascertain if and when to intervene to prevent harm befalling a Victim (as outlined above in §4.2. to §4.5.), special concern for a Victim is a reasonable contributory factor in the secondary narrative that might bring forward the threshold of harm (although divorced from an assessment of the probability of harm eventuating).

Can we therefore say that the threshold of harm should account for special concern by permitting a Deliberator to favour those individuals? The issue of prioritising those of special concern to Deliberator implies that Deliberator has a choice to save either an agent for whom they have special concern or another agent for whom they have no special concern. But as I will argue in §4.6.1.5., moral sequencing is concerned only with whether a Deliberator should intervene to save a Victim, regardless of whether the Victim has a special relationship to the Deliberator. The threshold of harm therefore cannot include a special concern requirement (that a Deliberator prioritise an agent for whom they have special concern) for the simple reason that such a decision—choosing between agents—is beyond the parameters of moral sequencing. However, it seems plausible that the reasons provided by Taurek (and other authors) for why one might or should prioritise those for whom they have a special concern could be factored into a Deliberator's decision-making process (via the secondary narrative), even if a special concern claim cannot be made a part of the threshold of harm itself.

4.6.1.3. MOBILISING NON-OWNED RESOURCES

Moreover, Taurek's account cannot explain how a Deliberator should mobilise resources that they do not own[87]. This is problematic for moral sequencing since many of the cases of moral sequencing we might put forward, including the case of Jones glassing Smith, involves a Deliberator who does not mobilise their own resources. To explain this further, consider that the Deliberator is an on-duty Police Officer (*PO*), part of whose job it is to

[87] It is worth noting that moral sequences are not concerned with the Victim becoming their *own* Intervener, and so the first case of deploying one's own recourses to save oneself is not an issue that will be discussed further.

prevent and tackle crime. *PO* might mobilise his own body (which we can consider as belonging to *PO*, see §4.1.) and thereby mobilise their own resources (*PO* might, for instance, restrain Jones or push Smith out of the way). However, *PO* might intervene by mobilising resources that are not theirs; *PO* might grab a nearby walking stick and hit and smash the glass that Jones has thrown at Smith, thereby preventing harm to Smith, but in the process damaging the walking stick that belongs to someone else. This highlights an interesting issue related to the resources that someone employs when intervening—we might say that the threshold of harm should be raised when mobilising resources that the Intervener does not own. However, this is problematic for a number of reasons. For instance, I might wet and use a cloth that I do not own to douse a fire that would kill someone or some people. (It is for such a reason that other factors not related to ownership must be considered—and such a discussion will take place in §4.6.4. concerning necessity, proportionality, and liability).

4.6.1.4. EQUAL CLAIMS TO RESOURCES

The issue is compounded if we consider a case in which *PO* can save only one of two Victims since both Victims might, according to Taurek, have equal claim to the resources at *PO*'s disposal—both Victims pay appropriate taxes, etc. that in effect fund the resource of public protection, thereby entitling both Victims to an equal claim of the resource of *PO*'s efforts to save *them* from harm. Taurek considers an analogous case in which a community live on an island with an active volcano which erupts. Islanders are stranded on the north side of the island where a large group of them await rescue from the island's only Coast Guard ship; a smaller group of islanders are also stranded on the south side of the island awaiting the same ship. The Coast Guard does not have time to save both groups, but

problematically both groups of islanders have 'equal claim to the use or benefit of that resource' (Taurek, 1977: 311)—we might say that they have paid their fair share of taxes that have funded the provision, or that they are all entitled to government resources since a government is charged with protecting its citizens, or some other reason. In such a case, the captain of the Coast Guard may flip a coin to decide which group to save (thereby respecting each individual islander's equal claim to the resource) and as a result set course for the south side. However, the captain may instead immediately set course for the north side without first flipping the coin if, for instance, it is Coast Guard policy to always save the greatest number of people possible. This case illustrates how, actually, the captain (here the person directing the resource provision) can have a 'duty-bound' obligation (via contractual obligations, institutional policy, legal guidelines, etc.) to save certain people (islanders on the north side) even in circumstances where each individual (each islander) has an equal claim to the resource.

So, coming back to *PO*'s dilemma, which Victim should *PO* save? If *PO* has no special obligation to either Victim and is not duty-bound to enforce certain procedural policy, etc., then Taurek would claim that *PO* has reason to flip a coin to decide how to deploy the resources to which both Victims have a claim. But if there are, for instance, Police guidelines on who to save (i.e. in situation *S* always deploy resource *R* in way *w*), then *PO* has reason to act according to those guidelines (it may, for instance, state that *PO* should always safeguard children, even if at the expense of an adult coming to harm). The problem is that this merely provides *a* reason, and not necessarily a *philosophically justified* reason, to deploy resources (*viz.* install a barrier to harm or intervene) in a certain way. In other words, *PO* can only state that they have a reason to act in a certain way (i.e. that guidelines, policy, legislation, etc. makes them duty-bound to act that way) but cannot necessarily state that

acting in that way is philosophically justifiable (i.e. that they have good philosophical reasons for acting in that way). Indeed, both Victims might have equal claim to the resource, and/or *PO* might have special consideration for one or both Victims, but the type of intervention itself (that is, how the resource is deployed) is arguably a larger contributory factor to the decision to intervene[88]. In other words, we might grant that *PO* might intervene earlier in moral sequences involving a threat to someone for whom *PO* has special concern (that is, the threshold of harm might, for *PO*, be comparatively lower when the harm threatens someone for whom *PO* has special concern over someone for whom *PO* has no special concern). We might further grant that, where there is no special concern for any intra-sequence agents, *PO* should act in a way that demonstrates equal concern for all relevant intra-sequence agents. Further still, we might grant that agents have greater say over how their resources are used. However, when agents have an equal claim to a resource (irrespective of whether the Deliberator has any special concern for the relevant intra-sequence agents) we are left rather empty-handed—what we require is an account of when intervention is justifiable (and in what ways and under what circumstances certain interventions are justifiable) that can help explain (and possibly vindicate) Taurek's claims but further help a Deliberator decide whether and how to intervene to prevent harm (I will call this the 'type-of-intervention problem'). Addressing this issue will drive the following sections.

---

[88] There is a subsidiary issue here too. What I will call the 'still-has-claim problem' is that the Victim that PO is not duty-bound (in terms of policy, etc.) to save might disagree with that policy (etc.) and therefore still feel entitled to their claim on the resource.

4.6.1.5. SOME PROBLEMS WITH TAUREK'S ACCOUNT AND THE RELEVANCE TO MORAL SEQUENCING

There are a number of issues that Taurek does not consider or does not sufficiently support—and these are relevant to moral sequencing in a number of ways.

The first concerns how we can make sense of the equal claim of two Victims (*V1* and *V2*) to resource *R*, but where *R* is an individual (*O*) who has their own resources (i.e. their body or property) at their disposal that can be deployed to save one but not both Victims; the problem is further compounded if *O* has a special concern for *V1* but not for *V2*. Such a consideration is not, admittedly, an issue for individual moral sequences *per se* since a moral sequence (as defined in chapter 2) involves a direct threat of harm to an individual Victim. It therefore cannot be the case that there exists a moral sequence in which there are two Victims competing for the harm-reducing or harm-eradicating resources of a Deliberator/Intervener (here, *O*). That said, this does pose an issue for the larger framework of moral sequencing, since it is quite possible that there exists a Deliberator who must choose between saving a Victim (*V1*) in one moral sequence (*MS1*) and saving another Victim (*V2*) in a different moral sequence (*MS2*). In such a case, Deliberator might choose to become an Intervener in *MS1* and thereby prevent harm from occurring to *V1*, but by doing so simultaneously becomes a Forbearer in *MS2* (since harm occurs to *V2*). *O* might have a special concern for *V1* that drives them to deploy their resources to save *V1* over *V2*, or *O* might not have any special concern for *V1* and *V2* and thereby flip a coin which dictates that *O* save *V1* over *V2*. The Deliberator can therefore be involved as a Deliberator in two different moral sequences and stands at a philosophical crossroads in their determination of which (if either) moral sequence to intrude into, with their decision: (i) impacting on whether to save *V1* or *V2* from harm (if either); and (ii) determining what sort of agent they

become (e.g. an Intervener, Forbearer, etc.). However, this seems to have more bearing on issues concerning the responsibility of the Deliberator for their choice of actions; indeed, if the Deliberator has special concern for *V1* and so becomes an Intervener in *MS1* but by doing so forbears to prevent harm to *V2* in *MS2*, this might impact on the type/level of responsibility a Moral Assessor might attribute to them for forbearing in *MS2*. Likewise, in a case in which the Deliberator does not have special concern for *V1* or *V2*, and so—to express equal concern for both *V1* and *V2*—flips a coin to decide who to aid (where the coin-flip determines that the Deliberator should save *V1* over *V2*), a Moral Assessor might also seek to adjust the type/level of responsibility attributed to them for forbearing to prevent harm to *V2* in *MS2*. In both cases, the Deliberator chooses to forbear to prevent harm to *V2* because of either their special concern for *V1* or the way in which they have demonstrated equal concern for both *V1* and *V2* if they have no special concern for either Victim, and not because they have negligently, maliciously, or otherwise decided to not intervene to prevent harm occurring to *V1*.

Second, Derek Parfit (1978) offers an insight into Taurek's argument and explains why I am *not* entitled to claim aid from a third-party individual to save my arm over saving another person's life. Taurek seems to hold that we can move from understanding why we might save my arm or my child's arm (*viz.* someone for whom I have a 'special concern' or a 'special obligation') over a stranger's life to the claim that a third-party individual can choose to save the life of one person rather than the life of five (in circumstances where that third-party individual has no special obligation, or rather equal concern, for each of the six). However, as Parfit points out, this takes (or rather presupposes) an 'agent-neutral' view of the reasons that this third-party individual has for acting when in fact some of the moral reasons in-play related to deciding to save *V1* over *V2* are 'agent-relative' by virtue of the

'agent-specific' moral permissions that individual has for acting in that way (Parfit, 1978). In other words, my saving of my arm over your life is an agent-relative moral permission tied to the idea that I am justified in prioritising my own wellbeing. This explains why my saving of my arm (or my life) over your life might be morally permissible, however this reasoning does not extend to third-parties—it cannot provide any moral permission for a third-party to save my arm (or my life) over your life. The only way that a third-party individual could reach a standpoint on the moral permissibility of saving my arm or life over your life is if they could provide relevant agent-neutral (rather than agent-relative) moral reasons for doing so. This agent-neutrality is presupposed by Taurek in his argument and is not clearly demarcated from the appeal to agent-relative moral reasons (e.g. prioritising my wellbeing over others, favouring those for whom I have special concern, etc.). It is exactly this sort of agent-neutrality that we seek from, and I think we can find in, a philosophically justifiable threshold of harm.[89]

Third, and following on from the second issue above, even if we agree with Taurek that we may give priority to our own welfare, and this therefore vindicates the choice to save my arm rather than a stranger's life, it is not clear what the parameters are. Might I choose to save my finger instead of a stranger's life? What about choosing to save myself from receiving a scratch? And what about choosing to save one of my belongings over saving another agent from harm? Parfit (1978) makes a similar point when asking whether we should choose to save our umbrella over saving a stranger from death. These questions are not addressed and remain unanswered. This rest of this chapter seeks to remedy this.

---

[89] For other criticisms of Taurek, see Gregory Kavka (1979) who argues that Taurek's argument does not ground the claim that the number of people harmed should count and Ken Dickey (1992) who challenges that Taurek's argument holds water 'for private citizens in real-life situations'.

However, before moving on, it is important that I acknowledge that the reader might suggest that the way in which I have presented the issues above skews the point Taurek is trying to make. Taurek asks who we should save (and at whose/what expense) whereas moral sequencing, as presented thus far, seeks to determine whether intervening in way x at t^n is justifiable in order to save a Victim from harm h. The way it achieves this is by looking to ascertain a threshold of harm, beyond which an intervention becomes morally permissible. However, it is plain to see that the decision that a Deliberator must make is similar to the decision agents in Taurek's examples are asked to make: in both cases, the decision to save is centred on the harm inflicted (or otherwise) on agents in order to prevent or alleviate harm to someone else (i.e. a Victim). Indeed, if in a moral sequence a Deliberator decides to not intervene (let us say that the Deliberator has decided that the only way to save the Victim from harm would be to drive through a crowded place, killing many), the Deliberator has in effect decided to save[90] the lives of those that would otherwise be killed during the process of saving the Victim. However, instead of seeking to determine *who* to save (and at what/whose expense), moral sequencing seeks to determine *whether* we should save the Victim (and the expense at which those decisions are made). Put this way, Taurek's argument offers a number of interesting conclusions for moral sequencing and understanding the threshold of harm. So, although moral sequencing is not concerned with how resources should be deployed in conflicting circumstances, it is concerned with whether

[90] I am aware that the use of the term 'save' here is problematic, and using this term in this way means we find ourselves in the territory of counterfactuals—an area far from pertinence to the point being made. However, I ask the reader to permit me the use of this term to capture the common-sense claim that, in deciding not to prevent harm from befalling the Victim, the Deliberator thereby avoids acting in a way that would have seen those Bystanders harmed in the process of saving the Victim. For present purposes, the idea of saving and preventing harm from befalling are conceptually similar.

they should be deployed at all (that is, whether the relevant threshold of harm has been reached).

4.6.2. DELINEATING TYPES OF INTERVENTION

So far, my account of moral sequencing has remained silent with respect to possible different types (and subsequent kinds) of interventions. However, as it will become clear, it is important to delineate different types of interventions for ascertaining the threshold of harm—the type/kind of intervention matters for if and when a Deliberator is justified in intervening in a moral sequence[91]. Looking back to the 'type-of-intervention problem' (see §4.6.1.4.), if intervening to save one Victim *V1* requires deploying resources that affect another agent (for instance, the deployment causes harm to a Bystander *B* or impinges on Initiator *I*'s autonomy to initiate a NPET), but intervening to save another Victim *V2* does not, we might believe that the type of intervention required (i.e. whether the intervention harms *B* or impinges on *I*'s autonomy) might affect whether and in what ways an Intervener deploys resources at their disposal to prevent harm from coming to a Victim (*V1* or *V2*). Delineating different types of interventions requires firstly recognising that some types of interventions come at a cost, and some agents therefore have a stake in the intervention itself. This section will: (i) comment on how interventions can either be costless or costly to an agent; (ii) outline two types of intervention, namely 'non-agent-affecting interventions' and 'agent-affecting interventions'; (iii) summarise the three different kinds of agent-affecting interventions ('autonomy-affecting', 'property-affecting', and 'harmful' interventions), associated sub-kinds, and whether they can be considered costly or costless

---

[91] I am grateful to Jonathan Parry and Patrick Tomlin for pressing me to more clearly delineate the different kinds of interventions.

interventions; and finally (iv) foreshadow how and why these different interventions matter for establishing the threshold of harm in light of the following sections' discussions of the importance of accounting for liability, necessity, and proportionality when establishing this threshold and thus when intervening. Importantly, then, throughout §4.6.2.2. to §4.6.3. some *prima facie* conclusions will be made with respect to the threshold of harm (e.g. whether certain factors lower or raise the threshold of harm), and these conclusions will later be tested against the emergent relevant concepts of liability, necessity, and proportionality.

4.6.2.1. COSTLESS AND COSTLY INTERVENTIONS

An Intervener can intervene in a moral sequence in a way that does not cause harm to an agent (including herself), does not infringe on an agent's autonomy (that is, does not affect a course of action decided on by another agent), or does not cause damage to an agent's property. I call such interventions 'costless interventions'. Contrastingly, an intervention might come at a price: it might infringe on the autonomy of an agent, it might damage property, or it might harm an agent—all thereby becoming 'costly interventions'. But let us not get too far ahead. First let us consider costless interventions, as illustrated in the following two cases.

4.6.2.2. NON-AGENT-AFFECTING INTERVENTIONS

Rolling Boulder 1.1

A small earthquake dislodges a large boulder from a steep, bumpy hill. At the foot of the hill is a small but popular carnival. The boulder bounces down the hill. Bailey is situated half-way down the hill. Seeing the rolling boulder and realising that it might roll into the carnival, crushing many people there, Bailey decides to push a sizeable rock in its path. The boulder hits the rock and stops.

Rolling Boulder 1.2

Sawyer is with some friends at the top of the hill and trips and falls into a large boulder which then bounces down the hill. Bailey decides to push a sizeable rock in its path. The boulder hits the rock and stops.

Both Rolling Boulder 1.1 and 1.2 are cases in which an intervention from Bailey occurs but at no cost. This 'costless intervention' is an intervention in which the intervention is not at the detriment of any agent (there is no agential cost of intervening). This is unlike other interventions, discussed shortly, which do come at some cost to an agent ('costly interventions'). The interventions in both of the above Rolling Boulder cases are costless in so far as the intervening act of rolling the rock into the path of the boulder and the subsequent destruction of the boulder does not come at an agential cost—no agent is harmed or infringed on by this intervention. I am aware that there are a number of potential clarificatory issues concerning how we might define and understand a costless intervention: for instance, what of the physical, mental, and emotional exertion/toll required by/of the intervening act—surely the intervention has cost something of the Intervener in one of these ways? I

acknowledge that certain interventions might involve certain physical, mental, and/or emotional costs to an Intervener (or indeed other intra-sequence agents). However, for the purposes of moral sequencing, I ask the reader to take a narrow understanding of what constitutes a costless intervention and understand an intervention as having a cost to an agent only if that intervention causes (physical) harm to an agent or, as shall be discussed shortly, infringes on an agent's autonomy (to act). A costless intervention is one that therefore does not have significant cost, since it neither causes (physical) harm to an agent nor infringes on an agent's autonomy (to act). In Rolling Boulder 1.1 and 1.2 no agent is harmed. It is in this sense that the intervention of rolling the sizeable rock into the path of the boulder is costless—it isn't costing the Intervener, or any other agent for that matter, anything of significance. There is, however, an important difference between the two cases: the first involves a non-agential initiator and the second involves an agent unintentionally/accidentally initiating the moral sequence. Differentiating between these two ways of initiating might seem overindulgent, but it is an important difference to make since both represent a distinct type of intervention, namely a 'non-agent-affecting intervention'. All non-agent-affecting interventions are costless: neither Rolling Boulder 1.1 nor 1.2 involves an intervention that affects an agent. The intervention in Rolling Boulder 1.1 does not affect any agent since no agent is involved in the initiation of the sequence nor harmed by the intervention itself; and in Rolling Boulder 1.2, although the Initiator is an agent (Sawyer), Sawyer does not sustain harm as a result of the intervention and Sawyer's autonomy has not been infringed since Sawyer did not choose to set the boulder in motion. Since the boulder was rolled unintentionally/accidentally, it makes no sense to say that Sawyer's decision to roll the boulder down the hill has been infringed. Sawyer therefore remains unaffected by the intervention.

We now have an understanding of what constitutes a costless intervention, and further how non-agent-affecting interventions are themselves costless interventions. However, in some situations, intervening in a moral sequence comes at a price. The next section will outline what constitutes a 'costly intervention', another type of intervention (namely an 'agent-affecting intervention'), and different kinds of agent-affecting interventions (namely 'autonomy-affecting', 'property-affecting', and 'harmful' interventions), and sub-kinds of these.

4.6.2.3. AGENT-AFFECTING INTERVENTIONS

A costly intervention is an intervention that is morally expensive in that it costs an intra- or extra-sequence agent something of their autonomy, property, and/or physical health. We might therefore think that any intervention that affects an agent (that is, any 'agent-affecting intervention') is costly. Sections 4.6.2.4. to 4.6.2.7. will outline three different kinds of agent-affecting interventions and will explain that and why they are costly interventions.

So how we can make sense of 'agent-affecting interventions', *viz.* interventions that affect an intra-sequence agent? Consider the following case, which illustrates the first kind of agent-affecting interventions.

4.6.2.4. AUTONOMY-AFFECTING INTERVENTIONS

> *Rolling Boulder 2*
>
> Sawyer pushes a large boulder down the hill. Seeing the possible harm to those at the carnival, Bailey pushes a sizable rock into its path. The rock halts the boulder.

Like Rolling Boulder 1.2, Rolling Boulder 2 involves Bailey pushing a rock into the path of the boulder to prevent the boulder from rolling into the carnival. However, the difference between the two is that, in the latter, Sawyer pushes the boulder of his own volition. Why does this matter? Sawyer's decision to push the boulder (that is, the exercising of his autonomy) renders any intervention that affects that action an infringement on Sawyer's autonomy; Sawyer decided to put the boulder in motion and an intervention that halts the boulder necessarily infringes on Sawyer's choice to roll the boulder[92]. It is for this reason that Rolling Boulder 2 illustrates a costly intervention—the intervention costs Sawyer his autonomous action of pushing the boulder—and further grounds the intuitive claim that intervening in a way that is autonomy-affecting requires justification. Rolling Boulder 2 is therefore the first case of an agent-affecting intervention, namely an 'autonomy-affecting intervention'.

The concept of an autonomy-affecting intervention will be developed further as a result of some philosophical issues that arise from some further Rolling Boulder questions that *prima*

---

[92] I am aware that agent *A* respecting another agent *B*'s prior action (*viz.* prior decision to act) *x* consequently thwarts *A*'s subsequent decision to act in way *w* that affects *x* and, problematically, by doing so *B* (or rather *x*) impinges on *A*'s autonomy (to act in way *w* that affects *x*). However, for current purposes, I ask the reader to put aside such considerations.

facie seem to be different kinds of intervention. Let us move on to consider further cases that, rather than involving a rock, involve an agent's property.

4.6.2.5. PROPERTY-AFFECTING INTERVENTIONS

> *Rolling Boulder 3.1*
>
> Sawyer's actions cause a large boulder to roll down the hill towards the carnival. Bailey manoeuvres her car in front of the boulder before removing herself from harm. The boulder collides with and crushes the car, stopping the boulder from rolling any further.

> *Rolling Boulder 3.2*
>
> This time, Bailey manoeuvres a nearby stranger's car in front of the boulder without the owner's permission before removing herself from harm. The boulder collides with and crushes the car, stopping the boulder from rolling any further.

Rolling Boulder 3.1 and 3.2 both illustrate the second kind of agent-affecting intervention, namely a 'property-affecting intervention'. The intervening act of rolling a car in front of the boulder affects an agent's property (either belonging to the Intervener or belonging to another agent); Rolling Boulder 3.1 illustrates a case in which it is the Intervener's property that is affected and Rolling Boulder 3.2 illustrates a case in which another intra- or extra-

sequence agent's property is affected[93,94]. Differentiating cases of Intervener-owning and non-Intervener-owning property is important in so far as we must ask whether the intervention using or employing certain property is justified and in what circumstances such property can be used. For instance, it seems intuitive that utilising one's own property is more justifiable than utilising property that belongs to another agent, especially without their consent. That is to say that using my own resource seems more justifiable than using another's resource without their permission since one can be reasonably expected to have a say in how the resource that one owns is used. After all, it is the property owner that risks their property being damaged. In short, autonomy-affecting interventions are those that infringe on a (third-party) agent's autonomous decision[95] to act (including utilising their resources) in the way they decide.

---

[93] 'Property' and 'resource' are used throughout this chapter in order to link the concept of property-affecting interventions to Taurek's discussion of utilising 'resources' and to demonstrate that the *use* of property makes that a *resource* in a Deliberator's or Intervener's arsenal.

[94] Although there are many ways in which property can be affected, and indeed a number of ways in which property can be defined, I ask the reader to understand property-affecting interventions as those interventions that utilise a resource that belongs to an agent (but excluding the agent's own body). The agent's own body is excluded from this category as harm to an agent's body forms the third kind of intervention, a 'harmful intervention', discussed shortly. I am also aware that specifying what constitutes an agent's ownership of a resource is philosophically problematic, but for current purposes (and to avoid a legal discourse) I ask the reader to take a layperson's understanding of ownership that resource *R* belongs to agent *A* if *A* has legitimately purchased, acquired, inherited, or otherwise legally owns *R*. I also ask the reader to restrict the scope of property to that pertaining to personal private property and exclude cases of common or collective property. This is to ensure that we avoid issues related to fringe cases in which two or more agents collectively or commonly own a resource but where one agent agrees to its use in the intervention but where other relevant agents do not, since such issues are not directly pertinent to the present aim of delineating the different types and kinds of intervention.

[95] Please note that I use the term 'autonomous decision' as a term of art. Becoming embroiled in a debate about free will, determinism, and so on is unimportant for current purposes and is tangential. I acknowledge that this term may seem loaded, but it is employed only to signify that the decision at hand relates to the agent's autonomy in so far as the agent is making a choice to act in way *x* rather than in way *y*. I ask any reader who wishes to baulk at this term to temporarily suspend their disbelief.

This leads to another sub-kind of property-affecting intervention in which the third-party agent does in fact give permission[96] for their property to be used in the intervention, illustrated by Rolling Boulder 3.3 below.

> *Rolling Boulder 3.3*
>
> This time, Bailey manoeuvres a nearby stranger's car in front of the boulder with the owner's permission before removing herself from harm. The boulder collides with and crushes the car, stopping the boulder from rolling any further.

This case highlights the importance of accounting for consent (and indeed, dissent or non-consent) to use another agent's property when delineating different kinds of intervention. Consent to use property is an important factor of consideration in the decision to intervene since an intuitive claim is that using someone's property with their permission is *ceteris paribus* more justifiable than using that property without permission. But is permission to use another agent's property required if that property will not be damaged? This introduces another interesting consideration for delineating intervention—namely whether the non-

---

[96] The third-party agent that agrees to their resource being used might say, after their property has been damaged/destroyed, that they did not agree to it being used in *that* way or gave permission for it to be used but not damaged or destroyed. Such considerations are interesting, but such a discussion would lead us too far astray—we would be led into a debate about the intentions of both parties, whether either party could have reasonably foreseen damage of or destruction to the property, the limits and restrictions of permission-giving, and so on. I therefore ask the reader to consider permission-giving at face-value (as-is) and take a case of permission to use one's property as providing the Intervener with a genuine right to use that resource as they see fit without limitations and immune from post-hoc claims about the use of that property. This is not intended to undervalue the philosophical intricacies of permission-giving, but serves our current purpose of delineating a distinct sub-kind of property-affecting interventions that has bearing on whether, when, and in what ways one intervenes in a moral sequence.

damaging of another agent's property can ground a claim to utilise that agent's property without their consent. Consider the following case.

> *Rolling Boulder 3.4*
>
> This time, Bailey manoeuvres a heavily armoured car in front of the boulder with the owner's permission before removing herself from harm. The boulder collides with the car. The car stops the boulder from rolling any further, and the car's armour is sufficiently robust to prevent the car from sustaining any damage.

In Rolling Boulder 3.4, the property used in the intervention is not damaged. Now, if the armoured car belongs to Bailey, then we might simply say "You decided to intervene by using your armoured car, and it hasn't been damaged by the boulder. How fortuitous!". Whether or not the Intervener's property is damaged as a result of their choice to use their property to intervene in the moral sequence is largely inconsequential for the purposes of this case; the fact that the Intervener decided to utilise their own property to install a barrier to harm in the moral sequence is not affected by whether or not or the extent to which that property is damaged—the potential damage to their property would likely have been a factor in their decision to intervene. Indeed, damage to property would likely seem a factor in deciding whether one is justified in intervening (*viz.* whether the threshold of harm has been reached). For instance, the prospect of my property being damaged or destroyed (whether this is at least foreseeable) would, *ceteris paribus*, likely justify a later response than if it was certain that my property would not be damaged. The difference, it seems, is that the former is a costly intervention whereas the latter is a costless intervention. But wait! What we now have is a seeming agent-affecting intervention that is actually a costless

intervention; if the car is not damaged, it seems to make no sense to talk about that intervention being property-affecting, since the property has not been affected (*viz.* damaged). And what of Rolling Boulder 3.4, which involves Bailey using another agent's property (but that does not sustain any damage)? Does the fact that Bailey utilised another agent's property with their consent impact on the kind of intervention? Could this be a costless intervention? After all, no agent's property has been affected by its use as an intervention. Because of this, one might think that cases in which property is utilised with permission and in which property is not damaged as a result of its employment as an intervention in a moral sequence (*viz.* non-property-affecting interventions) should be categorised and understood as a non-agent-affecting intervention (and thus a costless intervention). However, at the time of deciding to intervene Bailey can't be sure that there will be no damage. Taking someone's car and risking damage to it infringes their property rights even if no damage in fact occurs. This is what firmly grounds the claim that such cases can still be considered property-affecting—a *post hoc* justification that property was not damaged does not change the fact that the intervention was property-affecting in as much as the intervention was the kind of intervention that was enacted on the understanding that there was non-negligible risk of damage to that property. And, even if the car belonged to Bailey, it is still a property-affecting intervention if at the time of intervening there is a non-negligible risk of damage to Bailey's property. This will be important for §4.6.2. when attention will be turned to understanding how delineating different kinds of interventions matter for establishing the threshold of harm.

It therefore seems that one might increase the threshold of harm if the utilisation of property would result in its damage or destruction (in contrast to utilising property that would not be damaged through its use in the intervention). In other words, the magnitude of damage to

property is a factor relevant to ascertaining the threshold of harm. Indeed, we can make the reasonable assertion that an intervention that would destroy property would *ceteris paribus* (that is, without considering other relevant factors) require greater justification than an intervention that would only damage that property; and we might further say that damage to my property as a result of my actions or damage to another agent's property with their permission to use that property is more justifiable than damaging the property of another agent when permission has not been sought or granted to use that property. This becomes important when considering whether an agent is liable to have their property used or damaged/destroyed in the intervention. This issue will be discussed in §4.6.4.3. when examining the concept of liability in establishing the threshold of harm.

But, returning to the Rolling Boulder cases, what of cases in which the property owner dissents to the use of their property (that is, explicitly does not consent to their property being used)? Consider the following case.

> *Rolling Boulder 3.5*
>
> This time, Bailey asks the owner of a car if she can manoeuvre it in front of the boulder. The car owner dissents to Bailey using the car. Bailey uses the car anyway and manoeuvres it in the path of the rolling boulder. The boulder collides with the car, thereby stopping the boulder from rolling any further.

In Rolling Boulder 3.5, the car (property) owner denies Bailey the use of his car. Does Bailey's use of the car owner's car, regardless of the car owner's dissent, matter? The answer, I believe, relies on the extent to which the intervention is agent-affecting. If the car

is damaged as a result of its use, then the intervention was property-affecting (since it affected the car owner's property). However, even if this is so, this property-affecting intervention in which the property is used regardless of the property owner's dissent seems different to property-affecting interventions that are detrimental to only property that Bailey owns and different to those that are detrimental to another agent's property but where the property owner consents to its use. We therefore have reason to believe that not all property-affecting interventions are equal; some require more justification, and this will likely affect the threshold of harm. In other words, 'other-property-affecting interventions' (those interventions that affect the property of any agent apart from the Intervener) are different to, and in some cases require greater justification to install, than 'self-property-affecting interventions' (those interventions that affect the property of the Intervener), especially in those cases where permission to use another agent's property has not been sought or has been sought but denied. This requires some clarification. Cases in which I utilise my own property to intervene and cases in which I use another agent's property with their consent, and where in both cases the property is damaged, are both property-affecting interventions (the former a self-property-affecting intervention and the latter a property-affecting intervention); however for the purpose of establishing the threshold of harm, the two are, I believe, equal. Using one's own property and using another's property with permission are *ceteris paribus* equal with respect to their determining the threshold of harm in a given sequence since in neither has either agent's autonomy been affected and neither agent has been harmed, although both involve property being damaged—parity here is determined by ensuing permission to use property. An issue arises when another agent's property is affected and when either (i) permission has not been sought for their property to be used or (ii) permission is sought but denied. In such cases, there is, in the case of (i), an avoidance of obtaining or non-seeking of permission and, in the case of (ii), a disregarding of dissent—

both of which are different to the above case (iii) of an other-property-affecting intervention in which consent is sought/obtained from the property owner. Both (i) and (ii) are different from (iii) by virtue of the procurement of consent to use property.

But is either worse? Is (i) more justifiable than (ii) or *vice versa*? *Prima facie* we might think that (ii) is less justifiable than (i), for (i) simply involves not seeking permission to use another agent's property whereas (ii) involves blatantly disregarding that agent's refusal to give permission. Indeed, if a child asked whether they could eat a cookie and the parent forbade the child from eating the cookie, but the child ate it anyway, then the parent would *ceteris paribus* likely scold their child more vehemently for disobeying them had the child simply eaten the cookie without asking. Problematically, then, it seems that it would always be preferable for the Intervener to not seek permission to use another agent's property: if one believes that (ii) is worse than (i) then one would likely never ask for permission to use another agent's property—one would simply opt for (i) over (ii). In other words, if using the property of an agent who has dissented to its use is worse than not seeking permission to use it and then using it anyway, then it seems that an agent would always opt for not seeking permission to use another agent's property for fear that, if denied, their use of the property regardless would require far greater justification. The reason for this seems apparent: that ignoring the agent's dissent to use their property takes the intervening action up one notch—it makes the intervening action autonomy-affecting. As we have seen in previous Rolling Boulder cases, the use of property can make the intervention property-affecting (c.f. Rolling Boulder 3.1 to 3.3) or non-property-affecting (c.f. Rolling Boulder 3.4 to 3.5) (and a costly agent-affecting intervention in the former and a costless non-agent-affecting intervention in the latter), but what is interesting about Rolling Boulder 3.5 is that, although the intervention

is not property-affecting, it is autonomy-affecting since the case involves Bailey directly disregarding the car owner's dissent.

However, I think that (i) and (ii) should be considered equal with respect to the role they play in the threshold of harm. A solution to the above problem would be for the parents to have laid-down a rule that the child must always seek permission from them to eat cookies, and so if the child (aware of this rule) eats the cookie without permission and protests at their punishment for doing so, the child's parents can simply remind the child that they were/are required to ask permission and failed to do so. The same applies to general cases of using property. We can say that there is an expectation that agents seek permission from the property owner to use their property before using it—this is indeed the case in certain tort law and criminal law (e.g. trespass and theft, respectively)[97]. In other words, there needs to be (and arguably there currently exists through, for example, legislation) a constraint against using another agent's property without permission—there is an expectation and requirement that permission to use property is sought from the property owner[98]. Ignoring dissent, then, is always autonomy-affecting (since, to reiterate, that agent's autonomous decision to choose to use their property in the way they see fit has been infringed) and may

---

[97] Whether this should be the case, the extent to which this is enforceable, and so on are considerations that lie beyond the scope of this thesis. All that is important for our purposes is a recognition that such rules, laws, etc. exist and in such cases we are legally or otherwise bound to them.

[98] I am aware that there exist various caveats to this in various forms, including in legislation, where, for example, the state can use, destroy, or take ownership of property without permission. Indeed, it would seem odd for a convicted criminal to have the right to protest at the destruction of the illegal drugs that he owns. However, these cases are usually limited to the application of the law which, for current purposes, are not wholly relevant to ascertaining the threshold of harm; if anything, such problematic cases serve as evidence that there exists jurisprudential issues with respect to the intersection of the law (and application of it) and moral issues (including those relating to the decision to intervene in a moral sequence).

further be property-affecting if that agent's property is damaged as a result of the intervention.

Three questions, or rather potential issues, arise from this. These will be briefly presented and discussed in turn. First, what if, due to time-constraints, the Intervener did not have time to ask the property owner, and where if the Intervener had sought permission the threatened harm to Victim would have occurred? In such a case, the intervention is clearly still autonomy-affecting since the property owner was not consulted (and perhaps even property-affecting if that property was damaged as a result of its use). The apparent mitigating circumstance that the effectiveness of the intervention could only have been ensured by the non-seeking of permission to use the other agent's property might be taken into account by a Moral Assessor when attributing responsibility (or lack thereof), but it does not affect nor should it deter us from the claim that the intervention itself was costly and agent-affecting (autonomy-affecting, property-affecting, or both, depending on circumstances). Indeed, we can still say that the property owner had a right to have a say in whether and how their property is/was used, but this might not reflect whether and the extent to which such an Intervener should be attributed responsibility, nor whether the magnitude of the threatened harm to Victim could or should override that right in circumstances where it is impractical or not possible for permission to be sought. This introduces the concept of accounting for the magnitude of harm, which will be discussed in §4.6.4.

Second, what if the property owner was unavailable and so the Intervener was unable to seek permission to use that agent's property? This question seems similar to the first in that they both involve the Intervener knowing that the resource they want to utilise is the property of another agent and differs only in so far as each relate to two different situational

constraints, namely: asking for permission given time-constraints would render the use of the property in the intervention as redundant (in the first); and asking for permission cannot be sought since the property-owner is not available to be consulted (in the second). It is for this reason that we can treat both the same for the purposes of delineating different types/kinds of interventions—what is therefore said above in relation to the first question applies here to the second.

Third, what if the Intervener did not know that the property they were using belonged to another agent? Perhaps the car that Bailey pushed in front of the rolling boulder was so rusty and seemingly disused/abandoned that they thought that the car had been dumped/disowned. This seems another *prima facie* variation of the first question discussed above. The first and second involve the Intervener being situationally-constrained from seeking the property owner's consent, and the third question differs from the other two due to an epistemic difference, namely that, while in the first and second questions the Intervener knows that there is a property owner but is unable to seek their permission to use is, in the third the Intervener does not know that the resource being used is owned. This reflects a *prima facie* epistemic difference between the two questions and sets the scene for how these might be addressed. In such a case, we have to ask whether ignorance of ownership excuses the Intervener from utilising resources that they do not own. This is a problematic issue for a number of reasons and requires considering a number of related points, including: the context of the situation (whether the Intervener is in someone's home or in a deserted airfield makes all the difference to whether they have grounds to claim that they thought that the ornament they used did not have an owner); whether the Intervener could have been reasonably expected to believe that the resource had no owner; whether the Intervener acted recklessly or negligently (with respect to their use of the resource); whether the Intervener

was delusional (if, for example, the Intervener believed that the resource they took from someone's house did not belong to the owner of the house); and so on. I will not attempt to address these issues here. I simply wish to acknowledge that this objection, although not undermining the delineation of a property-affecting intervention, might cause some problems when such cases are encountered for ascertaining the threshold of harm. For current purposes, I ask the reader to limit such cases to those in which the Intervener can be reasonably expected to believe that a resource belongs to an owner—holding such a common-sense view will alleviate the sort of philosophical but tangential issues outlined above.

Returning to Rolling Boulder 3.5, what if the car is *not* damaged? What if the car, like in Rolling Boulder 3.4, is suitably armoured and thus does not sustain any damage (although the property owner has dissented to its use)? Does this make a difference? It seems not. The intervention cannot be considered property-affecting since the car is not damaged; but, importantly, this does not mean that the intervention was costless. The very ignoring of the car owner's dissent to use their car in the intervention is autonomy-affecting in as much as Bailey infringed on the car owner's autonomous decision to not allow their car to be used; the car owner acted in a way that ensured that their car would not be used, and by using their car regardless, Bailey's action (*viz.* intervention) was autonomy-affecting to the car owner. Non-property-affecting interventions can therefore be autonomy-affecting interventions.

The task of outlining property-affecting interventions has involved climbing a deceptively difficult philosophical conceptual mountain, however the fruits of our effort have culminated in an understanding of the sub-kinds of property-affecting interventions in moral

sequencing that are *prima facie* relevant to establishing the threshold of harm. Let us now consider another kind of agent-affecting intervention.

4.6.2.6. HARMFUL INTERVENTIONS

> *Rolling Boulder 4.1*
>
> Sawyer's actions cause a large boulder to roll down the hill towards the carnival. Bailey pushes Parker, a nearby stranger, in front of the boulder. The boulder crushes Parker's legs but provides just enough resistance to stop the boulder in its path.

Rolling Boulder 4.1 introduces the concept of a 'harmful intervention'[99]. This kind of intervention is harmful in so far as the intervention (the installation of the barrier to harm) itself causes physical[100] injury[101] to an agent; more specifically, the Intervener intervenes in a way that causes injury to a Victim. In the above case the Intervener (Bailey) pushes a Bystander (Parker) in front of the boulder, thereby using Parker as the barrier, resulting in injury to Parker. In short, Rolling Boulder 4.1 delineates a case in which an Intervener's

---

[99] Since moral sequencing is concerned with preventing harm to human agents, I will not discuss cases of interventions causing harm to non-human animals. There is a case to be made concerning whether harm caused to a pet, for example, would be property-affecting (if one believes that one owns one's pets) and/or harmful (at least to that animal). I leave such discussions to those who wish to extend moral sequencing to non-human animals.

[100] There are, of course, other harms, including mental, psychological, and emotional harm. These are further sub-kinds of harmful interventions, but which will not be discussed here. As discussed in chapter 2, for the purposes of the moral sequencing presented in this thesis, I limit the harm under consideration to physical harm. Hereafter, when I refer to 'harm' or 'injury', I am referring only to physical harm/injury.

[101] I use the term 'injury' to refer to any harm to an agent up to but not including death. I use the term 'harm' as a catch-all term to refer to any type of (physical) injury, including death.

actions cause harm (that is, injury but not death) to a Bystander. There is, of course, a related sub-kind of harmful intervention, presented in the following case.

Rolling Boulder 4.2

This time, the boulder crushes Parker to death. Parker's body halts the boulder.

Rolling Boulder 4.2 replicates the previous case with the exception that this time, instead of merely injuring a Bystander, the intervention causes the death of a Bystander. We can therefore separate those cases in which an Intervener injures, and those cases in which an Intervener kills, a Bystander by intervening. Both Rolling Boulder 4.1 and 4.2 are therefore both 'Bystander-harming interventions' since both involve either the injury or death of an agent (Bystander) other than the Intervener; the former is a 'Bystander-harming intervention' and the latter is a 'Bystander-sacrificial intervention' (both falling under the umbrella of a 'Bystander-harming intervention'). This leads us to consider contrasting cases of 'self-harming interventions' (with similar demarcations) in which the Intervener themselves sustains harm due to their intervention.

Rolling Boulder 4.3

Instead of grabbing Parker, Bailey throws herself into the path of the boulder. The boulder crushes Bailey's legs but this provides just enough resistance to stop the boulder in its path.

Rolling Boulder 4.4

This time, the boulder crushes Bailey to death. Bailey's body halts the boulder.

Rolling Boulder 4.3. and 4.4 are both 'self-harming interventions' since, in the former, as a result of their intervention the Intervener is injured and, in the latter, is killed. The former is therefore a 'self-injuring intervention' (whereby, as a result of their intervention, the Intervener is injured) and the latter is a 'self-sacrificial intervention' (whereby, as a result of their intervention, the Intervener dies). It is important to recognise that self-harming interventions can only come from the Intervener and no other agent—this might sound counter-intuitive, for it is possible, one might say, for Parker to throw himself in front of the boulder, and thereby injure/sacrifice himself. But this, importantly, is not a case of a non-Intervener undertaking a self-harming intervention; rather the act of throwing himself in front of the boulder makes Parker the Intervener and so it is entirely consistent to stipulate that self-harming interventions can only be enacted by an Intervener (this would therefore also make Bailey a Deliberator rather than an Intervener[102]).

---

[102] Unless, of course, Bailey threw herself in front of the boulder too. Then we would have multiple self-harming interventions and Bailey would also be considered an Intervener. It is worth remembering the discussion in this chapter on how there can be multiple Interveners (and so such a claim is consistent with previous discussions).

What *prima facie* conclusions can be drawn from these sub-kinds of harmful interventions? It seems intuitive to claim that, *ceteris paribus*, self-harming justifications would require a lower threshold than those interventions that harm other agents (i.e. Bystander-harming interventions). That is to say that (other things considered) if the Intervener was the only agent that would be harmed by intervening then they would likely be justified in intervening earlier than if the intervention would cause harm to another agent (e.g. Bystander); this is because the Intervener has themselves decided to risk harm to themselves by intervening whereas the Bystander has not—and, further, we can make the additional claim that such a Bystander-harming intervention would also be autonomy-affecting. It is not that these agent-affecting interventions are necessarily cumulative and this *ipso facto* raises the threshold, but intuitively we can see that more is at stake when such multiple agent-affecting interventions (whether at the detriment of one or multiple agents) are enacted. However, I do not think that we are entitled to claim that *all* interventions that harm other agents require greater justification than self-harming interventions. Consider the following example.

> *Rolling Boulder 4.5*
>
> Sawyer's actions cause a large boulder to roll down the hill towards the carnival. Sawyer rolls down the hill adjacent to the boulder, racing it to the bottom. As Sawyer rolls past Bailey, Bailey kicks Sawyer into the path of the boulder. Sawyer's body stops the boulder in its path but harms Sawyer in the process.

Whilst Rolling Boulder 4.1 to 4.4 all involve harm (either injury or death) to a Bystander or the Intervener, Rolling Boulder 4.5 provides a case in which the Intervener uses the Initiator

in the intervention and by doing so the Initiator is harmed[103]. We can call such cases 'Initiator-harming interventions' with the same two sub-kinds as the other two (Bystander- and self-) harming interventions (namely 'Initiator-injuring interventions' where the Initiator receives an injury and 'Initiator-sacrificial interventions' where the Initiator is killed). Moreover, and similarly to the *prima facie* conclusions concerning the lower threshold for damaging property than destroying it (c.f. §4.6.2.5.), it seems intuitive to think that an intervention resulting in death to an agent would require a higher threshold than an intervention that would only injure them[104]. I think such an intuition is sensible for Bystander-harming and self-harming interventions, but, as will become apparent, it is less intuitive to think that sacrificing the Initiator would require a higher threshold than the sacrificing of the Intervener—after all, it was the Initiator who initiated the moral sequence, and it is therefore as a direct result of their actions that the Intervener installs their body as a barrier to harm. It seems intuitively uncomfortable to say that the threshold should be higher if harm will be caused to the Initiator[105] than if harm would occur to the Intervener. I therefore think we have some *prima facie* grounds for rejecting the claim that *all*

---

[103] Of course, the Initiator might intervene in the moral sequence that they initiated of their own accord and, by doing so, become the Intervener in their own moral sequence (c.f. the previous footnote). Although I find this a philosophically uncomfortable idea, it is one that seems nonetheless possible—and certainly the moral status (including blame for initiating, praise for intervening, etc.) of this will be assessed post-sequence by a Moral Assessor. I will, however, not discuss this fringe case any further so as to remain focussed on the most salient issues.

[104] We might also say the same about degree of injury. I will defer such a discussion to §4.6.2.6.

[105] It is worth noting here that because the Intervener causes harm to the Initiator, and the harm to the initiator did not exist prior to the Intervener acting, it might be said that by intervening the Intervener initiates a non-pre-existing threat to Initiator and thereby initiates a new moral sequence. As uncomfortable as this might be, moral sequencing clearly states that this is so. However, importantly, the joint Intervener in the first moral sequence and Initiator of the second moral sequence is (a) not liable to defensive harm against the threat they pose to the Initiator of the first moral sequence (see the discussion in §4.6.4.3.) and (b) will *ceteris paribus* likely be attributed less responsibility for initiating the second moral sequence than the Initiator of the first (see the discussion in chapter 5).

interventions that harm other agents require greater justification than self-harming interventions—this perhaps is so for Bystander-harming interventions, but not for Initiator-harming interventions. This claim will be supported during the discussion of liability in §4.6.4.3. in which we will see that the reason for this rests on the idea that, unlike Bystander, Initiator is *ceteris paribus* liable to harm and therefore such Initiator-harming interventions actually requires a lower threshold than those interventions that harm a Bystander or a (non-initiating) Intervener.

But what about Victim-harming interventions, where the intervention to Victim is harmful but is less harmful than the harm they would have sustained without the intervention? Consider the following case.

> *Kneecapping*
>
> Gunman spots Rival across a large field. Gunman aims at Rival's head and shoots with precision—the bullet would certainly kill Rival. Helper shoots Rival in the kneecap, thereby making Rival fall to the floor and miss Gunman's bullet.

In Kneecapping, Rival (here, the Victim) stands to be killed by Gunman (here, the Initiator)—let us say that receiving a bullet to the head will cost Rival 100 units of harm. However, Helper (here, the Intervener) sees an opportunity to reduce the harm that befalls Rival by shooting Rival in the knee. This kind of intervention is harmful since it involves harming Rival, but it is different from the other kinds of harmful interventions since it is the only intervention that seeks to avert or reduce harm to Rival *by* using Rival in the intervention to save them from Gunman's threatened harm. Let us say that Rival receives

50 units of harm by being shot in the knee by Helper. The question is whether Helper inflicting harm on Rival in order to avert or reduce the harm that Rival would otherwise receive (from Gunman) without that intervention is justified. It seems intuitive to claim that any action that averts or reduces the harm to Rival would be justifiable, even though doing so involves harming them—this is a cost-benefit calculation or trade-off between two harms. What remains to be seen is whether this intuition holds philosophical weight; this will be discussed in more detail in §4.6.4.

But what of Victim-sacrificial interventions, where the Intervener kills Victim? Are those cases *prima facie* justifiable too? Consider the following case.

> *Tracheotomy*
>
> Blake notices that Addison is choking. The restaurateur attempts the Heimlich manoeuvre without success. Addison will soon die of asphyxiation. Remembering a recent film scene, Addison rummages through his bag for a sharp implement and asks the restaurateur for a straw. Using the implements available, Blake performs an emergency tracheotomy. The tracheotomy does not go well and Blake's actions cause Addison to die of exsanguination.

In Tracheotomy, Blake is the Intervener since he attempted to install a barrier to prevent/mitigate the threatened harm of death to Addison (here, the Victim). However, Blake's intervention was the cause of Addison's death—after all, Addison dies of exsanguination, not asphyxiation. This is an interesting case since Blake's actions intended to avert or mitigate the threat of death to Addison, yet Blake's actions brought about death

to Addison regardless. So, although it might seem strange to say that an Intervener can enact a Victim-sacrificial intervention (that is, enact an intervention that kills Victim in order to avert or mitigate the threatened harm to Victim), for completeness we must differentiate between those Victim-harming interventions that are Victim-injuring interventions (that bring harm but not death to Victim) and Victim-sacrificial interventions (that kill Victim). It is entirely plausible that an Initiator can act in a way that seeks to injure Victim in order to prevent Victim from being killed, but by doing so enacts an intervention that itself kills Victim. It does, however, seem problematic to say that an Intervener can justifiably intervene in a way that causes death to Victim. However, it seems intuitive that although we might not wish to justify Intervener enacting a Victim-sacrificing intervention, since such an intervention cannot by definition be an intervention that seeks to avert or mitigate harm to Victim, an Intervener might justifiably enact a *prima facie* or intended Victim-injuring intervention that later transpires to be or transforms into (*viz.* snowballs into) a Victim-sacrificial intervention. The issue is that *intentional* Victim-sacrificial interventions simply cannot be considered an (appropriate) barrier to the threatened harm (see §2.2.2.4.) and therefore seem unjustifiable —it simply makes no sense to say that Intervener intentionally kills Victim in order to save Victim. Indeed, a Victim-sacrificing intervention cannot, by definition, be an intervention that seeks to avert or mitigate harm to Victim since: (a) it makes no sense to say that such an intervention will avert death to Victim, since a Victim-sacrificial intervention will kill Victim too; and (b) it makes no sense to say that such an intervention will mitigate harm to Victim since there is no greater or more permanent harm

than death [106] . However, unintentional Victim-sacrificial interventions (*viz.* those interventions that are intended merely to harm Victim (be a Victim-injuring intervention) but which in fact kill Victim) might be justifiable in some cases. For instance, where threat of death to Victim is certain, it seems that any intervention that seeks to avert the death of Victim would be justifiable, even if those transpire to be Victim-sacrificial interventions, since the attempt to avert death to Victim is surely justifiable even if such an attempt ends in the death of Victim. However, in a situation in which Victim is not at threat of being killed, it would seem unjustifiable to enact a potentially Victim-sacrificing intervention to avert the threatened harm of, say, a broken leg or a scratch. We will return to this in §4.6.4.2. when discussing the role of proportionality in the threshold of harm.

What is common among all harmful interventions, then, is that the intervention necessarily involves the use of an agent's physical body, where the intervention is or risks being detrimental to that or some other agent's body. This is why, although an agent's body can reasonably be considered the property of that agent (see §1.2.5.), we need to delineate material property (property-affecting interventions) from bodily property (harmful interventions). However, there is an interesting related case that arises: similar to Rolling Boulder 3.4 in which the car's armour is sufficiently robust to prevent the car from sustaining any damage during its use in the intervention, there are those cases in which an agent's body is used in an intervention but where the agent is neither injured nor killed. It therefore cannot be the case that *all* instances of an agent's body being used in an

---

[106] One might argue that I am mistaken in asserting (b); one might say there are a number of things worse than death, including being tortured or sexually assaulted/abused on its own or before eventually being killed. Although I understand that such prolonged harms might indeed be more experientially harmful for the Victim, I think it is still reasonable to state that the permanence of death, prevention of future endeavours, etc. make death the greatest and most permanent harm to Victim.

intervention counts as a harmful intervention, since in such cases no agent sustains harm. Consider the following case.

> *Biker 1*
>
> Jones is about to glass Smith. Anticipating this, Bailey grabs Biker who, having only just entered the pub, is still wearing his biking helmet, and pushes him between Jones and Smith at the moment that Jones throws a glass at Smith. The glass hits Biker's helmet and smashes. Biker's helmet is sturdy enough to mean that they are not harmed (and the helmet is not damaged).

In Biker 1, Biker is not harmed and so the intervention cannot be considered a harmful intervention (nor a property-affecting intervention since the helmet is not damaged). What we therefore have is a case in which an agent's body is used in an intervention but without that agent sustaining any harm. So is this a costless intervention? In short, no. Biker 1 is a clear case of an autonomy-affecting intervention since Biker's choice to act in a way other than to intervene in the moral sequence was infringed—Bailey used Biker's body without their permission, and this grounds the claim that such an intervention is autonomy-affecting. Let's change the case slightly to see if all cases of an agent's body being used in an intervention are autonomy-affecting.

Biker 2

This time, Biker voluntarily stands between Jones and Smith at the moment that Jones throws a glass at Smith. The glass hits Biker's helmet and smashes. Biker's helmet is sturdy enough to mean that they are not harmed (and the helmet is not damaged).

The difference in Biker 2 is that this time Biker *chooses* to intervene in the moral sequence (and, like in Biker 1, they are not harmed nor is their property damaged as a result). As such, the intervention cannot be said to be autonomy-affecting. Here we have an intervention that is not autonomy-affecting, is not property-affecting, and is not harmful; for this reason, Biker 2 is a clear case of a non-agent affecting intervention and is thus a costless intervention—the intervention is not costing Biker anything.

From this discussion it is clear that not all interventions involving an agent's body are harmful, although some might be property-affecting and/or[107] autonomy-affecting (and are therefore costly) (as in Biker 1); however, in some cases, an intervention involving an agent's body can be wholly non-agent-affecting and therefore costless (as in Biker 2).

Moreover, it seems that self-harming interventions cannot also be autonomy-affecting, since the Intervener chooses to intervene in a way that is self-harming and this choice removes

---

[107] It is worth stating plainly that agent-affecting interventions are not mutually exclusive; they can occur within the same moral sequence and/or at the same time. Indeed, it is quite plausible that an intervention can be any combination of autonomy-affecting, property-affecting, and/or harmful. If, for example, Biker was pushed in-between Jones and Smith then the intervention would be autonomy-affecting, if Biker's helmet broke on contact with the glass then the intervention would be property-affecting, and if the integrity of the helmet was thus damaged by the glass and the effectiveness of the helmet was compromised thus causing Biker to be harmed by the glass then the intervention would be harmful.

any claim that the intervention is autonomy-affecting. This might be a bitter pill to swallow for some readers since, they might claim, the Intervener might not *feel* as though they had a choice—if the only available barrier to install involved self-harm then that Intervener might feel obliged, even duty-bound, to act in such a way. It is for this reason, some might claim, that we should permit such self-harming interventions to be autonomy-affecting—the Intervener's autonomy is affected in so far as their self-harming action is demanded by the moral sequence of which they find them self a part. But this is simply not true. The threshold of harm serves only as a guide to enable the Deliberator (and Moral Assessor post-sequence) to assess whether a certain kind of intervention at time *t* is, would be, or was justifiable. Moreover, the same claim (that a moral sequence can demand action) can be said of *any* intervention, including Bystander-harming, Initiator-harming, and Victim-harming interventions. But it is simply not the case that moral sequences demand that Deliberators act in any certain way (including those situations in which there is only one available intervention that is harmful to an agent), nor is it the case that the reaching of the threshold of harm in a given moral sequence morally demands or morally pressures the Deliberator to intervene (in a certain way or at all)—the reaching of the threshold of harm simply says that acting in way *w* is or was justified given circumstances C^n.

It is now appropriate to consider an intuitive conclusion regarding the role of such costless interventions for establishing the threshold of harm. Costless interventions are those that are non-agent-affecting interventions and, since they do not cost any agent anything (in terms of autonomy, property, and harm), it seems intuitive to claim that a costless intervention requires a lower threshold than a costly intervention. It further seems that, since such an intervention is costless (in the ways outlined above), certain other considerations are not relevant either; for instance, it seems intuitive that harm to a Victim does not need to be

imminent, nor the intervention necessary to prevent that harm, if the intervention is costless. As such, we seem entitled to the *prima facie* claim that an Intervener can install a barrier to harm that is costless very early on in a moral sequence. This claim, and the associated issues of necessity and imminence, will be discussed in §4.6.4.1.

4.6.2.7. SUMMARISING AGENT-AFFECTING INTERVENTIONS

Agent-affecting interventions can therefore be summarised and categorised as follows:

I. Autonomy-affecting intervention
    i. Intervening by affecting the autonomy of another agent
    ii. Intervening by using the resource of another agent without detriment to the resource
        a. without their permission to use their resource
        b. ignoring their dissent to use their resource
    iii. Intervening by using the resource of another agent with detriment to the resource
        a. without their permission to use their resource
        b. ignoring their dissent to use their resource

II. Property-affecting intervention
    i. Intervening by using my own resource with detriment to the resource
    ii. Intervening by using the resource of another agent with detriment to the resource
        a. with their permission to use their resource
        b. without their permission to use their resource
        c. ignoring their dissent to use their resource

III. Harmful intervention

i. Bystander-harming intervention

a. Bystander-injuring intervention

b. Bystander-sacrificial intervention

ii. Self-harming intervention

a. Self-injuring intervention

b. Self-sacrificial intervention

iii. Initiator-harming intervention

a. Initiator-injuring intervention

b. Initiator-sacrificial intervention

iv. Victim-harming intervention

a. Victim-injuring intervention

b. Victim-sacrificial intervention[108]

We can further distinguish that costless interventions are non-agent-affecting interventions, which include property-affecting interventions where either an Intervener utilises their own property or utilises another agent's property with their permission but without detriment to any property (see Rolling Boulder 3.4 and 3.4), and that all agent-affecting interventions are costly interventions. It is also worth noting that I.iii.a–b and II.ii.b–c are duplicates since these interventions can be both autonomy-affecting and property-affecting. However, there is still one final kind of intervention that requires delineating—'hopeless interventions'.

---

[108] As previous discussed, strictly speaking a Victim-sacrificial intervention isn't an intervention since it cannot be the case that Intervener intervenes to avert injury or death to Victim by killing them. I therefore use this term only as a representation of how a Victim can be killed by an Intervener's actions (e.g. via snowballing a Victim-injuring intervention).

4.6.2.8. HOPELESS INTERVENTIONS

Rolling Boulder 5

Sawyer's actions cause an extremely large boulder to roll down the hill towards the carnival. The sheer size of the boulder makes it clear to Bailey that no available intervention will halt the boulder or otherwise mitigate or lessen the destruction it will bring to the carnival. Desperate, Bailey attempts to intervene regardless. The intervention inevitably fails and the boulder continues on its path towards the carnival, harming many there.

In Rolling Boulder 5, there is no hope of installing an appropriate and effective barrier to harm. Any intervention (available to Bailey) is bound to fail, and so any such intervention can be considered 'hopeless'. Hopeless interventions are a strange kind of intervention, for they can be either non-agent-affecting or agent-affecting (and indeed include any of the related sub-kinds) in any given moral sequence. An Intervener can intervene in a way that is non-agent-affecting (e.g. if Sawyer accidentally rolled the large boulder and Bailey placed a nearby small rock in front of the rolling boulder), autonomy-affecting (e.g. if Bailey pushed a well-armoured agent in front of the rolling boulder), property-affecting (e.g. if Bailey placed his or another agent's car in front of the rolling boulder), and/or harmful (e.g. if Bailey threw herself or another agent in front of the rolling boulder). However, what unifies each of these kinds of interventions under the umbrella of a hopeless intervention is that each intervention is knowingly enacted without any chance of success—such interventions are enacted largely out of desperation to prevent or otherwise alleviate the threatened harm to a Victim. The issue for the threshold of harm is that hopeless interventions that are in some way agent-affecting are intuitively unjustifiable. If the

Intervener hopelessly intervenes then they are unnecessarily affecting an agent's autonomy, unnecessarily affecting an agent's property, and/or unnecessarily harming an agent. It is the unnecessariness of the intervention that seems to ground claims of their unjustifiability (see §4.6.4.1. for a discussion on the concept of the necessity of an intervention)[109]. Conversely, it is unlikely that one would begrudge or even prohibit an Intervener for enacting a hopeless intervention that is non-agent-affecting, even though such an intervention is just as needless (or unnecessary) as those hopeless interventions that are agent-affecting. The difference, however, is that non-agent-affecting interventions are costless whereas agent-affecting interventions are costly, and this intuitively matters. I therefore think that *prima facie* we have reason to think that non-agent-affecting hopeless interventions require little or no justification (since it does not cost any agent anything of value), and that agent-affecting hopeless interventions are unjustifiable (since it does cost an agent something, and possibly something of great value, for no reason or benefit—and are also unnecessary).

4.6.3. RELEVANT KINDS OF INTERVENTIONS AND REFINING THE THRESHOLD UNDER CONSIDERATION

Now that various kinds of interventions have been delineated, it is important to reflect on the kinds of interventions that have bearing on moral sequencing and, importantly for the

---

[109] One might disagree that agent-affecting hopeless interventions are intuitively unjustifiable. What if the only agent affected by an agent-affecting hopeless intervention is Victim, and where Victim is certain to die from the threatened harm from Initiator? Would Intervener not be justified in enacting a million-to-one chance of success (and thus a seemingly hopeless) intervention that is agent-affecting? The answer lies in the fact that there was *some* chance, although a slim chance, that harm to Victim could be averted. It is for this reason that such an intervention cannot be considered a hopeless intervention, but rather a *speculative intervention*, and these are justifiable so long as the threshold of harm is passed (e.g. that the intervention was necessary, proportionate to the threatened harm, and that any agent harmed by the intervention was liable to defensive harm).

task of this chapter, ascertaining which of these interventions are relevant to the threshold of harm. I will outline my case for refining the threshold of harm under consideration to a threshold of physical harm, which will, in the next section, pave the way to assessing some relevant philosophical concepts related to the three emergent key kinds of intervention that rely on establishing a threshold of *physical* harm to justify their enactment.

Moral sequencing is concerned with only those moral sequences in which a non-pre-existing threat *of harm* to a Victim has been initiated (see chapter 2). Moral sequencing, then, is primarily concerned with harmful actions ('harmful' in the sense described in §4.6.2.6.) since the initiation of a moral sequence begins only when an Initiator brings about a non-pre-existing threat (NPET) *of harm* against a Victim. But this seems to, on our understanding of 'harmful', preclude agent-affecting actions that are not *physically* harmful, such as autonomy-affecting actions and property-affecting actions.

Indeed, until now, the threshold of harm has been discussed in the broadest sense of the term 'harm' that tracked the kinds of costly interventions outlined in the previous section. Indeed, we can say that any kind of agent-affecting intervention can 'harm' an agent in the broadest sense—an agent might be 'harmed' physically, psychologically, emotionally, financially, socially, or morally[110] by any autonomy-affecting, property-affecting, and/or harmful intervention. However, since a moral sequence is concerned only with an Initiator initiating a NPET against a Victim, where the NPET relates only to posing physical harm to the Victim (see §4.6.3. in which the harm to a Victim in a moral sequence is limited to physical harm), I propose that we limit our assessment of such a threshold of harm that forms a part

[110] This list is not exhaustive. I imagine there are numerous other ways in which an agent can be harmed non-physically.

of the larger framework of moral sequencing to only physical harms. In other words, I propose that we focus our energy on establishing only a *threshold of physical harm* and not a threshold of harm broadly construed which would necessarily have to include sub-thresholds including a threshold of autonomy-restriction, a threshold of property rights, and a threshold of non-physical harm, amongst others.

It is important to note that I am not suggesting that issues around, for instance, autonomy-affecting interventions and property-affecting interventions (and indeed non-physical harmful interventions) are not at all relevant to establishing whether an intervention in a moral sequence is justifiable. Indeed I think that considerations relating to a threshold of autonomy-restriction, a threshold of property rights, and a threshold of non-physical harm would and do form part of considerations relevant to establishing a comprehensive and robust picture of a larger threshold requirement that, when combined with the threshold of physical harm, establishes the point at which *any* intervention is justified. However, such a task would require a thesis—possibly multiple theses—of its own. I therefore, for the reasons stated above, seek to only outline one such threshold, namely the threshold of physical harm, which I believe to be the most salient issue due to its relation to potential unjustified threat of physical harm to an agent that does not choose or agree to be harmed. My hope is that establishing a threshold of physical harm in moral sequencing will pave the way to, in future work, establishing and importantly synthesising other thresholds that can be used to determine whether it is justifiable to employ other interventions in moral sequences to avert or reduce harm. For now, however, I will concentration on harmful interventions since they involve physical harm.

In establishing the threshold of physical harm, I therefore propose that moral sequencing should concern itself with only certain (non-self-harming) harmful interventions, namely Bystander-harming, Initiator-harming, and Victim-injuring interventions. I rule out the following kinds of interventions from being relevant to our line of enquiry related to establishing the threshold of physical harm for the following reasons:

1. Any non-agent-affecting intervention, since all such interventions are costless and so are justifiable without any limitations. If an intervention is non-agent-affecting and this is costless, it simply makes no sense to ask that an agent wait until a certain threshold has been reached before acting in this way. For example, if Intervener could stop a rolling boulder from killing many people at the carnival by acting (using a rock) in a way that is costless—it is non-agent-affecting because it is not autonomy-affecting (since the boulder rolled of its own accord the Intervener's decision to stop it does not impinge on any Initiator's autonomy), not property-affecting (since the rock used to intervene does to belong to any agent and it is not used in a way that causes damage to any agent's property), and not harmful (since no agent is harmed by the intervention)—it would seem intuitively wrong to require the Intervener to wait until some threshold has been reached. Why should, for example, the Intervener act only when the boulder is progressing in such a way that it seems likely, or perhaps more likely than not, that the boulder will harm carnival-goers rather than rolling into the adjacent barren land? Indeed, the concept of non-agent-affecting interventions are not an issue for moral sequencing for the very reason that they are costless, and so we should not consider such when establishing

the threshold of physical harm[111]. I therefore propose that we not discuss non-agent-affecting interventions any further.

2. Any autonomy-affecting intervention, since infringing on an agent's autonomy is not harmful to an agent. Although infringing on an agent's autonomy is philosophically and potentially also morally severe (it does, after all, in some cases involve restricting the liberty of an agent, negating freedom of choice and/or freedom of action, and so on), it does not physically harm an agent. I therefore propose that, since our focus is on physical harm, we omit this kind of intervention from consideration and not discuss autonomy-affecting interventions any further.

3. Any property-affecting intervention, since damage to property in the parameters described (*viz.* excluding one's body) are not physically harmful to an agent (as defined above), although destruction to property might be non-physically harmful (e.g. losing one's house, one's beloved possessions, family heirlooms, or even the car that one uses to commute to work might have serious and significant psychological, emotional, financial, and/or related health impacts that can be said to "harm" an agent affected by such interventions—although these harms are indirect).

---

[111] I acknowledge the possible objection that *prima facie* costless interventions are not in every case costless under a much broader understanding of that term. Although non-agent-affecting interventions are costless in the ways cashed-out above, we cannot, some might say, cash-out this costlessness in other types of interventions. Picking up a rock off the ground and using it in a moral sequence, although non-agent-affecting, might be 'other-non-agent-affecting' which could be considered costly under a broader definition. For instance, the rock might be part of a non-human animal's home, it might be a unique rock that if destroyed could not be replaced, it could be a decorative rock that brings joy to passers-by, and so on. I acknowledge such reasons but ask the reader to keep from their mind such other-non-agent-affecting interventions since they involve fringe cases and cases that are not directly relevant to establishing the threshold of physical harm.

However, since our focus is on physical harm, I will therefore not discuss property-affecting interventions any further.

4. Any self-harming intervention (*viz.* self-injuring or self-sacrificial intervention), since, although such an intervention is physically harmful (and thus a harmful intervention), it is not autonomy-affecting, making it a separate and special kind of intervention (see §4.6.2.4. in which I claimed that an agent's use of *their own* body to avert the threatened harm to Victim cannot be considered autonomy-affecting since they chose to enact such an intervention)[112]. Self-harming interventions cannot be autonomy-affecting since that agent has chosen to intervene (and chosen to intervene in a way that is self-harming, even if such an intervention was the only available barrier, etc.[113]). I therefore propose that we isolate considerations relating to self-harm from discussions related to the threshold of physical harm, since what seems most salient are those cases in which harm comes to an agent that does not agree to or choose to be harmed. This, in effect, directs our attention away from an agent choosing to act in a way that harms themselves and towards focussing on those interventions that harm another agent. Such discussions about self-harming

---

[112] I am purposely avoiding discussions of accidental self-harming interventions, such as self-sacrificial interventions. Such instances are peculiar since they are neither strictly-speaking autonomy-affecting (since the agent's choice to act hasn't been infringed) nor strictly-speaking non-autonomy-affecting (since they have not chosen to act in that way). I therefore ask the reader to put aside such questions relating to such 'accidental interventions' which I will defer to a future paper.

[113] Problematically there are those cases in which an agent might ask another agent to harm them—such as in cases of BDSM. There is certainly a case for including such cases as instances of self-harm, although I am reluctant to agree since the harm can only be actualised by another (harming) agent. It is possible in moral sequencing for an agent to initiate a NPET against themselves (and thus initiate a self-harming moral sequence) and it is possible for an agent to ask another agent to initiate a NPET against them. However both fringe cases are beyond the scope of this thesis; I will return to the more mainstream and salient issues.

interventions I defer to a future paper so as to focus on the more significant issues related to the threshold of physical harm.

5. Any hopeless intervention, since: (i) hopeless non-agent-affecting interventions are costless and so are justifiable without any limitations and without considering any threshold (see 1. in this list); and (ii) hopeless agent-affecting interventions are unnecessarily costly (that is, they cost an agent something—of their autonomy, property, and/or physical health—without potential or actual gain) and so are unjustifiable in every circumstance. First, those hopeless interventions that are non-agent-affecting are *prima facie* justifiable in every circumstance since, by definition, such interventions are costless and therefore there cannot be any threshold of physical harm beyond which an intervention is or would be justifiable. There is simply no harm done to, and no agent affected by, such a hopeless non-agent-affecting intervention. Second, those hopeless interventions that are agent-affecting are *prima facie* unjustifiable in every circumstance since, by definition, there is no chance that the intervention would avert or diminish harm[114]. It is because such hopeless interventions, when coupled with the fact that they are agent-affecting, are costly but simultaneously without benefit or gain (that is, without chance of averting or alleviating harm to Victim) that grounds my claim that these interventions are

---

[114] There might, of course, be some extremely small chance, but a chance nonetheless, that an apparent hopeless intervention could avert or alleviate harm—the happening of which we, if it did avert or alleviate harm, would likely attribute to some divine intervention or extremely good luck. However these are usually very far-fetched examples. For example, the apparent hopeless agent-affecting intervention of throwing a small coin at car that has been pushed by Initiator and is rolling towards Victim could by chance ping off the car and hit the temple of a fat Bystander causing them to fall in front of the car and thereby halting the car in its track. However I propose that, for sake of argument, we understand hopeless interventions as those that have zero chance of success.

unjustifiable and therefore require no consideration of any threshold. The final nail in the coffin is that the very idea of acting in a way that is hopeless simply cannot strictly-speaking be considered an intervention at all (since one is simply not attempting to avert any harm to a Victim) (see §2.2.2.4. in which I discuss the installation of barriers and what constitutes the installation of an available barrier, an effective barrier, and so on). I therefore propose that when considering the threshold of physical harm that we not concern ourselves with hopeless interventions.

6. Any Victim-sacrificing intervention, since the concept of such an intervention is philosophically problematic. Although an intervention might transpire to be or snowball into a Victim-sacrificing intervention, as we have seen it does not make sense to say that an Intervener can enact a Victim-sacrificial intervention in order to save Victim—it simply makes no sense to say that Intervener intentionally kills Victim in order to save that Victim, even though we might say that Victim died by the hands of Intervener via a Victim-injuring intervention that snowballed (into a Victim-sacrificial intervention). Indeed, the concept of 'sacrificing' Victim (as per the terminology) to save Victim from being killed is absurd. As such, we should not focus on Victim-harming interventions broadly construed but only on those Victim-injuring interventions that seek to avert or mitigate the threatened harm to Victim. That said, we can still account for those cases in which Victim-injuring interventions transpire to be or snowball into a Victim-sacrificial intervention—that is, we can still account for an Intervener's enacting an intervention that seeks to injure Victim to avert Victim's death, but where Intervener's actions unintentionally cause the death of Victim. I therefore propose that when considering Victim-harming

interventions in relation to the threshold of physical harm that we focus on Victim-injuring interventions (whilst understanding that these, in certain circumstances, have the potential to snowball into a Victim-sacrificial intervention).

To summarise, only those interventions that are other-harmful (that is, are Bystander-harming, Initiator-harming, or Victim-injuring) are relevant to establishing the threshold of physical harm, since only these interventions pose risk of harm to an agent (other than those that harm an Intervener, which have been ruled-out due to their autonomous choice to act in a self-harming way). Moral sequencing and the threshold of physical harm therefore focus on the issue of an Intervener harming or potentially harming another agent (apart from themselves) in defence of a Victim. The three kinds of interventions that we are concerned with, and that we shall focus on, for establishing the threshold of physical harm are Bystander-harming, Initiator-harming, and Victim-injuring interventions.

Before moving on, it is important to note that the list of interventions that I have presented is not intended to be an exhaustive list—I do not claim to have delineated every type/kind of intervention. Such a task would be worthy of a tome in itself. The interventions that I have delineated are those that are most relevant to moral sequencing and, importantly, raise and outline some interesting and important considerations and issues for establishing the threshold of physical harm. The rest of this chapter will bring to the fore some of these considerations and will show how the threshold of physical harm can be determined in relation to Bystander-harming, Initiator-harming, and Victim-injuring interventions.

4.6.4. RELEVANT CONCEPTS IN THE PHILOSOPHICAL LITERATURE

This section will outline some salient philosophical concepts from diverse literature that are relevant for the tertiary narrative. The tertiary narrative, seeking relevant philosophical considerations for ascertaining the threshold of physical harm, will, at the end of §4.6., enable us to more clearly explain where and why the threshold of physical harm should be in Bystander-harming, Initiator-harming, and Victim-injuring interventions.

There are three concepts that are most relevant for establishing the threshold of physical harm, namely: necessity, proportionality, and liability. To elaborate, these concepts are relevant to ascertaining: whether intervening in a moral sequence is necessary; whether an intervention is proportionate to the magnitude of the threatened harm to Victim; and whether the agent(s) who are harmed by an intervention are liable to be harmed. These concepts are most frequently discussed in the literature on self-defence and other-defence[115], particularly in wartime ethics, and it is this body of literature that I will draw on to establish the three claims—the proportionality claim, the necessity claim, and the liability claim—which are all claims relevant to establishing the threshold of physical harm and thus ascertaining if and when an intervention in a moral sequence is justifiable (by acting beyond this threshold).

The reason that these claims are required revolve around the need to address some of the issues raised in the examples in §4.5. For instance: Scratch 3 highlights the need to account

---

[115] Some terminological clarifications are in order. The concept of 'self-defence' relates to harm caused by an Initiator who threatens harm to a Victim and where this Victim intervenes in self-defence to avert or mitigate the threatened harm to themselves. 'Other-defence' relates to harm caused by an Initiator who threatens harm to a Victim and where another agent other than Victim, say Intervener, intervenes in other-defence (in defence of Victim) to avert or mitigate the threatened harm to Victim.

for proportionality—killing Aggressor to prevent them from scratching Victim seems disproportionate; Wait and See 2 highlights the need to account for necessity—if harm to Victim could successfully be averted by intervening in two ways, each with varying degrees of harm to Aggressor, it seems unnecessary to choose the intervention with the greatest magnitude of harm to Aggressor; and all of these cases (discussed in §4.5.) highlight the need to account for whether an agent is liable to be harmed by an intervention—although Aggressor might be liable to be harmed to prevent harm coming to Victim (since Aggressor initiated the non-pre-existing threat against Victim), it does not seem that causing harm to a Bystander to prevent harm to Victim would be justifiable.

This section will not present a review of the literature on the areas of necessity, proportionality, and liability, nor will it evaluate the arguments made there—my aim here is not necessarily to advance the literature in these areas (although I think their implication for moral sequencing does extend the literature), but rather to draw on some relevant authors and accounts relating to the necessity, proportionality, and liability claims. I use some of the relevant literature to outline its relevance based on the delineation of interventions above and employ this in moral sequencing to offer a way of conceptualising and understanding the threshold of physical harm based on that relevant literature.

However, two initial and immediate difficulties arise when this literature is applied to moral sequencing. First, much of the literature discusses necessity, proportionality, and liability in relation to self-defence, meaning that they are primarily discussed in relation to the harm an Initiator threatens to a Victim whereby it is the claims relating to the concepts of necessity, proportionality, and liability (and the interconnectedness of them) of a Victim's self-defensive action against the Initiator that are under scrutiny. In other words, it is the action(s)

of a Victim against the Initiator to avoid the harm that the Initiator has threated to the Victim that is under consideration. However, moral sequencing is primarily concerned with other-defence in which an Intervener acts in (other-) defence of a Victim. We therefore need to be careful in how the literature on self-defence is applied to interventions that involve other-defence and how it is brought to bear on moral sequencing. Second, much of the literature is preoccupied with the use of lethal force in self-defence, especially those that discuss the issues in relation to wartime ethics. However, as presented and discussed in §4.2. to §4.5., the threshold of physical harm is relative to the situation (that is, relative to each moral sequence) and the kind of intervention(s) available—as well as an Intervener's relationship to the threatened harm to a Victim (including the intervention's necessity and proportionality and the liability of certain agents to be affected by that intervention). We therefore need to be careful when applying these claims to moral sequencing (or more specifically when conceptualising the threshold of physical harm) to ensure that the literature maintains its relevance whilst also preserving the framework of moral sequencing as provided in chapter 2 and the scope of the threshold as outlined in §4.5.

What follows is a discussion of each of the three concepts.

4.6.4.1. IMMINENCE AND NECESSITY

Let us start the discussion of necessity by contrasting it to a related (and often conflated) issue concerning 'imminence'—I introduce this to firstly explain why it is worth considering for establishing the threshold of physical harm, and secondly to explain that and why I agree with Marcia Baron (2011: 228) that the imminence requirement 'should be jettisoned' and 'replaced with the necessity requirement'. In short, the imminence claim is

that an agent (Intervener) should only bring about harm to an aggressor (Initiator) that has threatened a Victim (in order to avert or reduce the harm that would be caused to Victim) if harm to that Victim from the aggressor (Initiator) is imminent. But what, exactly, counts as an imminent threat? Much of the literature on imminence looks to cases involving 'battered women'[116] to kick-start the idea that the concept of 'imminence' is an important factor of consideration in any discussion of justified (or unjustified) harm in response to threatened harm from an aggressor.

Baron (2011: 265) argues that '[t]he importance of imminence has been overrated' due to a general misunderstanding or 'conflation' of terms; that *x* is imminent is often taken to mean that *x* will occur soon ('in its strictly temporal sense') and that *x* is imminent is also often understood to mean that *x* 'is sure to happen'. This conflation of understandings has led to a situation in which battered women are denied the right to claim self-defence after killing their husbands. In the case of Judy Norman and her abusive husband JT (outlined in Baron, 2011: 232–233), Norman suffered decades of abuse, degradation, and violence from JT and shot him in the head and killed him while he slept[117]. However, because Norman was not under imminent threat of harm from JT, it seems that we are unable to say that killing JT was necessary (and this, it seems, was the reason that the judge refused to give a self-defence instruction to the jury). The issue, however, is that Norman's actions were clearly in self-defence and were, all things considered, necessary to prevent herself from enduring any

---

[116] I do not like this term for numerous reasons, including that the term 'battered' trivialises the serious violence that domestic abuse survivors often endure and seemingly singling-out 'women' as being the victims needlessly neglects the fact that men (and indeed others, including children) can also be victims of domestic violence. I therefore begrudgingly use this term only in order to adhere to the terminology used in the literature, for example in Baron (2011) and Ferzan (2004).

[117] For a good overview of the case of Judy Norman, see Ferzan (2004).

further harm. JT beat her, tortured her, starved her, prostituted her, and ensured that she suffered a multitude of serious harms often involving suffering and degradation. She had presented with injuries at hospital on numerous occasions, the police were called to a number of incidents involving JT harming her, and she confided in the sheriff's deputies that she was being beaten and was desperate; on one occasion, she failed to take her own life by overdosing. Her attempts to avert JT from inflicting more harm on her failed at every stage. It is for these reasons that we are led to the intuitive belief that killing JT while he slept was necessary, even though harm to her was not imminent; killing her husband was the only way, at least in her mind[118], that she could prevent him inflicting more harm on her. There are numerous other similar cases too, such as those outlined by Ferzan (2004: 235–236); in addition to Norman, Victoria Sands and Nelda Lane both killed their violent and abusive husbands but each killed their husbands in circumstances in which neither Sands nor Lane were in imminent danger of being harmed by their respective husbands. Baron uses the case of Judy Norman (amongst other arguments discussed throughout her chapter) to argue that 'imminence should not be regarded as essential to self-defence' (Baron, 2011: 265)[119].

However, according to some authors, Norman, Sands, and Lane's killing of their husbands, in which they acted before there was an imminent threat of harm to themselves, made them fall into what Anthony Sebok (1996) calls a '*Cape Fear* gap': 'there is a gap between the risks that society demands we accept and the amount of protection that society is capable of

---

[118] I echo Baron's (2011: 229–231) comments about the need to leave aside the debate on objective *vs.* subjective imminence, and the difference between harms that are imminent and those that seem imminent.

[119] Other authors support the notion that the imminence requirement should be jettisoned for self-defence, including Horder (2002), Rosen (1993), and Robinson (1984).

providing' (Sebok, 1996: 744). Kimberly Kessler Ferzan (2004) builds on this (Ferzan, 2004: 236–237) and discusses the United States of America (USA) (under the Bush Administration) in the context of international law and its decision to launch Operation Iraqi Freedom as a response to possible weapons of mass destruction (WMDs); the USA cited self-defence as one reason for going to war, despite the USA being under no imminent threat of harm. Ferzan interesting draws a parallel between the USA's decision to go to war and battered women's decision to kill their husbands (since neither were under imminent threat of harm yet claimed that their self-defensive actions were necessary), and argues against what she calls the 'immediately necessary' standard which is the fusing of the imminence requirement with necessity—defined as an entitlement 'to act when they *must* so they may protect themselves [or others]' (Ferzan, 2004: 246). Ferzan argues that the imminence requirement, far from presupposing that it is used to establish necessity, 'informs our understanding of the types of threats that trigger a legitimate defensive response' (Ferzan, 2004: 218). In short, Ferzan (2004: 250) believes that, by itself, the immediately necessary standard ignores a number of relevant factors, including 'the intentions, capabilities, or actions of a putative aggressor [*viz.* Initiator]', and preserving the imminence requirement (divorced from the issue of necessity) ensures that such factors are considered in cases of self-defence (or rather, for our purposes, the decision to intervene to prevent harm to a Victim)[120]. Let us consider a case that Ferzan (2004: 250) discusses (I will call this case Affair):

---

[120] This forms part of Ferzan's (2004) larger argument that the immediately necessary standard 'fails to distinguish self-defence from other self-preferential acts'. However, since what we are concerned with are cases of other-defence, this line of argument won't be discussed. All that is needed is an understanding that Ferzan thinks that the immediately necessary standard wrongly focusses on the needs of the defender; what needs to take centre-stage is an understanding that 'self-defence is an action against a *threat*' (Ferzan, 2004: 252), and from this she builds her defence of imminence divorced from the immediately necessary standard (*viz.* necessity).

Affair

'Assume that A is a friend of B. A always carries a gun and is a quick shot. B likewise carries a gun, but cannot draw quickly. Unbeknownst to A, B is having an affair with A's wife. B also believes that sooner or later A's wife will confess to the affair, and that A is quite hot-tempered and jealous. Thus, if B is around A when A finds out, B knows that he is dead. May B kill A now?'

Ferzan argues that we should answer this question in the negative; although it is necessary that *B* kill *A* to ensure that *B* is not killed by *A*, there was no imminent threat of harm to *B* from *A* since, amongst other things, *A* has not planned an attack on *B* and *A* did not yet know about the affair. So to prevent legitimising *B* killing *A*, we must ensure that the imminence requirement forms part of our understanding of what constitutes a justifiable action (or, for our purposes, a justifiable intervention to prevent harm). This is an interesting claim; does this mean that the threshold of physical harm should include an imminence requirement?

Let us look at how we might respond to such a claim[121]. Fritz Allhoff (2019: 1547) confronts Ferzan's argument and explains: 'my move would generally be to simply deny that necessity is likely to be satisfied in the sorts of cases she [Ferzan] envisions'. It was simply not

[121] It is worth noting that Baron (2011) responds to Ferzan in her paper. Baron (2011: 242–245) discusses Ferzan's (2004: 241–242) argument related to the justifiability of using deadly force in a kidnapping scenario (Robinson's hypothetical). Baron argues that '[w]hat makes it permissible to use lethal force in cases such as Robinson's hypothetical is primarily the threat of death or serious bodily harm, not [as Ferzan argues] the kidnapping, and therefore his hypothetical lends very strong support to the position that imminence should not be crucial to a self-defense claim' (Baron, 2011: 245). However, as Robinson's hypothetical is largely tangential to the focus of this section, this response will not be discussed any further.

necessary for *B* to kill *A now*, since *B* could, for example, have moved to another country, sought a restraining order against *A*, or any other course of action apart from killing *A* that would have ensured that *B* would not be killed by *A*—and, importantly, 'any of these possibilities undermines necessity' (Allhoff, 2019: 1547). So even in the example provided by Ferzan, we don't need to look to imminence since acting to avert the threat of harm was simply not necessary. But what of the case of Judy Norman? In Allhoff's (2019: 1541, 1543) own words, 'Norman is the hardest case' since she could have left the house rather than kill her husband; she even ensured that her grandchild would not be in the house since she took the child to a neighbour's house before killing JT. This case highlights that we must draw a distinction between physical restrictions (e.g. not being physically able to flee) and psychological restrictions (e.g. 'where, even if there was a doorway leading out of the house, she felt that there was no escape from an abusive relationship' (Allhoff, 2019: 1543)). We must therefore read Allhoff's objection to Ferzan as a general objection against the fact that some cases, including cases provided by Ferzan, do not undermine necessity since an agent was neither physically nor psychologically restricted yet chose to harm unnecessarily. Furthermore, even if Ferzan asked us to assume, for sake of argument, that it simply *was* necessary for *B* to kill *A* now, I agree with Allhoff (2019: 1547) that we should 'accept the consequences and maintain, against her [Ferzan's] intuitions, that B would be justified in killing A now […] because otherwise B is assuredly dead at A's hands'. We need to avoid a situation in which *B* must twiddle his thumbs and wait for death. What we have are therefore cases, even practical cases (c.f. Judy Norman), in which it seems that defensive action (e.g. an intervention) is justified in cases where there is no imminent threat. I agree with Allhoff (2019: 1548) that, in a case such as Affair, 'we should not say that the defender must wait for a certain, but future, death'—requiring that a threat be imminent is simply too restrictive. In short, 'imminence therefore matters—at least in practice—because it helps us

see when self-defense is necessary' (Allhoff, 2019: 1546). As a threat fades, it becomes less likely that a necessity requirement (discussed shortly) will be fulfilled since the increasing distance of the threat affords new prospects of various interventions that would avert that threat. The upshot of this is that the concepts of imminence and necessity 'can come apart and, when they do, imminence yields to necessity; the dual requirement simply does not make sense' (Allhoff, 2019: 1546).

So should we jettison the imminence requirement when establishing the threshold of physical harm? Yes, but with caveats—we should be careful to not throw the baby out with the bathwater. Although there are those cases in which the imminence of the threat of harm seems to matter to a decision to intervene to prevent that harm, the fact that there are those cases—and practical cases at that—in which there is no imminent threat of harm to a Victim but where an intervention is necessary to avert harm to that Victim (and where no other non-harmful interventions are available), we cannot make a requirement that imminence forms an integral part of the threshold of physical harm in and of itself, divorced from other considerations such as necessity. In other words, although imminence is certainly a consideration in the legal system for ascertaining the justifiability of defensive harm (c.f. Ferzan, 2004), including the imminence requirement would be too philosophically restrictive in the threshold of physical harm. Cases of battered women as presented in the literature often involve no imminent threat of harm, yet there remains the necessity of acting (very often by using lethal force) to avert any future harm from their husbands. So whilst I acknowledge that imminence is a philosophically interesting and rich consideration—and might often form part of a larger decision-making process beyond the boundaries discussed here and outside of the scope of the threshold of physical harm—we cannot require that the threshold of physical harm include the stipulation that a threat of harm must be imminent

for an intervention to be justifiable. We can therefore keep the baby (and orphan it to the legal system) whilst throwing out the bathwater.

Jettisoning imminence as a consideration in establishing the threshold of physical harm (in the way outlined above) tallies with the previous discussion in §4.5. on why establishing the probability of harm occurring to a Victim only gets us so far. If we were to require that the threshold include the imminence requirement, that threat of harm to a Victim is imminent, then we would be stuck back in (a) attempting to establish the probability, in numerical terms, of the risk of harm to a Victim, and (b) ascertaining the threshold in terms of the percentage likelihood of harm that would constitute harm being imminent—would we say that a 51% chance of harm means that harm is imminent, or a 75% chance, or a 95% chance? And if one of these percentages, why not one percent below that? One might argue that *x* is simply imminent when harm to a Victim is certain (a 100% chance of harm). But this would be *too* restrictive for the threshold of physical harm, since we want to be able to prevent harm without having to needlessly wait too long (c.f. Wait and See 1). And, further problematically, is seems like such an interpretation (where imminence requires that harm to a Victim be certain) falls short of being able to explain why we intuitively want to say that Norman was justified in killing her husband—after all, at the moment she killed him, there was a 0% chance that he was about to harm her since he was sleeping. I therefore propose, in line with Baron (2011), that we should jettison the imminence requirement and instead look to the necessity requirement to help explain our intuitions in the case of Norman and to help establish a philosophically informed threshold of physical harm.

I wish to anticipate a possible objection. One might argue that I want to have my cake and eat it since I claim both that the probability of harm to a Victim eventuating is an important

factor in the threshold of physical harm (via the primary and secondary narratives) and that the same threshold need not require that the harm be imminent. However, imminence and the probability (or risk) or harm eventuating are separate considerations. It is quite plausible that harm *h* has a 30% chance of eventuating to Victim, even though *h* is not imminent. For instance, imagine that a small but heavy boulder is rolling down the road towards Victim. The boulder is rolling so slowly that it will take over a day to reach Victim. So although there is a 30% chance that this boulder will harm Victim, the threat is not imminent. This highlights how we can differentiate imminence from the chance of harm eventuating. But wait! Surely the 30% chance of harm eventuating has factored-in the proximity of harm to Victim, and so the risk of harm to Victim necessarily incorporates whether and the extent to which harm is imminent? Admittedly, the probability of harm eventuating to Victim might include whether and the extent to which harm to Victim is imminent, and this might form part of the primary narrative. But the primary narrative does not require that we focus on whether and the extent to which harm is imminent, rather it simply involves an assessment of past sequence-events and an anticipation of how future sequence-events might unfold. This is precisely the reason I do not want to *fully* jettison imminence from consideration; the imminence requirement just cannot be a distinct requirement or claim in establishing the threshold of physical harm (for the reasons outlined above). So although one might be partially justified in claiming that I wish to have my cake and eat it, it is rather that I want to respect that imminence plays a role in some but not all decisions to intervene, and so although imminence might play a role in the primary narrative, the imminence requirement does not form part of the tertiary narrative.

There is also an upshot to jettisoning imminence in this way; it can help to address a concern raised by Baron (2011: 266): ‘If the danger is *not* imminent it may be appropriate to expect

more careful thought as to just how one might be able to escape the threat without resorting to violence'. This seems true, and something that the threshold of physical harm can (and will, over the course of the remainder of this chapter) capture. In those cases where a threat of harm to a Victim is imminent, the primary (and secondary) narrative might be such that a quick response is required. As I've already argued, this doesn't necessarily mean that the decision to intervene will be less reasoned or less justified, but in those cases where threat of harm to a Victim is not imminent, the luxury of more time before harm becomes imminent will enable more time to engage in the moral decision-making procedure which could ensure that more evidence is taken into account. This may not yield a more reasoned decision, but the possibility of utilising evidence that would be unavailable or otherwise unnoticeable in more imminent cases of threat would likely make the decision more *accurate* with respect to ascertaining the threshold of physical harm.

Let us now turn attention to understanding the concept of necessity (separated from the issue of imminence). Simply put, the necessity claim tracks a dominant principle in the self-and other-defence literature that a Deliberator should only resort to installing barrier *b1* if the harm *h1* caused to Initiator/Victim/Bystander as a result of installing *b1* is necessary to avert, mitigate, or diminish the harm *h2* that Victim would sustain without *b1*. There are two main considerations discussed in the literature that deal with what this principle should look like. The first claims that *b1* should be enacted/implemented only if *b1* averts, mitigates, or diminishes *h2* in a way that some other available barrier *b2* that causes less harm *h3* would not (*the least harmful claim / the lesser evil claim*). The second claims that *b1* is only justified if *h1* is a 'last resort' (*the last resort claim*). Let us assess each in turn.

In relation to the least harmful claim, Helen Frowe (2016: 12) describes such a consideration as a claim that 'forbids using more force than one has to in the course of defending oneself'[122]. Here, the (least harmful) necessity claim requires that where a Victim could either act in a way that causes harm *h1* (say, death) or in a way that would cause (less) harm *h2* (say, injury) in order to defend themselves against an aggressor, they should choose to install barrier *b2* since it is least harmful—this is why it is sometimes called the lesser evil claim, since Victim acts to bring about the lesser evil. We can restate the necessity claim so that it fits with moral sequencing thus: where Intervener has the option of installing either barrier *b1* or *b2* at the expense of agent *A* (be that themselves, Bystander, or any other intra-sequence agent), where *b1* results in the death of *A* and *b2* results in *A* being maimed, Intervener should install *b2* since it is least harmful. Put simply, the barrier installed by an Intervener should be the (available) barrier that causes the least harm—killing *A* is unnecessary since harm to Victim can be averted by only maiming *A*. In relation to the last resort claim, Frowe (2016: 13) describes this as a claim on why '[o]ne should use force only if one has no other option'. It is this sort of claim that underpins much of the legal system's insistence (at least in the UK) on an agent retreating from an aggressor and only if retreat is unavailable or futile should that agent inflict harm on the aggressor[123,124]. Although there

---

[122] An issue, briefly mentioned before, is that the literature is often primarily concerned with resorting to the use of force, and often lethal force. I have taken a broader view of the harm in question since moral sequencing is concerned not just with the most lethal kinds of force but a range of harm that causes physical harm to a Victim.

[123] There are a number of related issues, including the 'castle exception' which questions whether an agent is required to retreat from their own home, and further complications arising such as retreat requirements where the aggressor co-habits with the victim (e.g. Ferzan, 2004: 223, 232; Allhoff, 2019: 1534–1535). Frowe (2016: 13) provides a brief example via the case of the British man Tony Martin who shot two burglars, killing one; for a good discussion on this, see Squires (2006) and Burnside (2001).

[124] As an aside, empirical studies interestingly highlight how taking a life in self-defence as a last resort is an intuitively morally acceptable course of action (Robinson and Kurzban, 2007: 1843) (although some authors, such as Kaufman (2009), disagree with this intuition).

are subtle differences between the two claims, and although these are sometimes discussed apart from each other (for instance, in the just war and wartime ethics literature, the least harmful claim is often discussed in relation to *jus in bello* whereas the last resort claim is often discussed in relation to *jus ad bellum*), for current purposes (related to our enquiry concerning other-defence and not wartime ethics) we can understand the necessity claim as encompassing both the least harmful and the last resort claims where it is necessary for Intervener to install barrier *b* to harm *h* (*viz.* intervene to avert harm to Victim) only if no other non-harmful or less-harmful *h** barrier *b** was available to install (*viz.* if the Intervener could not have acted in a non- or less-harmful way to avert harm to Victim).

Seth Lazar (2012; 2016) offers a rich and philosophically useful understanding of necessity (that is not concerned with demarcating the least harmful and last resort claims), and argues that '[w]here an option *O* aims to avert a threat *T*, we determine *O*'s necessity by comparing it with all the other options that will either mitigate or avert *T*' (Lazar, 2016). A more fleshed-out version of this, and what I propose we adopt as the necessity claim, is that 'a self-[/other-]defensive harm H is necessary to avert the threat T if and only if the expected reduction in risk to the prospective victims of T [*viz.* Victim] outweighs the expected marginal morally weighted harms that Defender [*viz.* Intervener] inflicts on others [*viz.* Victim, Initiator, or Bystander]' (Lazar, 2012: 44) (also see Lazar, 2012: 13)[125]. In other words, it is necessary for Intervener to install barrier *b* to avert harm *h* to Victim only if the expected reduction in the risk of *h* occurring to Victim is greater than the expected harm *h** that Intervener will bring to Initiator, Victim, or Bystander by installing *b*.

---

[125] I also agree with Lazar that employing his understanding of necessity is also useful since it tracks the ordinary use/meaning of 'necessity' (see Lazar, 2012: 14–15).

The reason that the necessity claim is required of the threshold of physical harm is to ensure that (harmful) interventions are enacted only in those circumstances where another non-harmful or less harmful intervention was not available. If the threshold of physical harm did not include the necessity claim then an intervention would be justified in those cases where an Intervener could have chosen a non-harmful or less harmful intervention (yet chose not to), and would, for instance, in Wait and See 2 (c.f. §4.5.), permit Deliberator to choose to act now and kill Aggressor rather than waiting for the chance that will later present itself to avert the same harm to Victim by breaking Aggressor's leg—in this case we would, I think fairly uncontentiously, want to say that if the Deliberator is furnished with evidence that harm to Victim can be averted later by breaking Aggressor's leg, Deliberator would not be justified in intervening by acting now and killing Aggressor.

But what *sort* of harm is justifiable? Although the necessity claim ensures that an Intervener chooses to enact a harmful intervention only if he must, and then enacts the least harmful intervention, this does not on its own ensure that the intervention is justifiable, for the least harmful intervention might be disproportionate to the threat that the intervention seeks to avert. Two questions arise. First, does the sort or severity of harm matter—that is, does the magnitude of the threatened harm need to be acknowledged and accounted for when deciding how to intervene and the justifiability of that intervention bringing with it certain levels of harm? Second, does the type or severity of harm that can be imposed differ based on, for instance, the type of agent sustaining defensive harm? Does it matter whether the agent sustaining defensive harm is an Initiator, Victim, or Bystander—that is, does the enacting of an Initiator-harming, Victim-injuring, or Bystander-harming intervention matter in terms of the types of harms that each agent is liable for? The second question relates to issues surrounding whether an agent is liable to be harmed (by defensive action, that is, by

an intervention); this will be discussed in §4.6.4.3. The first question relates to proportionality. The threshold of physical harm is currently missing the stipulation that harm must not only be necessary (in the ways outlined in this section) but the defensive harm inflicted via the intervention must be proportionate to the harm threatened to Victim. The next section will look to the philosophical literature to ascertain how we can understand such a proportionality claim before incorporating it into the threshold of physical harm.

4.6.4.2. PROPORTIONALITY

Why is incorporating a proportionality claim into the threshold of physical harm required? Think back to the discussion of Taurek earlier in this chapter concerning whether the number of people harmed should count in a decision concerning which party to save. Although I argued that such a consideration is not directly relevant to moral sequencing (since moral sequencing is not concerned with deciding who to save but rather whether and in what ways an intervention should be enacted to save a Victim), it does introduce the concept of proportionality nicely and can factor into the decision to intervene. For example, when confronted with the choice to submit to a threatened harm from an aggressor that will surely kill you, or kill 10 people but thereby save yourself, it seems *ceteris paribus* disproportionate to opt for the latter, even though this ensures your demise and even though killing those 10 people is necessary to avert harm to yourself. But proportionality does not just have to be a numbers game—it need not solely concern itself with lives saved, gained, or lost. Indeed many discussions of proportionality in the literature deal not with the *number* of people harmed but the *kind* of harm or the *severity* of harm inflicted or endured. In other words, what is at stake is not how many agents will be harmed but the ways in which agents will be harmed. Although my killing an aggressor might be necessary to save me from receiving

a scratch (see the Scratch cases in §4.5.), it would seem *prima facie* disproportionate for me to kill the aggressor to avert that scratch. Likewise it would seem *prima facie* disproportionate, even if it is necessary, to kill the aggressor to prevent them from chopping off my leg. The proportionality claim is therefore important since it helps to account for whether necessary harm is also proportionate to the harm threatened, and this can and should be used to establish the threshold of physical harm.

Let us start by continuing the discussion of the last section. Many authors argue that necessity and proportionality are linked. For instance: Ferzan (2004: 222) states that the 'necessity component has two parts: the actor may only act *when* necessary and *to the extent* necessary. The latter requirement is one of proportionality'; Lazar (2012: 17–23) also argues that there is a 'deep connection between necessity and proportionality' to the extent that necessity 'entails proportionality' (Lazar, 2012: 44); and Rodin (2011: 109) argues that liability and the lesser evil justification 'share much in common' and are 'complex forms of proportionality relationship between a shared set of underlying factors'. Although I do not agree with some of these arguments, it does seem fairly intuitive that in claiming that, in line with Ferzan's (2004) claim, the extent to which an intervention was necessary is itself a proportionality claim; the extent to which defensive harm *d* is necessary to avert threatened harm *t* is that *d* is proportionate to *t*. But let us not get ahead of ourselves.

The proportionality claim can be loosely understood as a requirement that 'the harm I inflict upon my attacker must not significantly outweigh the good that I hope to secure thereby' (Frowe, 2016: 11). Many authors have discussed the number of considerations relevant to making judgements on the proportionality of certain acts. David Rodin (2011), for example, identifies fourteen considerations relevant to deciding if/whether an act is/was (narrowly)

proportional, whereas Quong (2015: 149) outlines four conditions that require satisfying by an account of proportionality[126] (and further outlines how McMahan's (2005; 2009a; 2011; 2017a; 2017b) 'responsibility account' (discussed shortly) fulfils his conditions (Quong, 2015: 149ff.)).

Quong (2015) argues that many authors seem to endorse the idea that whether harm to an agent *A* is proportionate is dependent on the extent to which *A* is '*morally responsible* for creating a situation where someone suffering harm is unavoidable', and so the level of defensive harm that is proportionate to inflict on *A* is dependent on and is often equivalent to the level of moral responsibility that is or can be attributed to *A* (Quong, 2015: 145). Quong (2015) rejects this idea and instead offers a 'positive account' of proportionality, what he calls 'the stringency principle', whereby the stringency of the right that an aggressor threatens to infringe determines whether a certain defensive harm was proportionate. In Quong's (2015: 145) words: 'The more stringent the right that is threatened, the greater the degree of defensive harm that is proportionate'. In a case where the magnitude of (other-) defensive harm outweighs the magnitude of threatened harm (to a Victim), 'the harm you impose on him far exceeds the stringency of the right of [Victim] that he threatens to violate' (Quong, 2015: 146) and thus the response is not proportionate. Although Quong's account

---

[126] Quong (2015: 149) states that 'a successful account of proportionality should, ideally, do all of the following:

1. Explain our intuitive judgments about proportionality in paradigm cases.
2. Be sensitive to the way multiple considerations bear on proportionality judgments, and not misrepresent or distort the way these different considerations matter for proportionality.
3. Offer a coherent framework that unifies these different considerations, explaining why *these* considerations, rather than others, belong together in an account of proportionality in defensive harm.
4. Explain the relationship between (a) the necessary and sufficient conditions for liability to defensive harm, and (b) the considerations that determine how much harm a person is liable to bear.'

of proportionality is appealing—in that it tracks an understanding that an increased severity of threatened harm to a stringent right (e.g. the negative right to not be killed) justifies an increased response proportionate to the stringency of that right (this justifying defensive action including killing the aggressor that jeopardises your negative right)—it does not differentiate between the type of agent that is, might, or will be the subject of defensive harm. In other words, it does not differentiate between, for example, Initiator-harming interventions (which threaten defensive harm against an Initiator), Victim-injuring interventions (which threaten defensive harm against a Victim), or Bystander-harming interventions (which threaten defensive harm against a Bystander). Why is this important? Because, as we shall see in the next section, and without intending to jump ahead, whether an intervention is proportionate seems tied-up with whether and the extent to which an agent is liable to (a level of) defensive harm.

Jeff McMahan offers what Rodin (2011: 78) calls an 'at once overly simple and overly complex' distinction between different forms of proportionality (McMahan, 2009a: 20–21; McMahan, 2017a: 135–138). The first, *wide proportionality*, is 'proportionality in harms caused to people who are not liable to those harms […] [in] that they have done nothing to forfeit their right not to be caused those harms' (McMahan, 2015: 2). This is, essentially, a lesser-evil justification for harming (McMahan, 2017a: 135); agents who are not liable to be harmed are wronged when harming them results in less harm to other agents who are also not liable to be harmed. In other words, wide proportionality is concerned with whether defensive harm (via, for example, an intervention) is justified by accounting for the harms that might be caused to non-liable (intra-sequence) agents by the intervention (e.g. harm caused to Bystander). McMahan's second and contrasting type of proportionality is *narrow proportionality*, which is 'proportionality in harm inflicted on people who are potentially

liable to be harmed […] [meaning that an agent] who poses a threat of wrongful harm may be potentially liable to be caused some degree of harm in defense of his victim' (McMahan, 2015: 2–3). This is, essentially, a liability-based justification for harming (McMahan, 2017a: 135); an agent is wronged when harm that comes to that agent exceeds the harm for which that agent is liable[127]. Lazar (2018: 863) explains why McMahan's account of wide proportionality is useful: 'Although N's lacking the protection of his right to life is sufficient to ensure that he is not wronged by being killed, it does not ensure that killing him is objectively permissible. For there might be other adverse consequences of doing so, which rule it out. Suppose, for example, that killing N would also kill a large number of bystanders. Or suppose N, and only N, knows the cure for cancer. Killing in such cases might be 'widely disproportionate''.

[127] Interestingly, McMahan also mentions a third form of proportionality, namely 'proportionality in the aggregate' (McMahan, 2017a: 142–143; also see McMahan, 2017b), which accounts for the number of wrongful aggressors as an aggregate. McMahan (2017a: 142) uses the example of the 'intuitively clear' belief that 'killing half a million minimally culpable combatants would have been disproportionate in the aggregate in relation to the importance of preserving British sovereignty over the Falkland Islands'. McMahan (2017a: 143) claims that '[a]lthough killing all half-million combatants would have made the killing of each effective and narrowly proportionate, it would nevertheless have been disproportionate in the aggregate'. However, as moral sequencing is concerned only with harm occurring to the Victim of a moral sequence (and not aggregates), this will not be discussed further.

The difference between narrow and wide proportionality can be cashed-out in a different way. Rodin (2011: 78) offers four distinct forms of proportionality:

> '1. Acts that intentionally harm those who are potentially liable to be harmed […]
>
> 2. Acts that unintentionally but foreseeably harm those who are potentially liable.
>
> 3. Acts that intentionally harm those who are not liable.
>
> 4. Acts that unintentionally but foreseeably harm those who are not liable to be harmed […]'.

Rodin (2011: 78) claims that 1. and 2. align with McMahan's concept of narrow proportionality, whilst 3. and 4. align with wide proportionality (and Rodin further claims that the former two are relevant to liability justifications and the latter two are relevant to lesser evil justifications). Laying out the forms of proportionality in this way draws attention to two of Rodin's claims: first, that 'directly intending harm is harder to justify and therefore subject to a more demanding proportionality constraint than harm that is foreseen but unintended'; and second, 'harm brought about through something we do […] is, other things being equal, more difficult to justify than harm brought about by something we allow to happen' (Rodin, 2011: 78) (also see McMahan, 2017a: 152–153 for his discussion of wide proportionality in relation to the distinction between doing harm and allowing harm; and see chapter 2 for a discussion of the distinction between doing harm and allowing harm, specifically §1.2.6. which discusses the concepts of positive and negative agency).

Since the concepts of narrow and wide proportionality rely on an understanding of liability (the definitions of each employ liability too), I propose that we fast-forward to the next

section which discusses liability. That said, and before doing so, we can make some brief preliminary comments on the proportionality claim in relation to Initiator-harming, Victim-injuring, and Bystander-harming interventions. Using a pre-theoretical understanding of liability, Initiator-harming interventions seem to best fit narrow proportionality whereas Victim-injuring and Bystander-harming interventions seem to best fit wide proportionality. This is because only Initiator seems liable to be harmed since they initiated the NPET against Victim and thus (unjustly) threatened harm to Victim. Neither Victim nor Bystander seem liable to be harmed (by defensive action, e.g. an intervention) since neither have threatened harm against another agent. However, it seems that there is also a *prima facie* difference between Victim and Bystander here; although Victim might be harmed by an Intervener (and thus harmed through Intervener's defensive action), harm caused to Victim here likely involves a least harm / lesser evil justification such that the defensive harm would cause less harm to Victim than if the original threatened harm had eventuated (the defensive harm might, for example, involve shooting Victim in the kneecap to avert death to Victim). In this way, Victim stands to gain by the intervention even though they are not, it seems, liable to be harmed. However, the same cannot be said for Bystander, who does not stand to gain anything by being used in an intervention to avert/mitigate harm to Victim. Bystander is the ultimate loser; they stand to lose something (in the case of a Bystander-injuring intervention) or stand to lose everything (in the case of a Bystander-sacrificial intervention), and stand to gain nothing.

Put another way, it is possible that we are presented with a case in which defensive harm to an agent to avert/mitigate harm to a Victim is both necessary and proportionate, but where we might want to say that enacting that defensive harm is not justifiable. Consider the following:

Cannibal

Jill is stuck on a barren desert island with Jack. Both are starving and will not survive another day without food. Using the last bit of battery on their radio, they receive a message that they will be rescued in several days. Jill kills Jack and eats him, thus enabling Jill to survive until she is rescued.

In Cannibal, is seems that Jill's action is both necessary and proportionate—eating Jack is necessary since it is the only way Jill can survive and eating Jack is proportionate to the threatened harm since without acting Jill will die too. However, it is intuitively uncomfortable to suggest that Jill murdering and cannibalising Jack is justified. What is missing from this case, and what is missing from our understanding of the relevant considerations for the threshold of physical harm, is an understanding of how and in what ways an agent's liability has bearing on the justifiability of defensive harm. The next section will outline liability and how a liability claim forms part of the threshold of physical harm.

4.6.4.3. LIABILITY

It is appropriate to start with some clarificatory and scoping remarks on liability before assessing the concept of liability itself. It is important to recognise that the concepts of liability [128] and (moral) desert are separate and should be kept distinct for current

[128] I wish to make the same stipulation as McMahan (2005: 386) regarding the use of the term 'liability': 'Although I borrow the notion of liability from legal theory […] my concern in this article [here, chapter] is with moral rather than legal liability'.

purposes[129]. An agent might *deserve* to be killed if 'there is a reason to kill her even if it is possible for no one to be killed', whereas an agent might be said to be *liable* to be harmed if 'the person to be killed [or otherwise harmed] has acted in such a way that to kill [or otherwise harm] him would neither wrong him nor violate his right' (McMahan, 2005: 386); Jonathan Quong (2015: 146–147) also makes the same point. Quong (2015: 147) also notes that liability should not be confused with moral permissibility; whether an agent is liable to defensive harm is a relevant consideration when determining if it is morally permissible to harm that agent, although liability cannot determine whether it is morally permissible to impose that harm.

The concept of liability also seems to have various relationships with necessity and proportionality. For instance, Quong (2015: 144–145) argues that there is a nuanced relationship between liability and proportionality: 'a person is never liable to defensive harm generally; rather, a person can only be liable to some particular proportionate level of defensive harm'. What's more, some authors draw a link between liability and necessity: Firth and Quong (2012) offer a pluralist account of liability in which there is a relationship between necessity and liability; Rodin (2011: 74) offers 'a general explanatory model of the liability and lesser evil justification [*viz.* necessity] of harm'; Kai Draper (2016: 178) links necessity with (amongst other things) 'greater enforcement costs for at least one nonliable party'; and McMahan (2009a: 10ff.) claims that 'necessity is internal to liability'—although not all self-defence theorists believe this (see Frowe, 2014; Firth and Quong, 2012). Moreover, as we have seen, McMahan's (2009a; 2015; 2017a; 2017b) account of narrow

[129] That said, the two can be linked if one believes or can demonstrate that an agent's liability to *x* in some way implies or entails desert to *x* (for example, desert by way of punishment). However, the issue of punishment will not be discussed.

and wide proportionality incorporates liability. Other authors agree that the concept of liability cannot be separated from the consideration of proportionality. Rodin (2011) argues that an agent's liability is tied-up in narrow proportionality; that an agent is liable to harm involves that harm being narrowly proportionate. Rodin (2011: 79) neatly summarises his position thus: 'For a person to be liable to a harm, just is for that harm to be narrowly proportionate in the circumstances. Proportionality and liability, far from being independent factors, are two manifestations of the same underlying normative relations'. With that in mind, let us now focus on the concept of liability.

In short, what we are trying to ascertain when looking to liability is *who* (that is, which, if any, intra-sequence agent) is liable to defensive harm (from Intervener in cases of other-defence, but also from Victim in cases of self-defence). This ultimately helps with the concepts of necessity and proportionality too, since necessity seeks to ascertain whether it is *necessary* to inflict defensive harm and proportionality seeks to determine the *magnitude* of harm they are liable to receive.

But what, exactly, makes an agent liable to (self- or other-) defensive harm? The best way to understand the concept of liability is to look at some of the accounts of liability from the literature. Let us focus on three of the most widely discussed accounts in turn before using that discussion to understand how the liability claim can be understood in the threshold of physical harm.

The Culpability Account

The culpability account of liability to defensive harm is popular amongst philosophers and legal scholars alike and, although much of the new literature seems to favour other

approaches, some of those authors started their journey as proponents of the culpability account (e.g. McMahan, 1994: 268–271). The term 'culpable' can be and has been defined and interpreted in a number of ways, but it can be loosely defined thus: agent *A* is culpable if *A* acts in a way *w* such that *w* threatens to harm another agent *B* and where this threat of harm is unjustified—this follows Alexander and Ferzan's (2008: 378–379) understanding of culpability. The culpability account therefore states that agent *A* is liable to defensive harm[130] *h* if *A* is culpable for unjustly threatening harm against a Victim. To elucidate this account, consider the following case, adapted from Thomson (1991: 283):

> *Truck*
>
> Micah is a pedestrian walking on the pavement down a deserted road. Villain is hell-bent on running over Micah and drives towards them. There is nowhere that Micah can go to hide from or avoid Villain. It just so happens that Micah is carrying a bazooka which they can use to destroy the truck, killing Villain, and thereby saving their own life.

Here we can call Villain a Culpable Aggressor[131] since it seems intuitively justifiable for Micah to destroy the truck and kill Villain. Since Villain poses an unjustified threat of harm to Micah, Micah is justified in using defensive force to thwart Villain's attempts to harm

---

[130] This might sound like an odd turn of phrase. However, to clarify, to say that an agent *A* is liable to defensive harm is simply to say that *A* has acted in a way that means that *A* is now liable to be harmed by another agent *B* whom *A*'s actions have caused *B* to be under a threat of harm and where *B* now seeks to harm *A* in self-defence to avert the harm that *A* has threatened to *B*. Agent *C* can also act in defence of *B*; this is an act of other-defence to avert the harm that *A* has threatened to *B*.

[131] As I present these cases and attribute a term to the type of agent under consideration, here Culpable Aggressor, I am aware that other authors sometimes use different terms, however I present and utilise these terms consistent with Frowe's (2011) usage.

them (Doggett, 2011; Ferzan, 2005). On the culpability account, 'it is Villain's malicious intention [to harm Micah] that makes it permissible to kill him (along with the necessity and proportionality of doing so)' (Frowe, 2011: 14; McMahan, 2005: 394 also agrees). This is a case of *intentionally* threatening harm, and so Micah's response here seems to be narrowly proportionate to the threat of harm—although this is not to say that the culpability account only involves cases of intentional harming. Even if Villain did not intend to harm Micah, let us say Villain took some hallucinogenic drugs that impaired his ability to drive, we would still likely believe that this makes Villain culpable and that Micah was justified in killing Villain. Importantly, the culpability account can also explain why Villain is not entitled to defend themselves against Micah's defensive harm—Villain is not justified in, say, speeding up to ensure that they hit and kill Micah before Micah can shoot the bazooka. In short, it can help to explain 'our intuition that should you try to defend yourself against Villain, Villain cannot then invoke a right of self-defence against *you*' (Frowe, 2011: 14). The fact that Micah is innocent protects them from retaliatory harm from Villain. However, the culpability account faces some challenges. Consider the following example, adapted from Fletcher (1973: 371):

Car

Micah and Stranger are passengers of a car, sharing a taxi ride into town. Stranger is schizophrenic and, without any fault on their part[132], suffers a schizophrenic episode and attacks Micah. If Stranger is not stopped, they will certainly kill Micah. Micah can only stop Stranger by killing them[133].

Like in Truck, it seems intuitive that Micah would be justified in killing Stranger in Car. Here, Stranger is what is often called an Innocent Threat (since they are seemingly not responsible for the harm they cause). Stranger is therefore not culpable (they are not a Culpable Aggressor) but we would still likely think that Micah is justified in killing Stranger—therein lies the problem for the culpability account. Also, here, Micah's killing Stranger seems widely proportionate. This has led some authors to accept that the culpability account therefore prohibits Micah from killing Stranger, whilst others see this as proof that the account is too restrictive (Leverick, 2006: 44; McMahan, 2014: 113). Other authors have suggested that cases involving Innocent Threat point to the need to move away from culpability and towards other considerations that can explain why Micah would be justified in killing Stranger. Simply put, because the culpability account cannot be employed to justify the intuition that Micah should be permitted to kill Stranger, we should look elsewhere. One suggestion is that, similar to what Taurek (1977) argues about prioritising oneself (see §4.6.1.1. for a discussion of this), Micah could kill Stranger in Car if we permit

---

[132] We must assume that Stranger has not has a schizophrenic episode before and could not have otherwise been aware of this episode occurring.

[133] I am aware that this example is insensitive and in no way reflective of genuine 'schizophrenic' episodes. I ask the reader to forgive its inclusion on the grounds that this is an important example from the literature.

that greater value is placed on the preservation of Micah's life than on the Stranger's life (Alexander and Ferzan, 2009: 136–141; Ferzan, 2012: 685; Ferzan, 2016: 221; Alexander, 2016: 26–28). On this view, the 'personal prerogative' of the Victim (e.g. Micah) outweighs the value of the Innocent Threat's (e.g. Stranger's) life. However, the problems are not yet over for the culpability account. Consider the following case, adapted from McMahan (2005: 393):

> *Vehicle*
>
> Conscientious Driver keeps their car well maintained and always drives cautiously and alertly. On one occasion, however, freak circumstances cause the car to go out of control. The car veers in the direction of Micah whom it will kill unless Micah blows it up by using their bazooka.

Vehicle presents a problem for the culpability account since although we would likely think that Micah is not justified in killing Stranger in Car, we would likely intuitively claim that Micah is justified in killing Conscientious Driver in Vehicle since Conscientious Driver has at least some responsibility (unlike Stranger) for the threat of harm to Micah *even though* Conscientious Driver is not a Culpable Aggressor (since they were not acting maliciously). The culpability account would therefore require that Micah cannot kill Conscientious Driver. The issue, then, is that if we want to justify Micah's killing of Stranger in Car and Conscientious Driver in Vehicle then we need to look to another account of liability to defensive harm. The culpability account simply cannot ground claims of liability to defensive harm in the ways that seem most intuitive.

The Rights-Based (or Rights-Forfeiture) Account

The rights-based account of liability to defensive harm centres on the claim that agents have a right to not be harmed. We can link this to the discussions in §1.2.4. and §4.1. which discuss the concept and importance of negative rights. So on this account, if we wish to state that an agent can be harmed then we must explain why their negative right to not be harmed has been forfeited and why harming them does not violate their negative right. On this account, '[t]he reason the victim is permitted to kill the aggressor, but the aggressor is not permitted to kill the victim, is that the aggressor, by virtue of her conduct in becoming an unjust immediate threat to the life of the victim that cannot be avoided by any less harmful means, forfeits her right to life' (Leverick, 2006: 66)[134]. Jonathan Quong (2012: 45) neatly summarises a general rights-based account, stating that you as an agent are liable to defensive harm when you have 'forfeited your rights against the harm being imposed, and thus you are not wronged if the harm is imposed on you'. This echoes what is arguably one of the most well-known rights-based accounts, that of Judith Jarvis Thomson (1991), who argues that an agent's right to not be killed is a 'claim right'; claim rights are generally held rights, such as my claim right to not be harmed, and go both ways: claim right *r* requires that other agents do not harm me and that other agents offer defensive action to thwart an unjust attack against me from an aggressor. In other words, 'other things being equal, every person Y has a right against X that X not kill [/harm] Y' (Thomson, 1991: 299). *Y*'s right to not be harmed is what justifies *Y* in harming *X* as a defensive action; *X* is liable to defensive harm only if *X* has forfeited their right to not be harmed by *Y* by virtue of the fact that *X* will otherwise violate *Y*'s right to not be harmed. Let us turn attention back to the three cases discussed above. First, Micah would be justified in killing Villain in Truck since

---

[134] Leverick (2006: 67–68) argues that the rights-based account 'is the most convincing account' and outlines five key advantages that it has over other accounts.

Villain has forfeited their right to not be killed and so Micah's killing Villain does not violate Villain's rights—and further Villain is unable to use defensive force against Micah. The rights-based account can therefore justify the defensive force against Culpable Aggressor. Second, Micah would be justified in killing Stranger in Car since Stranger does not have a right to kill Micah and so Stranger stands to violate Micah's right if they were to succeed in killing Micah; Stranger thus forfeits their right to not be killed by Micah. The rights-based account can therefore also justify the use of defensive force against Innocent Threat. Third, Micah would be justified in killing Conscientious Driver in Vehicle since Conscientious Driver does not have a right to kill Micah and so, to defend the violation of Micah's right, Micah can deploy defensive action against Conscientious Driver. The rights-based account can therefore overcome the issue inherent in the culpability account of not being able to take defensive action in Vehicle. Thomson's rights-based account therefore avoids the issue of culpability[135] whilst being able to make sense of our intuitions in the above cases[136]. However, Thomson's rights-based account inevitably runs into problems. Consider the following example, adapted from Nozick (1974: 34):

---

[135] It is also worth noting that Thomson's account avoids needing to account for the moral responsibility of the aggressor (Thomson, 1991: 287–289).
[136] Frowe (2011: 16–17) praises Thomson's account since it 'passes the test' Frowe sets out in the Hunger case (see Frowe, 2011: 16).

Well

> Falling Agent has been hoisted up into the air by a tornado and is on course to land in a well. Micah is taking refuge from the tornado in the well and is now trapped there. Unless Micah vaporises Falling Agent with a ray gun before he lands in the well, Falling Agent will crush Micah to death, but Micah's body will cushion his fall, thereby saving the life of Falling Agent.

Falling Agent is an Agency-Lacking Threat since they threaten harm to Micah but have not chosen to act in this way and is helpless to countervail their threat of harm to Micah and is powerless to act otherwise. In short, Falling Agent did not act, rather their body was moved by the tornado[137]. It is not the case that Falling Agent is trying, maliciously or otherwise, to kill Micah (unlike Villain in Truck and Stranger in Car). However, it cannot be the case that Falling Agent has a right to kill Micah, and so Falling Agent has a duty to not kill Micah. As a result, Falling Agent lacks the right to not be killed and Micah is therefore, according to Thomson's rights-based account, justified in killing Falling Agent (and for the same reasons that Micah is justified in killing Villain, Stranger, and Conscientious Driver). On this understanding, Falling Agent (an Agency-Lacking Threat) is liable to defensive harm. But this just seems counter-intuitive; even if Falling Agent will kill Micah, it seems absurd to claim that Falling Agent violates Micah's right by being dropped into the well, through no fault of their own, due to the tornado. Otsuka (1994) agrees and argues that the following

---

[137] It is for this reason that, as Frowe (2011: 17) points out, some authors (e.g. Otsuka, 1994; McMahan, 2002: 409; Rodin, 2002: 81–83) have compared Agency-Lacking Threats to animals and objects.

two cases (adapted from Otsuka (1994: 80)), highlight a problem with Thomson's rights-based account:

> *Stone*
>
> A stone is hoisted-up into the air by the force of a tornado and will land in a well in which Micah is taking refuge. Unless Micah uses a ray gun to vaporise the stone before it lands in the well, the velocity of the falling stone will kill Micah.

> *Unconscious*
>
> An Unconscious Agent hoisted-up into the air by the force of a tornado and will land in a well in which Micah is taking refuge. Unless Micah vaporises Unconscious Agent with a ray gun before he lands in the well, Unconscious Agent will crush Micah to death, but Micah's body will cushion his fall, thereby saving the life of Unconscious Agent.

Both Stone and Unconscious involve an Agency-Lacking Threat similar to Well, the only difference being that Well involves a conscious threat, Unconscious involves an unconscious threat, and Stone involves a non-conscious (and, realistically, non-agent) threat. Yet it seems intuitively wrong to deny that Unconscious Agent and the stone have violated Micah's rights. Otsuka (1994: 80) argues: 'I do not see how the rights-violating power of such a human object [in Unconscious] could be any greater than the rights-violating power of a chunk of granite [in Stone]'. In other words, although Thomson wants us to believe that Falling Agent in Well violates Micah's rights (and this grounds a justification for Micah to take defensive action against Falling Agent), the two parallel cases

of Stone and Unconscious involving similar Agency-Lacking Threats illustrate that we cannot consider Falling Agent to have violated Micah's rights for same reason that it seems absurd to claim that the stone and Unconscious Agent have violated Micah's rights in Stone and Unconscious respectively. This ultimately highlights how 'talk of rights violations has, I think, gone too far if it is based on a theory that implies that a falling stone can violate a human right' (Otsuka, 1994: 80). Otsuka (1994) therefore rejects Thomson's rights-based account and the conclusion that Agency-Lacking Threats are liable to defensive harm. However, although I agree with Otsuka that falling stones cannot violate an agent's rights, and we therefore have reason to challenge Thomson's account, it does in fact seem intuitive to claim that Micah should be justified in defending themselves against an Agency-Lacking Threat.

Can we rescue Thomson's account (in the sense that we surely do not want to simply shout down the well to Micah and declare that they must prepare for imminent death without recourse)? Rodin (2002) makes an attempt by considering a case of a stone rolling down a hill (Rodin, 2002: 86); he claims that even if we submit to Thomson's claim that the stone has no right to fall on Micah, Thomson is not entitled to infer that the stone has a duty to not fall on Micah since 'it is not a moral subject at all' (Rodin, 2002: 86). The same conclusion extends to Falling Agent and Unconscious Agent—neither have a duty to not fall on Micah. To paraphrase Rodin's (2002: 86) conclusion: the falling is not something either Falling Agent or Unconscious Agent does so their falling cannot be in violation of any duties he owes Micah. Hence, in crushing Micah, neither Falling Agent nor Unconscious Agent violates any of Micah's rights. Although Rodin's argument against Thomson seems to address the issue, it does not. As Frowe (2011: 18) points out, if we agree with Rodin's conclusion that the stone is not bound by rights and duties then 'if the reason

for this is that the stone is not a moral subject at all, this argument fails to show that [Falling Agent] cannot be the subject of rights and duties. After all, if she were really just like the falling stone, it would obviously be true that [Micah] is permitted to vaporise her, since vaporising stones is morally unproblematic'. The only way that Rodin can dodge such criticism is by claiming that Falling Agent 'is enough of a moral subject to have a right not to be killed', but then this would require understanding Falling Agent as 'a sort of quasi-moral subject, with a right not to be killed but no duty not to kill others'—this simply won't do.

Moreover, Thomson doesn't differentiate between cases in which agents intentionally threaten harm and cases in which agents unintentionally threaten harm to a Victim. This matters because, as Draper (2009: 72) argues, '[i]f you have a right against someone that he behave in a certain way, then surely it must be at least logically possible that he has a moral reason to behave in that way'. However, since it is not within Falling Agent's power to not fall down the well, they cannot have a duty to abstain from killing Micah, and so it doesn't make sense to claim that Micah has a right to defensive action against Falling Agent since Falling Agent's negative rights have not been forfeited.

Perhaps a larger problem is that Thomson's account does not permit Falling Agent to employ defensive action against Micah's defensive action against him. In other words, even if we agree that Micah is justified in killing Falling Agent, this does not preclude or prohibit Falling Agent from employing his *own* defensive actions against the threat of harm that Micah now poses to him. Agents in other cases, e.g. Truck, cannot permissibly retaliate against Micah's defensive actions since the Initiator, Villain in the case of Truck, has by virtue of their actions forfeited their right to not be harmed. However, as we have seen,

Falling Agent (and any other Agency-Lacking Threat) cannot be said to have forfeited their right to not be harmed since they are morally innocent of any threatened harm towards Micah, and this grounds the claim that Falling Agent should be justified in deploying their own defensive actions against Micah (should Micah choose to deploy defensive action against Falling Agent). However, Thomson's account prohibits Falling Agent from engaging in any defensive action of their own since they are, under the rights-based account, liable to be harmed.

It therefore seems that Thomson's rights-based account encounters a number of issues that prevent it being able to sufficiently deal with the issue of Agency-Lacking Threats, and this makes the rights-based account 'less appealing intuitively than it would be if it could justify the common-sense intuition' (McMahan, 2009b: 389). But what makes the account even *less* appealing is how it handles cases of Innocent Bystanders (simply called Bystanders in moral sequencing). Consider the following case, adapted from Thomson (1991: 290):

> *Railway*
>
> A runaway train is heading towards Micah, whose foot has become stuck in the tracks. Micah can only avert imminent death by pulling a nearby lever, causing the train to divert onto another track where Jules' foot is also stuck in that track. Diverting the trolley will kill Jules.

Jules is an Innocent Bystander, namely 'a person whose actions, movements, or presence do not endanger Victim [Micah]' (Frowe, 2014: 22). However, Jules has not violated Micah's right to not be harmed, Jules has not forfeited their right to not be harmed. Because Jules is not liable to be harmed, Micah therefore cannot deploy defensive actions against

Jules to divert the train (and thereby transfer imminent death to Jules). What is the difference between Railway and Falling Agent? Under Thomson's rights-based account, Micah is still bound by the duty to not harm Jules, since Jules poses no risk of harm to Micah. Jules' negative rights remain intact. Because of this Jules would be justified in deploying defensive action against Micah should Micah go to pull the lever. But under Thomson's account it won't get that far, since there is no scenario in which Micah can permissibly harm Jules even though Micah is in imminent danger of death. So whilst Thomson's rights-based account can justify defensive harm to Culpable Aggressors, Innocent Threats, and Agency-Lacking Threats, it cannot justify harming Innocent Bystanders—neither Jules nor any other Innocent Bystander is liable to harm.

The Responsibility Account

The responsibility account of liability to defensive harm refocuses the discussion back on a moral assessment of agents, like in the culpability account, but in a way that is less restrictive and in a way that offers an understanding of some as-yet problematic upshots of the other two accounts. Many authors have introduced a responsibility account of liability (Otsuka, 1994; Coady, 2004a; Coady, 2004b; Rodin, 2008; Frowe, 2014), although McMahan's (2005; 2009a; 2009b; 2011; 2014; 2017a; 2017b) is by far the most developed. McMahan's responsibility account states that for agent *A* to be liable to defensive harm (that is, harm from another agent *B* acting in self- or other-defence), *A* must have 'foreseeably imposed risks' (McMahan, 2005: 394) of non-negligible harm that causes 'significant unjust harm' (McMahan, 2005: 394) to a Victim. Liability to harm rests on ascertaining whether an agent has 'moral responsibility for an objectively unjustified threat of harm' (McMahan, 2009b: 392), which can be gauged by appealing to a sense of justice: 'we may think of liability to defensive action as a matter of preventive justice, or justice in the distribution of harm ex

ante' (McMahan, 2005: 395). So how does McMahan's account apportion responsibility in the cases discussed in relation to the culpability account and rights-based account above? In his own words (McMahan, 2005: 395):

> 'You know that if you drive you impose a very small risk on other innocent people. If you choose to drive, the consequences are your responsibility unless others also contribute to the outcome through their own risk-imposing activities. You will be liable to defensive action even if you satisfy the relevant standards of due care'.

To elucidate McMahan's claim, re-consider the Vehicle case (discussed in the culpability account above). Conscientious Driver chose to drive the car, regardless of the foreseeable chance that their driving of it could impose significant unjust risk on another agent, and their actions resulted in an unjust threat of harm to Micah. This furnishes Conscientious Driver with responsibility for the outcomes of their decision to drive the car, and so Conscientious Driver is liable to defensive harm from Micah. Conscientious Driver's responsibility means that Micah can take defensive action to ensure that Conscientious Driver bears harm for their actions and not Micah; and Micah, in return, is not liable to any defensive action from Conscientious Driver since Micah is not responsible for the threatened harm they are trying to avert[138]. But how, exactly, does McMahan's insistence on ascertaining liability in relation to whether an agent is morally responsible for the threatened harm to Victim differ from the

---

[138] Even though, in the strictest sense, Micah is now responsible for the harm, via their defensive action, that comes to Conscientious Driver—although the use of 'responsibility' here is not in line with McMahan's technical definition. This hints at an issue related to employing an everyday term and how this sits with technical uses of the term—this will be discussed more in chapter 5.

culpability account? The culpability account would say that Conscientious Driver is not liable to defensive harm since Conscientious Driver is not culpable for their actions (they took due care, etc. to ensure that they would not threaten harm by driving the car), and this sort of prohibition on imposing liability to defensive harm in these sorts of situations led us to challenge and ultimately reject the culpability account as a full account of liability to defensive harm. McMahan therefore denies that culpability is necessary for liability. So far so good—but how does McMahan's account deal with some of the other cases?

Villain in Truck is still, like on the other two accounts, liable to defensive harm—this is, I think, the most uncontentious of the cases. However one looks at the case, liability to defensive harm is justified—and, on McMahan's responsibility account, this is because Villain foresees the imposed risks of their driving and moreover intends to harm Micah and so imposed an unjust threat of harm to Micah that makes Villain morally responsible for the threatened harm to Micah. Micah is therefore justified in deploying defensive action against Villain without recourse from Villain (*viz.* Villain is not justified in taking defensive steps to avert Micah's defensive harm). Stranger in Car and Falling Agent in Well both pose an unjust threat of harm to Micah, but neither Stranger nor Falling Agent can be considered, on McMahan's account, to be morally responsible for their unjust threat of harm to Micah and so neither are liable to defensive harm from Micah. Problematically, then, Stranger and Falling Agent are permitted to harm Micah (under the considerations outlined in their respective cases) and cannot be said to have wronged Micah if Micah is harmed by them, but if Micah attempts to deploy defensive action against their threat of harm then Micah has wronged them (and Micah would in such a case be morally responsible for the apparently unjust threat of harm posed to or inflicted on them). McMahan's account also gets us no further in being able to potentially defend Micah's decision to divert the train to Jules in

Railway—Micah is stuck, it seems, with simply waiting for the impending train and imminent death since Jules is not morally responsible for the unjust threat of harm that the train poses to Micah. To those who wish to justify Micah's use of defensive force in Railway by diverting the train to Jules, McMahan (2009b: 389) simply says that 'it is our intuition that is mistaken, not the theory'. Interestingly, then, McMahan's responsibility account is more permissive than the culpability account but is more restrictive[139] than Thomson's rights-based account, and, moreover, seems to yield unintuitive results in Car and Well (and possible also in Railway, too).

Let us leave the discussion of the different accounts of liability to defensive harm here. There are, of course other accounts of liability that we could delve into—e.g. Cécile Fabre's (2009; 2012; 2014; 2016) partiality-based account[140]—but the drive of this section is not to canvass the whole literature, nor provide a literature review on the different accounts of liability. This section seeks only to (a) draw the reader's attention to an on-going and lively debate in the literature that seems pertinent to a determination of establishing a threshold of physical harm, and (b) use and cash-in on some of the most relevant discussions in that literature to help establish the liability claim for the threshold of physical harm. To those ends, this section has presented three accounts of liability to defensive harm that attempt to

---

[139] Authors such as Benbaji (2007), Steinhoff (2008), Haque (2009), and Tadros (2012) think that McMahan's account is too restrictive.

[140] Such an account might, admittedly, get us one step closer to being able to defend Micah taking defensive action against Jules in Railway (since Micah has 'a personal prerogative to confer greater weight on their own projects and goals than on other agents [including Jules]' (Fabre, 2014: 101). However, to simply say that an agent has a personal prerogative that outweighs any other relevant factors (similar to what Taurek (1977) argued, discussed in §4.6.1.1.) diverts attention away from liability. Such accounts don't seem to be concerned with whether an agent is liable to defensive harm, but rather with providing an account (or perhaps excuse) for why an agent might be excused from other considerations to justify prioritising their own wellbeing. This is one of the reasons that I will not discuss any such accounts here.

deal with some problem cases, none of which can by themselves account for our intuitions on the various liabilities of agents therein. What is now required is a separate assessment of how we utilise this discussion to form the liability claim. The next section will do so and will bring the other claims of necessity and proportionality (discussed in §4.6.4.1. and §4.6.4.2. respectively) to bear on creating the threshold of physical harm.

4.7. ESTABLISHING THE THRESHOLD OF PHYSICAL HARM IN A MORAL SEQUENCE

This section will draw on the discussions of the primary and secondary narratives and the various aspects of the tertiary narrative to inform a philosophically relevant understanding of the threshold of physical harm. The threshold of physical harm can be understood as incorporating the following elements in the ways outlined here:

i. The primary narrative, which tracks the unfolding of a moral sequence, sequence-event by sequence-event, in a way that enables a Deliberator to predict the likelihood of harm eventuating based on what is happening in the moral sequence (and further based on a prediction of how that moral sequence might further unfold) (see §4.3. and its subsequent discussion in §4.5.).

ii. The secondary narrative, which helps to isolate what sort of other various epistemically available information, outside of the progression of the moral sequence, a Deliberator can use to inform and update the probability of harm eventuating—such epistemic factors include: knowledge of intra-sequence agents (including the desires, dispositions, personality, etc. of those agents); situational

> knowledge, including what is known about the circumstance of the moral sequence and happenings/events prior to the moral sequence, etc.; contextual information, including the social, cultural, and political context of the moral sequence; and expert knowledge of the Deliberator, e.g. the level of expert knowledge pertinent to a particular moral sequence, including the Deliberator's past experiences and prior involvements (see §4.3. and its subsequent discussion in §4.5.).

Jointly, the primary and secondary narratives enable a Deliberator to ascertain the probability that harm will eventuate to a Victim without an intervention (where any secondary narrative supplements the primary narrative). Jointly, the primary and secondary narratives are utilised to identify the risk of harm to Victim. It is for this reason that we can understand the primary and secondary narratives as useful for ascertaining a *probability element* of the threshold of physical harm. But as I argued in §4.4. and §4.5., ascertaining the probability of harm eventuating to a Victim only gets the Deliberator so far. Even though the probability element can furnish a Deliberator with an understanding of the risk that an Initiator poses to a Victim, it cannot, for instance, help a Deliberator to determine whether intervening is morally justifiable—it cannot help to gauge whether a Deliberator should intervene in those cases (i.e. Russian Roulette, the three Scratch cases, and the two Wait and See cases) discussed in §4.5. What the threshold of physical harm needs is an additional *moral element*—and this is what the tertiary narrative provides.

iii. The tertiary narrative seeks to incorporate a number of philosophically relevant considerations required to ensure that an intervention is morally legitimate, and includes the following considerations:

   a. The necessity to take defensive action (see §4.4.).
   b. The proportionality of defensive action (in terms of the magnitude of harm caused by defensive action) (see §4.5.).
   c. The liability of intra-sequence agents to sustain defensive harm (see §4.6.).

Each of iii.a–c corresponds to a respective claim that is required of the tertiary narrative. The moral element of the threshold of physical harm comprises these three claims. And, ultimately, the threshold of physical harm comprises the probability element (primary and secondary narratives) and the moral element (tertiary narrative). But how can we understand each of the three claims of the moral element? And what does a Deliberator do with their understanding of the probability of harm occurring to a Victim? The rest of this section will consider these questions. To kick-start the discussion, let us consider the role that risk of harm plays in the threshold of physical harm.

Risk of Harm to Victim Eventuating

The primary and secondary narratives have been discussed at length in this chapter (see §4.2. and §4.5.) and the decision-making procedure was discussed in chapter 3; I also made some preliminary comments in the discussion on costly and costless interventions in §4.6.2.1. To avoid unnecessarily repeating earlier comments, I will not go over those discussions; here I will simply bring them together.

If the Deliberator judges that the combination of the magnitude and probability of harm occurring to a Victim has reached a certain minimum—dependent on the situation as judged by the Deliberator (but roughly commensurate to or inversely proportionate to the magnitude of the threatened harm to the Victim but with some caveats as outlined shortly)—then the Deliberator can say that they have arrived at a preliminary verdict, based on the primary and secondary narratives, that *ceteris paribus* an intervention would prevent harm to the Victim given the circumstances (although further judgements relating to the tertiary narrative are required to justify an intervention).

Importantly, updating their decision-making with the available evidence and continually assessing the moral sequence in ways that both cashes-in on an understanding of how the moral sequence has progressed and is progressing and the information that the Deliberator possesses might, in cases where the magnitude of harm is high or increases, permit an earlier intervention than would otherwise have been permitted; and likewise in cases where the magnitude of harm is low or reduces, the Deliberator might have to wait. In short, the primary and secondary narratives therefore help to gauge *when* an intervention might be justifiable although they do not by themselves help the Deliberator to assess *whether* an intervention at that point would be justifiable. The Deliberator needs to supplement the primary and secondary narratives (that have led the Deliberator to the belief that *ceteris paribus* an intervention would be justified to avert threatened harm to a Victim) with information about whether harm is necessary, whether the threatened harm can be averted by intervening in a way that is proportionate to the threatened harm, and whether they can intervene in a way that harms only those agents that are liable to harm—which itself is partially directed by the proportionality of the intervention in relation to whether the agent that would be harmed by the intervention is the type of agent that is liable to the proportion

of harm that is necessary in the given circumstances. This will form the basis of the discussion in the rest of this section.

Now that we have an understanding of the role that risk plays in determining the threshold of physical harm (*viz.* when the intervention should take place)—in so far as the probability of harm eventuating to a Victim enables the Deliberator to assess *when* an intervention would be justifiable and it can help ascertain whether an intervention is *probably justifiable*—it does not provide the Deliberator with an understanding of *how* they should intervene nor whether the decision to intervene is *morally justifiable*. What follows is an attempt to define and understand each of the claims required of the tertiary narrative by cashing-in on the relevant discussions from §4.4., §4.5., and §4.6.

The Necessity Claim

The necessity claim, constructed from the discussion in §4.4., is that it is necessary for Intervener to install barrier *b1* to avert harm *h* to Victim only if the expected reduction in the risk of *h* occurring to Victim is greater than the expected harm *h** that Intervener will bring to Initiator, Victim, or Bystander by installing *b1 and* no other less harmful but equally appropriate barrier *b2* is available to install.

To reiterate what I said in §4.4., the reason that the necessity claim is required of the threshold of physical harm is to ensure that (harmful) interventions are enacted only in those circumstances where another non-harmful or less harmful intervention was not available. If the threshold of physical harm did not include the necessity claim then an intervention would be justified in those cases where an Intervener could have chosen a non-harmful or less harmful intervention (yet chose not to), and would, in Wait and See 2 (c.f. §4.5.) for

instance, permit a Deliberator to choose to act now and kill Aggressor rather than waiting for the chance that will later present itself to avert the same harm to Victim by breaking Aggressor's leg; in this case we would, I think fairly uncontentiously, want to say that if the Deliberator is furnished with evidence that harm to Victim can be averted later by breaking Aggressor's leg, Deliberator would not be justified in intervening by acting now and killing Aggressor.

The Proportionality Claim

The proportionality claim, constructed from the discussion in §4.5., is that defensive harm is proportionate only if the expected magnitude of the defensive harm h is less than or equal to the expected threatened harm h^* to Victim *and* the defensive harm is proportionate to the liability of the type of agent (i.e. Initiator, Victim, and Bystander) (as presented in the liability claim below).

The second stipulation in the proportionality claim is required in order to acknowledge the difference between wide and narrow proportionality (discussed in §4.5.) and that, in line with the liability claim (discussed below), defensive harm must be proportionate to the liability of affected agents to defensive harm. Cast the net too narrow and we are stuck with only narrow proportionality which brings with it the *prima facie* issue that an Intervener could not deploy a Victim-injuring or Bystander-harming intervention (since neither is liable to harm), although we surely want to enable an Intervener to deploy a Victim-injuring intervention that, although inflicting harm on a Victim for which they are not liable, saves them from the greater harm that they would have sustained had the Intervener not intervened. This is why we cannot restrict the proportionality claim to only narrow proportionality; we intuitively want to say that an Intervener would be justified in

intervening in a way that causes harm to a non-liable Victim in order to prevent the Victim from sustaining greater harm—and for this, the proportionality claim needs to cash-in on the lesser-evil justification afforded by wide proportionality. Doing so also explains the intuition that enacting a Victim-sacrificing intervention to avert lesser harm to the Victim, say a broken leg, would be widely disproportionate to the threatened harm—killing a Victim to 'save' that Victim from injury would be morally egregious. Cast the net too wide, however, and we get stuck only with wide proportionality which might rule-out interventions against a liable Initiator on the grounds that an Initiator-harming intervention in a particular case would not be less harmful than their threatened harm to the Victim. Cast the net up in the air to catch both and we open the gates to the somewhat uncomfortable position of potentially justifying an intervention against a non-liable Bystander who (unlike a non-liable Victim in the same position) stands to gain nothing from sustaining defensive harm (unlike a Victim who stands to sustain less defensive harm from an Intervener than the otherwise greater harm threatened by the Initiator). This is why I do not think that we can (or should) pin only narrow or wide proportionality to the flag of the proportionality claim. And this is also why, as we shall soon see, I think we should employ a wide proportionality caveat in the case of Bystander-harming interventions, whilst keeping one eye on the liability claim (for reasons that will become apparent). In short, the proportionality claim needs to include narrow proportionality to account for Initiator-harming interventions, but it also needs wide proportionality to account for Victim-injuring and Bystander-harming interventions.

The Liability Claim

As mentioned in §4.6., discussions of liability seek to determine *who* is liable to defensive harm. It therefore makes sense to treat each of the three types of agents (Initiators, Victims,

and Bystander) relevant to this thesis separately (see §4.6.3. in which I justified focussing on Initiator-harming, Victim-injuring, and Bystander-harming interventions). To elaborate: the liability claim, constructed from the discussion in §4.6., is complex and requires variable apportioning of liability to defensive harm on an agent-type by agent-type basis, where the deciding factor in such apportioning is the liability a certain agent has to receive a certain magnitude of harm in virtue of the type of agent they are. In other words, agent *A* is liable to defensive harm *h* only if *A* is the type of agent *T* that is liable to the level of the magnitude of harm that is apportioned for their type of agency (*A*'s being *T*). What this means is that whether defensive harm is proportionate will in part depend on whether the agent is liable to that defensive harm. Although the proportionality claim's partial reliance on the liability claim and the liability claim's partial reliance on the probability claim might make the argument sound circular, it is not—the two are simply heavily reliant on each other. Proportionality and liability are irrevocably intertwined. This will hopefully become clearer in the discussion that follows.

Let us start by considering the four kinds of agents discussed throughout the accounts of liability: Culpable Aggressor, Innocent Threat, Agency-Lacking Threat, and Innocent Bystander. We can map these kinds of agents onto the different types of agents in moral sequencing. Culpable Aggressor and Innocent Threat both seem to fit our understanding of Initiator since both initiate a non-pre-existing threat (NPET) to Victim (c.f. the definition of initiation in §2.2.2.1.); however, the two differ with regards to whether they intend to initiate the moral sequence. Although there is much to be said about the difference between those agents who intend to harm and those who harm unintentionally—and indeed all the various branches that sprout from those, such as the concepts of recklessness, negligence, etc., and whether the harm was foreseen even if unintended—I propose that, for the purpose of the

threshold of physical harm, we treat these as being both cases of initiation that differ only with regards to intentionality. Although a Moral Assessor may wish to attribute greater moral responsibility to Culpable Threat than to Innocent Threat due to their intention to harm, the secondary narrative made it clear that an agent's intentions fall under a consideration of the likelihood of harm occurring; however, as we saw in the discussion of the culpability account, liability to defensive harm relies on making an assessment of intentionality (at least on that account). Our assessment of each will therefore employ an understanding of that agent's intentions for establishing their liability to defensive harm. Importantly, then, intentionality has use both in the secondary narrative and in the liability claim of the tertiary narrative. Moving on, the final two seem to be Bystanders: both Agency-Lacking Threat and Innocent Bystander have no intention of becoming involved in the moral sequence, do not consent to their participation in the moral sequence, and are otherwise disengaged and unmotivated to engage in the moral sequence. The only difference between the two is that Innocent Bystander is *used by* an Intervener in an intervention (*viz.* in a Bystander-harming intervention) whereas Agency-Lacking Threat involves their body simply acting as a projectile with no other agent involvement.

Culpable Aggressor (as an Initiator) is liable to defensive harm in any moral sequence. This tallies with all three accounts of liability, which argue that Initiator-harming interventions specifically against a Culpable Aggressor are justified since that Initiator has intentionally threatened harm to a Victim and so is culpable (on the culpability account), has forfeited his right to not be harmed (on the rights-based account), and is morally responsible for the threatened harm to the Victim (on the responsibility account). However, Innocent Threat diverges from Culpable Aggressors in the literature by virtue of the fact that they do not intend to harm the Victim (although they threaten harm to the Victim regardless). It is this

lack of intentionality that drives the rights-based account and responsibility account to suggest that an Intervener would not be justified in deploying an Initiator-harming intervention specifically against an Innocent Threat. However, I do not think that intention here makes Innocent Threat any less liable to defensive harm than Culpable Aggressor. I think that the latter might for reasons related to a lack of intentional action be less morally responsible for the threatened harm to the Victim than Culpable Aggressor, but I do not think this would affect a Deliberator's assessment of the threshold of physical harm—the harm to the Victim remains the same in both cases and it is through *their* agency (that is, the agency of the Initiator) that harm is threatened to the Victim. I therefore propose that both Culpable Aggressors and Innocent Threats, as Initiators, are equally liable to defensive harm.

It is worth noting that even though we might think that an Initiator's intentionally threatening harm to a Victim means that Culpable Aggressor deserves to sustain a greater magnitude of defensive harm than, say, Innocent Threat, we cannot say that an Intervener would be justified in inflicting greater defensive harm. This stipulation is required to preserve the notion that an Intervener is not justified in acting in *any* way to avert the harm that an Initiator has threatened to a Victim. This is why the liability claim rests on, or rather relies on, the necessity and proportionality claims simultaneously being satisfied. Where an Intervener has a choice of either killing or scratching an Initiator to avert harm to a Victim, it would be unnecessary for an Intervener to opt for the former intervention and would not be justified; moreover, it would seem almost psychopathic for an Intervener to choose to intervene by killing an Initiator when harm to the Victim could have been averted by a non-harmful intervention. Equally, killing an Initiator to avert the threatened harm of a scratch to a Victim is wholly disproportionate and should not be justified either. Although I admit

that this is unsatisfactory for a number of reasons, including that (as I imagine a number of opponents to this view might exclaim) it seems unfair and unjust to require that a Victim endure harm from an Initiator if there is no (necessary and) proportionate intervention available to an Intervener, the requirement that Initiator-harming interventions are only justified is that they are both necessary and proportionate, even though they are wholly liable to defensive harm. Another way of putting this is that Initiators are only liable to defensive harm if it is necessary in the circumstances and proportionate to their threatened harm. So although this view does not permit an Initiator to "get away with" murder (since I think most harmful Initiator-harming interventions would be justified in cases of averting a threat of death to a Victim), it might enable them to "get away with" battery or assault. In such a situation, a Victim can take solace in the fact that a Moral Assessor will likely attribute a sufficient level of responsibility to the Initiator to warrant the sort of desert (be that incarceration, compensation, etc.) that might be expected or demanded by the Victim for the harm they sustained. We can therefore say that an Initiator is liable to *any* defensive harm *so long as* it is necessary and proportionate to the harm they threatened to the Victim.

Before moving on to discuss the liability claim in relation to Bystanders (both Agency-Lacking Threats and Innocent Bystanders) and an as-yet neglected type of agent, Victim, it is important to bring to the fore three issues with discussing Bystander-harming and Victim-injuring interventions in relation to the three accounts of liability discussed in §4.6.: the first relates to the apparent unjustifiability of Victim-injuring and Bystander-harming interventions; the second relates to their inability to justify Victim-injuring and Bystander-harming interventions in non-zero-sum situations (i.e. a Victim sustains harm or they do not); and the third relates to their tendency to seemingly narrowly focus on an exchange of

similar harms (*viz.* situations in which a Victim is saved from death by killing another agent, such as an Initiator).

The first issue is that if we adopt any one of the three accounts of liability discussed in §4.6. then we are unable to account for or justify Victim-injuring and Bystander-harming interventions which seem to be intuitively permissible. Although the culpability, rights-based, and responsibility accounts of liability to defensive harm are undoubtedly useful in cases of self-defence, they are less useful—and downright problematic—for certain cases of other-defence. Although certain readings of these accounts, and indeed some brief nodding in the direction of other-defence by those authors (e.g. Thomson's (1991: 308) defence of a third-party defending Micah from Falling Agent in Well), the accounts cannot explain what I think is an intuitive belief that in certain cases Victim-injuring and Bystander-harming interventions are justifiable. Currently, none of the three accounts of liability can vindicate this intuition. We therefore need to take a novel approach to tackling the issues surrounding Victim-injuring and Bystander-harming interventions, whilst drawing on the accounts of liability for inspiration and philosophical grounding.

The second issue is that the three accounts of liability seem to primarily deal with zero-sum games: either an Initiator's threatened harm to a Victim eventuates or it does not. Moral sequencing, however, requires that we consider more practical cases of intervention, most of which do not often involve zero-sum games. In other words, we need to make sure that we account for Victim-harmful interventions which, if we were to focus only on zero-sum games (where the threatened harm to a Victim eventuates or it does not) would be ruled-out. We need to be able to assess cases such as those in which an Intervener averts the

Initiator's threatened harm of death to a Victim by deploying a Victim-injuring intervention that breaks the Victim's leg, but which thereby saves the Victim's life.

The third issue is that the three accounts of liability seem to focus on an exchange of similar harms—e.g. where a Victim is saved from death (magnitude of harm *h*) by and Intervener killing the Initiator (the same magnitude of harm *h*). However, moral sequencing needs to be able to account for more complex cases in which harm to the Victim can be averted by other less harmful interventions. An Intervener requires the freedom to assess whether and *the extent to which* defensive harm is justifiable—not just whether the same magnitude of harm is justifiable. This further demonstrates the need to look back at and incorporate the literature on necessity and proportionality—both which discuss at great length exchanges of unequal harm—into the discussion for liability to open it up to more practical issues (and more real-life moral sequences).

With that in mind, let us turn attention back to assessing the liability claim in relation to the two remaining types of intervention, Bystander-harming and Victim-injuring interventions.

Bystanders (both Agency-Lacking Threats and Innocent Bystanders) are not liable to defensive harm. As I mentioned above, the only difference between the types of agents is that Innocent Bystander is *used by* an agent in an intervention (*viz.* in a Bystander-harming intervention) whereas Agency-Lacking Threat involves their body simply acting as a projectile with no other agent involvement. Although we might *prima facie* wish to justify the imposing of negligible harm to a Bystander if this averts a much larger harm from being imposed on a Victim (one might say the net-gain is a morally attractive trade-off), I think that we should bite the bullet and say that Bystanders simply cannot be considered liable to

defensive harm (even if the risk of defensive harm is negligible, and even if the defensive harm is necessary to avert greater harm to a Victim). This is in line with all three accounts of liability discussed in §4.6. which demonstrate that and why Bystanders are not liable to defensive harm: they are not culpable since they have no intention of becoming involved in the moral sequence, do not consent to their participation in the moral sequence, and are otherwise disengaged and unmotivated to engage in the moral sequence (and so are not liable to defensive harm on the culpability account); they do not forfeit their negative rights, and by virtue of their type of agency (they are not, after all, a Deliberator) they can be said to have a duty of assistance (the Intervener is employing *them* in the intervention, after all) (and so are not liable to defensive harm on the rights-based account); and they are not morally responsible for the threat of harm to the Victim (and so are not liable to defensive harm on the responsibility account). Both the complete lack of agency in Agency-Lacking Threat and lack of consent of Innocent Bystander to participate in the moral sequence (let alone the intervention) renders both non-liable to defensive harm, and thus Bystander-harming interventions are unjustifiable. But perhaps this bullet isn't too big to bite; in the case of Bystanders, perhaps we can agree with McMahan that 'it is our intuition that is mistaken, not the theory' (McMahan, 2009b: 389)—it is not that we are yet to find an account that can vindicate our intuition that Bystander-harming interventions are justifiable, rather it is that this intuition is mistaken (and so such intuitionists are, perhaps, to forever remain empty-handed).

Finally, let us discuss (the seemingly overlooked issue of) Victim. A Victim is liable to defensive harm so long as it fulfils the lesser-evil justification. Although Victim-injuring interventions cannot be justified on any of the three accounts of liability (a Victim is not culpable, has not forfeited their rights, and is not morally responsible for the threatened

harm to themselves), a Victim stands to gain something whereas a Bystander does not[141]. In other words, the fact that a Victim benefits from a Victim-injuring intervention in a way that a Bystander does not from a Bystander-harming intervention drives the lesser-evil justification that a Deliberator would be justified in intervening by causing less defensive harm to the Victim than they would have received from the Initiator's original threatened harm in a way that that a Deliberator would not in the case of a Bystander (who stands to bear the brunt of harm). But am I having my cake and eating it? I claim that an Intervener is entitled to cash-in on the lesser-evil justification in relation to Victim-injuring interventions (so long as the Victim will with the Intervener's intervention sustain less harm than they would otherwise), yet claim that the Intervener is not entitled to the same lesser-evil justification in relation to Bystander-harming interventions. However, I believe that I can have my cake and eat it. The difference, and important difference, is that the former is justified due to the personal gain (of the Victim) of sustaining a harmful (Victim-injuring) intervention, whereas the latter is not justified since the Bystander does not stand to gain anything—and in fact, in the worst cases, they stand to lose everything. One might think that this trivialises the seriousness of practical cases or that it boils liability down to a self-centred view on the reasons for preventing harm. I don't think this is an implication of this view. After all, it is not that the Victim or Bystander are in these cases choosing to be involved in the intervention, rather the Intervener chooses for them. (Let us not forget that, after all, we are (in line with the discussion in §4.6.3.) only considering those interventions that are other-harming (*viz.* not self-harming). If a Bystander wanted to enact a self-harming intervention to avert greater harm to a Victim, they might have done so. But in the cases we

---

[141] Notwithstanding the Bystander being the Victim's spouse, parent, friend, etc. who might gain something from their death, be that financial reward, moral delight, etc. However, these complexities are peripheral issues that are not salient to the aim or drive of this chapter and so will not be discussed.

are considering, they haven't.) An Intervener is stuck with the unenviable task of deciding if an intervention is justifiable, when an intervention is justifiable, and in what ways they are justified in intervening. Part of this justification can acknowledge that neither the Bystander nor the Victim are liable to harm, but the Victim was *already* under a threat of harm—if they weren't, there would be no moral sequence and the Deliberator would not *be* a Deliberator. So I think that it is reasonable to suggest that although the lesser-evil justification cannot be employed to justify Bystander-harming interventions, it can be employed to justify Victim-injuring interventions. Let us think of this the other way around and think of it not from the perspective of the Deliberator but from the perspective of the agent under threat of defensive harm. The Bystander would likely balk at the idea of being harmed to avert harm to the Victim, whereas the Victim would likely be relieved that there is an option that they will sustain less harm. That's not to say that the Bystander would in all cases balk at them being harmed to prevent greater harm occurring to the Victim—indeed, the Bystander might be a very altruistic person. But even if the reader thinks that this *perspectival justification* for the use of the lesser-evil justification in Victim-injuring but not Bystander-harming interventions is not convincing, the Deliberator can still fall-back on the idea that the net personal gain of the Victim far outweighs the gain (and expected net loss) of the Bystander. This also addresses the issue of Victim-sacrificial interventions (discussed in §4.6.2.6.) since Victim-sacrificial interventions do not fulfil the lesser-evil justification; to 'sacrifice' (*viz.* kill) a Victim in order to avert a lesser threat of harm (say, breaking a bone) is not only disproportionate to the threat of harm, it also cannot be said to have been enacted on the grounds that it is the lesser evil (since it brings about more harm to the Victim).

On their own, each of the three accounts of liability to defensive harm discussed run into insurmountable problems, and none on its own can vindicate our intuitions nor conclusively demonstrate that such intuitions are misguided. In providing one way of understanding the threshold of physical harm, I have taken the more salient discussions related to the concept of liability to defensive harm and demonstrated how they can be mapped onto various agents and types of interventions to provide an understanding of whether and in what ways those agents and associated types of interventions are justifiable. Together, these discussions help a Deliberator to understand the liability claim of the threshold of physical harm.

From this discussion, we can therefore see the threshold of physical harm as lying on a continuum—there is no fixed point beyond which an intervention is justifiable, rather the justifiability of an intervention is subjectively dependent on the evidence available to a Deliberator (that informs the primary and secondary narratives) and the way in which the tertiary narrative is employed to ensure that an intervention is morally justifiable (ascertained by ensuring that the intervention is necessary, that the intervention is proportionate to the threat of harm, and that any agent harmed by the intervention is liable to that harm in the way appropriate for their type of agency). The threshold of physical harm slides along this scale according to (a) the evidence available to the Deliberator and (b) the type of agent that would sustain defensive harm.

The same moral sequence might present itself to two Deliberators who might have different thresholds of physical harm; in other words, although the tertiary narrative remains steadfast in its demands, one Deliberator might consider deploying a Victim-injuring intervention whilst the other might consider deploying an Initiator-harming intervention. Moreover, one Deliberator might have noticed, been afforded, or otherwise accounted for more evidence

(as part of the primary narrative) than the other Deliberator, which might affect the probability of harm occurring to the Victim. And one Deliberator might have more personal, inter-agential, situational, contextual, and/or expert knowledge than the other. All of these differences might change the threshold of physical harm *for that* Deliberator. This, I submit, is entirely plausible as moral sequencing does not seek to identify *the most* justifiable way to intervene, it simply seeks to establish whether a Deliberator in a particular spatio-temporal location, with the evidence available to them and their knowledge, skills, and experience, would be justified (or was justified) in intervening in way *W*.

So how might this play out? Consider a modified case of Jones glassing outlined in the Introduction and another similar case, both presented below.

> *Unknown*
>
> Jones is regular guy. He is not known to the Police and has no criminal record. After a difficult day at work, Jones decides to go to his local bar for a drink. Jones enters the bar, walks to the counter, and orders his favourite drink. After taking a few sips, he notices a co-worker, Harper, at the other side of the bar. Harper is the cause of Jones having had a bad day. After a few more drinks, Jones decides to confront Harper. Jones grabs his drink, walks over to Harper, and after a short but heated exchange of words, Jones raises his glass and throws it. The glass smashes against Harper's face, causing him serious harm.

So what would happen in such a moral sequence where a Deliberator could not identify a secondary narrative—such as a Rookie Police Officer (*RPO*) with little practical

experience? The sequence starts when a NPET is initiated by an Initiator; we might, for instance, say that this moral sequence is initiated the moment Jones took ownership of the glass. However, one might plausibly argue that the moral sequence was initiated at another moment, such as when Jones starts moving towards Harper, or when he raises his hand holding the glass, or even when he entered the bar. Indeed, determining whether a moral sequence has been initiated is reliant on an assessment of the available narratives. For instance, if Jones is a known troublemaker then one might say that the moral sequence was initiated when he entered the bar, or if Jones is a known glasser and has frequently quarrelled with Harper then one might say that the moral sequence was initiated the moment Jones walked towards Harper with a glass in his hand. This is not to say that the initiation of a moral sequence is arbitrary, rather it is down to an assessment of the available narratives—and this might arguably change depending on what information is available, and may also change post-sequence depending on whether a (post-sequence) Moral Assessor is in possession of a fuller narrative.

Known

> Jones is known for glassing people at his local bar. The Police have been tipped-off that Jones will be visiting the bar this evening; they believe that once he is there Jones will glass a man, Smith, with whom he has previously quarrelled. Jones enters the bar, walks to the counter, orders a drink, and notices Smith. After a few more drinks, Jones picks an argument with Smith. Jones grabs his drink, walks over to Smith, and after a short but heated exchange of words, Jones raises his glass and throws it. The glass smashes against Smith's face, causing him serious harm.

How do the three concurrent narratives look for both Known and Unknown?

Primary narrative:

In Known and Unknown, Jones acts in exactly the same way. The sequence-events of the two cases are identical: Jones enters the bar, orders a drink, walks over to Harper/Smith, and glasses Harper/Smith. This sequence of events can be assessed on the basis of the sequence proceeding in such a way that the available evidence enables a Deliberator (e.g. the Police) to assess the probability of harm eventuating.

Secondary narrative:

In Unknown, there is no secondary narrative; the Deliberator (e.g. the Police) has not encountered Jones before, he is not known to be a violent man, and his dispositions, personality, etc. are unknown to the Deliberator. Even if there is no secondary narrative available to the Deliberator, for instance if there is no information on whether the agent has

harmed before, the primary narrative should still be assessed. If the Deliberator has no knowledge of the Initiator's (Jones') personality, dispositions, etc., then that Deliberator can, at best, only construct a pseudo-secondary narrative, both which (a) cannot provide a reliable method for ascertaining the probability of a harmful sequence outcome, and (b) would be hard to justify as reasons for intervening post-intervention (i.e. during a court trial). This is why the primary narrative must be prioritised and why the secondary narrative supplements that narrative/evidence, and not the other way around.

In Known, Jones' past violent conduct, his character, dispositions, etc. are known, or rather are accessible, to the Deliberator. In this case, and from this epistemic vantage point, the Deliberator can use this knowledge, extrapolate an informed hypothesis and understanding of Jones' intentions based on his previous conduct and personality, and supplement the primary narrative with this secondary information.

Tertiary narrative:

In both Known and Unknown, the following questions arise: *How* should Deliberator intervene? Should he shoot and kill Jones? Should he shoot Harper/Smith in the kneecap so he falls to the floor, avoiding the flying glass heading his way? Or should he push a nearby Bystander into Jones, harming the Bystander in the process? Are these interventions necessary, are they proportionate to the threatened harm, and is the agent involved in the intervention liable to sustain defensive harm?

The joining of these narratives (and answering the questions therein) yields the threshold of physical harm beyond which we can say that an intervention to prevent harm is or was

justifiable; but, importantly, this is a calculated (via the primary narrative), informed (via the secondary narrative), and moral (via the tertiary narrative) decision.

So when would an intervention have been justifiable[142]? A Deliberator in Unknown, say a Rookie Police Officer (*RPO*), who happens to be walking by the bar, has to decide at what point the threshold of physical harm has been passed. Determining when this threshold has been reached relies on assessing (a) the primary and secondary narrative of the sequence to establish the probability of harm to the Victim (Harper) eventuating and (b) the tertiary narrative to ensure that an intervention fulfils the necessity, proportionality, and liability claims. Importantly, if no secondary narrative is available to the *RPO*, this does not mean that the threshold of physical harm cannot be determined, nor does it mean that no somatic markers are used in the decision-making process. In this situation, without a secondary narrative, *RPO* would simply fall back on the primary narrative whilst still cashing-in on the tertiary narrative; moreover, *RPO*'s calculations are still subject to *RPO*'s somatic markers—for instance, if *RPO* has been the subject of domestic violence, a somatic marker could be in place that gives him a "gut feeling" that, judging by the expression on Jones' face and how it resembles his attacker's expression during a particular bad incident of domestic violence, something bad is likely going to happen. This "gut feeling" might therefore allow *RPO* to bypass certain considerations in his decision-making process and might provide more evidence that adds weight to the probability of harm occurring to Harper. This "gut feeling" can then either be updated in the face of counter-evidence or, in this case, is confirmed by the primary narrative. In this case, then, one might argue that

---

[142] Importantly, we are not concerned with when an intervention would have been *most* justifiable, we are simply concerned with assessing whether, based on the narratives available to the Deliberator, they would be or were justified in intervening in a particular moral sequence.

either the moment Jones walked over to Harper or the moment Jones raised his glass ready to throw it, is the moment the threshold of physical harm is crossed. *RPO* can make an informed decision that the possible actions available after this are of the type where harm will be caused; the scales become tipped in the favour of harm, and this primary narrative validates his "gut feelings".

Now consider Known. Should *RPO* intervene in the same place as the case before Jones was a known glasser, or does Jones' known predisposition for glassing make a difference for the point of intervention? As a Deliberator, *RPO* now has a unique epistemological vantage point. *RPO* has had a glimpse of Jones' personality: his dispositions, his temperament, his character, and can now use this as a profile of what Jones might do. From this vantage point, a NPET might now be recognised as being initiated when Jones enters the bar; it is a threatening action in as much as Jones: (a) knows that Smith will be at the bar, yet still chose to visit the same bar; (b) has, in his previous actions, demonstrated a disposition for glassing; and (c) has, by glassing before, revealed a violent streak in his character. Say that an Experienced Police Officer (*EPO*) has knowledge of Jones' criminal past, and this experience, knowledge of other criminals, and expertise in policing has resulted in *EPO* developing a number of somatic markers. *EPO* has seen the negative outcomes of Jones' previous glassing, and so, when confronted with a similar situation, has a "gut feeling" that harm will occur. Like in the case of *RPO*, *EPO*'s somatic markers indicate to him that harm will occur, and this narrows-down the number of factors *EPO* needs to consider when deciding the probability of harm occurring to Smith. These situation-relevant somatic markers significantly reduce the number of factors *EPO* must consider in his decision-making process, which thus enables *EPO* to make a quicker decision than *RPO* would in the same situation. This "gut feeling" is then used in *EPO*'s

assessment of the primary narrative in his decision-making process, from which he can decide that the risk of harm eventuating to Smith if Jones enters the bar is significant.

But what is required is that this understanding of the risk of harm is combined with the trio of philosophically relevant factors (necessity, proportionality, and liability) to assess whether the threshold of physical harm has been reached. Both *RPO* and *EPO* would have to weigh-up the situation in Known and Unknown to ascertain the risk of harm eventuating to the Victim (Harper/Smith). Based on the discussion above, it is likely that they would: be justified in enacting an Initiator-harming intervention (against Jones) with the threshold being the lowest of the three since Jones is liable to harm in three different ways (according to the culpability, rights-based, and responsibility account); be justified in enacting a Victim-injuring intervention (against Harper/Smith) on the proviso that it meet the lesser-evil justification; but not justified in enacting a Bystander-harming intervention (against any Bystander) since on no account of liability to defensive harm can harm to them be justified. All of these conclusions would, however, be determined and justified in relation to the moral sequence as it presented itself, and continued to present itself, to *RPO*/*EPO*. This is why it is difficult, if not impossible, to say that time *t* is or was *the* most justifiable time to intervene.

4.8. CONCLUDING REMARKS

This chapter has argued that if a moral sequence passes the threshold of physical harm then an intervention can be morally justified on the grounds that that intervention is (a) responsive to the evidence that there is a non-negligible risk of harm to a Victim, (b) necessary to prevent the Victim from sustaining harm, (c) proportionate to the threatened

harm, and (d) that the agent being harmed is either liable to that harm or is non-liable but harm to them is the 'lesser evil' (in the special case of Victim-injuring interventions).

The next chapter will show how moral sequencing can help to determine which intra-sequence agents are, by their actions in a particular moral sequence, responsible for the harm that was caused to a Victim or would have eventuated without an intervention.

CHAPTER 5:

RESPONSIBILITY IN MORAL SEQUENCING

To complete the picture of moral sequencing, this chapter will turn attention to posing and answering questions related to the responsibility of intra-sequence agents for the harm that occurred or would have occurred to a Victim as a result of their actions or inaction. Different intra-sequence agents—namely, Initiators, Forbearers, Sustainers-2, Interveners, and Snowballers—will be discussed in order to ascertain whether any are responsible for the harm that occurred or would have occurred to a Victim as a result of their actions—of initiating, forbearing, sustaining-2, intervening, or snowballing respectively.

It is not the aim of this chapter to assess or critique competing accounts of responsibility. To do so would detract from the drive of this thesis, namely to provide a novel system of moral sequencing that can be used to decide if, when, and in what way to intervene to prevent harm and, most pertinent to this chapter, explain how the connectedness of an intra-sequence agent to the threatened harm to a Victim can be used to determine their responsibility for their intra-sequence actions, although this is separate from a consideration of moral responsibility; questions about whether and how differing accounts of *moral* responsibility interact with or make a difference to the claim about the responsibility (as defined as a term of art in §5.2.) of Initiators, Forbearers, Sustainers-2, Interveners, and Snowballers are questions I will leave to future work.

However, it will become apparent that there is a potential problem related to responsibility that the system of moral sequencing that I have presented thus far, and this problem will

drive the discussion in chapter 6. After this is addressed, attention will be turned back to the issue of responsibility to show how the *prima facie* problem outlined (in chapter 6) can be accounted for by revising the system of moral sequencing presented thus far.

5.1. WHAT IS MEANT BY 'RESPONSIBILITY'?

This section will look to some of the most relevant and salient philosophical literature on the topic of responsibility to assess what is usually meant when one says that an agent is 'responsible' for an action or outcome. From this discussion, §5.2. will discuss what responsibility means in the context of moral sequencing and how it diverges from the traditional use in the literature, defining the term 'responsibility' as a term of art within the moral sequencing debate that understands responsibility in terms of the extent to which an agent is connected to the harm that is projected or that eventuates to a Victim. Although, as we shall see, the use of the term 'responsibility' in moral sequencing deviates from its ordinary use in the literature, the issue of the *moral* responsibility of an intra-sequence agent for their actions (or lack thereof) is more of an implication of moral sequencing than part of it (in that the framework of moral sequencing seeks to primarily determine if, when, and in what ways an Intervener might install a barrier to harm to avert harm to a Victim and then connect the eventuating or projected harm within that moral sequence to the actions or omissions of intra-sequence agents). Finally, §5.3. will look at how we might go about connecting responsibility to agents and assess how a Moral Assessor would likely determine

the responsibility of intra-sequence agents. So without any further ado, let us look to the philosophical literature[143].

Let us start our discussion with Strawson's (1962) account of 'reactive attitudes'; indeed, '[i]t is difficult to overestimate the influence of Strawson's work on the topic of moral responsibility' (Eshleman, 2014)[144]. Strawson's (1962) aim is to settle the apparent disagreement between what is often called the 'merit-based view' of responsibility and the 'consequentialist view' of responsibility: the former argues that responsibility consists in apportioning praise or blame based on whether such a response would be the appropriate reaction and on whether such praise or blame is deserved; the latter argues that a reaction of praise or blame is appropriate only if such a reaction would result in at least some positive or desired change in the agent under scrutiny (including that agent's future actions). According to Strawson (1962), neither view captures a full-bodied account of responsibility since both assume that an agent's responsibility is tied-up in a (Moral Assessor's) theoretical judgement of whether and to what extent that agent is responsible. In other words, on these two views, an agent being responsible relies on there already being an understanding of responsibility to which an agent's reactions, attitudes, emotions, etc. are mapped over.

---

[143] It would be easy to become embroiled in a discussion of free will, since it is generally opined that if an agent does not have free will (that is, their actions are determined to some extent by certain factors, including socio-economic factors, genetics, the laws of physics, etc.) then that agent cannot be responsible for their actions. However, whether or not and the extent to which agents have free will (*viz.* are free to choose their actions, are fully or partially determined, etc.) and the extent to which these are separate or compatible (*viz.* whether free will and determinism are compatible or incompatible with each other) are separate issues and will not be discussed here.

[144] Strawson's account of responsibility has been widely discussed (see Bennett, 1980; Watson, 1987; Fischer and Ravizza, 1993: 14–22; Wallace, 1994; Russell, 1995: chapter 5; Magill, 1997: 19–22; McKenna, 1998; Magill, 2000; Ekstrom, 2000: chapter 5; McKenna and Russell, 2008; Shoemaker, 2007; Russell, 2013; and Shoemaker and Tognazzini, 2015).

However, Strawson's (1962) view is that, when an agent (say, a Moral Assessor) *MA* declares that agent *A* is responsible for *x*, *MA*'s declaration is reliant on *MA* expressing an attitude (e.g. 'resentment, gratitude, forgiveness, anger', etc. (Strawson, 1962: 6)) towards *A*; the manifestation of this expressed attitude is dependent on the personal relationship between *MA* and *A*, and demonstrates 'how much we actually mind, how much it matters to us, whether the actions of other people—and particularly some other people—reflect attitudes towards us of good will, affection, or esteem on the one hand or contempt, indifference, or malevolence on the other' (Strawson, 1962: 5). Strawson summarises these 'participant reactive attitudes' as 'essentially natural human reactions to the good or ill will or indifference of others towards us, as displayed in their attitudes and actions' (Strawson, 1962: 4–6), and where these reactions are fundamentally related to an agent's 'involvement or participation with others in inter-personal human relationships' (Strawson, 1962: 6). So when *MA* and *A* form or have a relationship, each have certain attitudes that constitute it being a personal relationship. These reactive attitudes can be suspended under two conditions: *A*'s action *x* may be excused if *x* was an accident—although importantly *MA*'s attitude towards *A* does not change, only *MA*'s attitude towards what *A* did changes; or *MA* might claim that *x* was justified if *A* enacted *x* due to a crisis situation or under other mitigating circumstances or if *A* enacted *x* in order to facilitate the eventuating of some greater good—and this produces in *MA* an 'objective attitude' rather than a reactive one. In adopting an objective attitude, the kind brought about via the second suspension of reactive attitudes, *MA* 'ceases to regard the individual as capable of participating in genuine personal relations (either for some limited time or permanently)' (Eshleman, 2014). In Strawson's (1962: 9) words, 'if your attitude towards someone is wholly objective, then though you may fight him, you cannot quarrel with him, and though you may talk to him, even negotiate with him, you cannot reason with him. You can at most pretend to quarrel, or to reason, with

him'. In offering this account of reactive attitudes, Strawson moves the debate on responsibility to being concerned with understanding how *MA*'s judgements about *A* being responsible for *x* are based on the role of *MA*'s reactive attitudes in the way(s) in which *MA* holds *A* responsible for *x*. This is in direct contrast to the way in which the two other merit-based and consequentialist views understand responsibility and the attribution of responsibility. Andrew Eshleman (2014) nicely summarises the impact of Strawson's account thus: 'Whereas judgments are true or false and thereby can generate the need for justification, the desire for good will and those attitudes generated by it possess no truth value themselves, thereby eliminating any need for an external justification'.

The seeming strength of Strawson's view of responsibility is that it is constructed without assuming that attributing responsibility relies on *MA*'s comparing *A*'s actions against an objective, non-person-centred view of responsibility; but this is, in the same breath, a criticism of Strawson's view—that responsibility is determined only by *MA*'s *personal* (and seemingly subjective) reactive attitudes towards the actions of *A* makes *MA*'s determining that *A* is *objectively* responsible for that action impossible. In other words, *MA* cannot justify *A*'s reactive attitudes objectively and, because reactive attitudes are natural (*viz.* are a natural reaction to a situation presenting itself) and dependent on internal processes determined by the inter-personal relationship between agents, they cannot be challenged on theoretical grounds[145].

---

[145] Some authors have attempted to challenge Strawson's views on the grounds that external agents (to the inter-personal relationship between agents) can in fact question an agent's responsibility (see Fischer and Ravizza, 1993: 18; Ekstrom, 2000: 148–149). However, since the purpose of this chapter is not to defend a certain account of responsibility, but rather to provide a brief overview of some salient literature, Strawson's account and associated criticisms will not be discussed further.

Off the back of Strawson's account of responsibility, a number of authors have agreed with Strawson that responsibility is reliant on, or at least tied-up in, an agent having reactive attitudes (see Bennett 1980; Wallace 1994; Watson 1996; Fischer and Ravizza 1998; Darwall 2006). Such authors claim that, in short, (i) agent *A* is responsible for *x* only if another agent (say, a Moral Assessor) *MA* holds *A* responsible for *x* based on appropriate reactive attitudes, and further that (ii) Strawson (and such a view of responsibility) highlights that and how attributing responsibility is a social enterprise. *MA* communicates to other members of the moral community via *MA*'s reactive attitudes a 'demand [for] some degree of goodwill or regard on the part of those who stand in these [inter-]relationships to [them]'—and thus arises 'the reign of universal goodwill' (Strawson, 1962: 6)[146].

The debate on responsibility seemingly took on-board Strawson's views on reactive attitudes, but shifted focus; for instance, Wallace (1994) thinks that reactive attitudes correspond to negative attitudes that are 'second-personal' and in doing so leans towards the notion of responsibility as accountability—responsibility lies in whether and how one is accountable for their actions[147]. Although it is not important for the aim of this chapter to become embroiled in a discussion of what constitutes a reactive attitude, we are interested in the implications of this view for the trajectory of the topic of responsibility. One of these implications is that, as Eshleman (2014: 230) puts it, such attitudes 'presuppose that the agent's actions were her own, or attributable, in the sense that they were expressions of her unobstructed and judgment-sensitive will and that they reflect the moral quality of that will'.

---

[146] For some good discussions of Strawson and related supporting comments, see Stern (1974), Watson (1996), Darwall (2006), Shoemaker (2007), and McKenna (2012).
[147] Darwall (2006: 69) offers a more positive view of accountability (by looking at the concept of 'gratitude'), although he thinks that there may also be other plausible views of responsibility.

However, as Gary Watson (1996: 235–241) argues, we must separate the concepts of responsibility as attributability and responsibility as accountability since the former is not sufficient for the latter. In other words, it is plausible that responsibility *r* can be attributed to agent *A* for *x* without *A* being accountable for *r*—and, if *x* cannot be attributed to *A*, then it would seem unfair to hold *A* responsible for *x* via reactive attitudes. Eshleman (2014) uses the example of a bad career choice, while others have taken the opportunity to apply this to the literature on and our understanding of whether it is fair to hold psychopaths responsible for their actions (see Watson, 2011; Talbert, 2012).

Thomas Scanlon (2008) offers a rich understanding of the 'moral dimension' of blame. This is relevant for two reasons. First, Scanlon's account of blame seems to support the idea of responsibility as attributability (in terms of how blame is attributed). Second, Scanlon's discussion of blame reflects a folk understanding of responsibility, namely that responsibility involves being deserving of praise or blame (then perhaps corresponding to being deserving of punishment or reward)—that agent *A* can be blamed for *x* is to say (or at least implies) that *A* is responsible for *x* (and perhaps further that *A* can be punished for *x*). However, this folk understanding is different from Scanlon's account, so let us look more closely at Scanlon's argument. Scanlon argues that 'blame normally involves more than

evaluation but is not a kind of sanction' (Scanlon, 2008: 122)[148]. An agent *MA* (to keep this in line with the terminology of moral sequencing, I propose we understand this agent as the Moral Assessor, abbreviated here to *MA*) and agent *A* are friends. In this situation, *MA*'s blaming *A* for doing *x* (that wrongs *MA*) involves *MA* recognising that *A*'s doing *x* indicates that *MA*'s relationship with *A* (who *MA* blames for *x*) has been 'impaired' (Scanlon, 2008: 123, 131–138), and this, Scanlon argues, is the lens through which we should understand their relationship. In other words, we all have relationships with one another and they become impaired—that is, damaged—when agents perform certain actions[149]. Vallier (2010: 562) neatly summarises Scanlon's point thus: 'blame registers the damage done'. Interestingly, Scanlon (2008: 136–137) thinks third-party *P*—that is, those outside of the relationship between *A* and *MA*—is not permitted to blame *A* but merely 'disapprove' of *A* since *P* 'is not in a position to adjust his attitudes toward the guilty party [*A*] in the relevant way', although this does not prevent my blaming of strangers and vice versa since we have a 'moral relationship' even with strangers (Scanlon, 2008: 138). When a stranger engages in a morally impermissible behaviour against me, that stranger impairs my moral

---

[148] It is worth noting that, whilst Scanlon's (2008) discussion of blame is most pertinent to this chapter's focus on responsibility, he also discusses two other 'dimensions' of moral responsibility, namely 'permissibility' and 'meaning'. Scanlon's analysis of 'permissibility' involves a discussion of the Doctrine of Double Effect in so far as permissible action seems dependent on an agent's intentions for acting; in the famous "trolley problem" (Foot, 1984: 183), Scanlon argues that the deciding factor in whether pulling the level is permissible is dependent on what the agent's intentions are for pulling the level—only if the intention is to save, and not kill, is the action permissible—and that acting impermissibly—acting with ill-intent or the wrong attitude—can impair that agent's relationship with others. (This concept of permissibility also links back to Scanlon's (1998) previous work on 'What We Owe Each Other'.) Scanlon's analysis of 'meaning' is linked to his discussion of 'permissibility' since an action's 'meaning' is dependent on the intentions or attitudes of the agent. This is where 'blame' enters the picture: having the wrong intentions or attitudes is blameworthy.

[149] Scanlon (2008: 125) also differentiates between blame in the sense described and 'objective stigma'—the former is a more personal impairment whereas the latter is a more impersonal impairment which he rejects.

relationship with them; Scanlon (2008: 147) even goes as far as to say that a stranger plays 'a distinctive role in our lives' when they wrong us. Scanlon's account of blame, then, relies on (1) 'adjusting one's attitude toward the person [*A*]' (Scanlon, 2008: 130) or '*revising* one's [*MA*'s] attitude toward him [*A*]' (Scanlon, 2008: 131) and (2) an impairment in one's relationship with another (e.g. between *A* and *MA* (as described above))[150]— *MA*'s blaming of *A* for *x* essentially 'marks a change in that relationship and hence is a form of blame' (Scanlon, 2008: 130) and, in addition, a stranger wronging us impairs our relationship with them as a 'fellow human being' (Scanlon, 2008: 140). Scanlon's account therefore seems to be aligned to what is usually called an attributability view of responsibility[151]—namely, that *x* is attributable to *A* grounds the claim that *A* is responsible for *x*—although it is by no means a "neat fit" since the account is primarily focussed on inter-personal and third-party attributions of blame to agents.

Since Scanlon (2008)—or more specifically since Angela Smith's (2005) earlier paper in which she, like Scanlon, argued for an attributability view of responsibility—there has been a growing movement to reframe how responsibility is understood. David Shoemaker (2011), for instance, argues that 'Scanlonian responsibility is not comprehensive enough' since it 'conflates attributability and answerability, which are actually distinct conceptions of responsibility' (Shoemaker, 2011: 603); he also builds on Watson's (1996) account of responsibility by drawing a distinction between types of responsibility. Scanlon's argument

---

[150] For a good and more detailed explanation of these two claims, see Shoemaker (2011: 603).

[151] Interestingly, Scanlon (2015) later changes his terminology, changing 'responsibility as attribution' to 'moral reaction responsibility', and clarifies some subtle differences (Scanlon, 2015: 89–90) and how this differs from the second kind of responsibility, substantive responsibility, that he considers. Delving into a comprehensive review of the literature on responsibility and presenting a chronology of related arguments lies beyond the aims of this chapter and so this will not be discussed further.

assumes that *A*'s being responsible for *x* involves being answerable to *MA* for *x*, and the mistake, as Shoemaker (2011: 603) puts it, is that this implies that 'the conditions of answerability are just the conditions for attributability'. So, in short, on Scanlon's (2008) account, that *x* is attributable to *A* makes *A* answerable for *x*. However, Shoemaker (2011) argues that it is possible for *A* to be 'attributably-responsible' for *x* without *A* necessarily being 'answerability-responsible' for *x* (and uses the two cases of irrationality and emotional commitment to drive this claim). More specifically, Shoemaker (2011: 611) argues that on closer inspection we find the 'surprising result' of cases that involve 'attributability without answerability'—for instance, I can have attitudes that belong to me but for which I am not answerable (and uses the example of standing by your ex-partner ('a groundless emotional commitment'), unable to justify my attitudes towards my ex-partner, although the attitudes that I have clearly belong to me and 'reflect on me, on my deep self, and in particular on who I am as an agent in the world' even though 'they are not grounded in any evaluative reasons'); Shoemaker (2011: 612–615) uses the concept of 'aretaic appraisals'[152] to highlight this. Furthermore, Shoemaker (2011: 604) believes that Scanlon conflates the concepts of responsibility as answerability and responsibility as accountability since 'one may well be answerable for ϕ without being accountable for it'. Indeed, according to Shoemaker (2011: 623), 'there seems no reason in principle to think of them as inexorably intertwined'. To help explain that and why these three ways of understanding responsibility are distinct, Shoemaker (2011: 627–630) considers a case of a psychopath to test what each

[152] The notion of 'aretaic appraisal' was first introduced by Watson (1996: 231) as an evaluation of an agent's 'excellences and faults—or virtues and vices—as manifested in thought and action'. Shoemaker (2011: 612–613) has a slightly different interpretation and describes these appraisals as 'track[ing] expressions of evaluative commitments generally […] [and] may be warranted in the absence of answerability'.

view of responsibility says about the responsibility of a psychopath[153]. A psychopath *P*, Shoemaker (2011: 627–630) claims, is both attributability-responsible and answerability-responsible, but not accountability-responsible, for their actions. *P* is attributability-responsible since his actions 'surely reflect his evaluative commitments, however shallow they may be' and because 'his attitudes and actions are expressions of his self qua agent' (Shoemaker, 2011: 628). *P* is answerability-responsible since 'his actions and attitudes flow from evaluative commitments themselves grounded in his reasons' (Shoemaker, 2011: 628). But *P* is not accountability-responsible, *viz. P* cannot be susceptible to sanctions, since they 'lack a crucial capacity, namely, a certain sort of moral sense: they do not care about the justifiability of their actions, and they lack sensitivity to the sort of considerations to which they would be required to attend even if they did care about justifiability' (Shoemaker, 2011: 628). We therefore need to understand responsibility as not being restricted to 'responsibility as answerability', but rather extending beyond this to encompass 'responsibility as answerability' and the concepts of 'responsibility as attributability' and 'responsibility as accountability'[154].

---

[153] Shoemaker (2011: 627–628) defines the case of the psychopath as follows: 'Psychopaths, in general, have severe deficits in their emotional, interpersonal, and self-control capacities. They do not feel guilty about harming others, they lack empathy generally, and they lie to and manipulate others for personal amusement. They are also imprudent, repeatedly motivated to act on their impulses for short-term gain to the detriment of their long-term interests'. It is under these circumstances that Shoemaker asks: 'Are they morally responsible for their actions and attitudes?'

[154] For Shoemaker, to be attributability-responsible is for an action or attitude 'to express my practical commitments' (Shoemaker, 2011: 631), to be answerability-responsible is 'to be susceptible for assessment of, and response to, the reasons one takes to justify one's action' (Shoemaker, 2011: 623), and to be accountability-responsible is 'to be susceptible to being held to account if one flouts relationship-defining demands' (Shoemaker, 2011: 623).

Angela Smith (2012) offers a response to Shoemaker (2011) by arguing that 'our moral practices do not, in fact, embody three different conceptions of moral responsibility'; she argues that the responsibility as answerability account 'is indeed the only kind of moral responsibility there is' (Smith, 2012: 576). Although Shoemaker believes that an agent is responsible for their irrationality—since 'irrationality is the sort of thing that can be attributable to an agent' (Smith, 2012: 579)—according to Smith it 'makes little sense' to say they are responsible for their irrationality (Smith, 2012: 580). An (irrational) agent that has a phobia of spiders, whilst simultaneously acknowledging that spiders are not dangerous, is answerability-responsible for their irrationality since they are able to 'reconsider the judgments [they] take to support each of [their] conflicting attitudes, because those judgments conflict and therefore they cannot both be correct' (Smith, 2010: 581). In other words, a person is not responsible for being irrational because they have been attributed irrationality, they are responsible because they have the attitudes that make them irrational. It is for this reason that Smith believes that Shoemaker is wrong in claiming that an irrational agent is attributability-responsible but not answerability-responsible—'[a]n agent is answerable for her irrationality *because* she is answerable for the attitudes that together constitute her irrationality' (Smith, 2012: 580, emphasis added).

Essentially, the authors discussed above and similar authors offer a distinction between, and offer various accounts that try to make sense of, three different ways (and, on some accounts, intertwined ways) of understanding responsibility: responsibility as attributability, responsibility as answerability, and responsibility as accountability—although there is little agreement about whether these are separate, the ways in which they are interconnected, and the scope and limits of the definitions themselves.

Before moving on, it is important to note that there are other accounts of responsibility that do not rely on Strawson's reactive attitudes nor an underlying account of the relationships between attributability, answerability, and accountability. One of the most interesting is the 'ledger view' of responsibility, which works on the idea that each agent is in possession of a metaphorical ledger against which, depending on their actions, responsibility is credited or debited. On such views (e.g. Feinberg, 1970: 30–1; Glover, 1970: 64; Zimmerman, 1988: 38–9)[155], agent *A* is responsible for action *x* if credit or fault is properly attributed to *A* for *x*. It will therefore come as no surprise that such ledger views of responsibility often adopt a responsibility as attributability view.

I will henceforth refer to the kinds of accounts of responsibility (in the philosophical literature) discussed above as accounts of *moral responsibility* in order to differentiate these discussions from the ways in which moral sequencing employs the term 'responsibility'. The next section will explain that and why this chapter will deviate from the sorts of discussions outlined above and why this chapter will not discuss moral responsibility.

5.2. LIMITING THE VIEW: 'RESPONSIBILITY' AS A TERM OF ART IN MORAL SEQUENCING

The last section briefly outlined some of the prevailing accounts of moral responsibility in the philosophical literature. This was provided to give some background to the concept of moral responsibility to situate the forthcoming discussion and to clearly demarcate the traditional philosophical debate in (moral) responsibility as presented in the literature above

---

[155] See Watson (1987: 261–2) and Fischer and Ravizza (1998: 8–10) for a discussion of these ledger views.

and the sort of responsibility that is relevant to moral sequencing and that will be discussed in this chapter. Instead of asking "Is intra-sequence agent *A morally responsible* for their actions (and to what extent and in what ways)?" (as asked by the philosophical literature), this chapter seeks to ask "How much *harm* is each intra-sequence agent responsible for (and to what extent and in what ways can harm be connected to the actions of intra-sequence agents)?" It is in this way that the term 'responsibility' and its derivatives are used *as terms of art* in moral sequencing, not to attribute moral responsibility to an intra-sequence agent but rather to attempt to *connect* their actions to the harm that eventuated (or would have eventuated) to a Victim as a result of their actions or inaction: it seems that the least connected an agent is to harm, the less responsible they are; the more connected they are, the more responsible they are.

Although some might think that *x* being connected to *y* requires a causal explanation for how or why *x* causes *y*, this is not what is under consideration. Instead, that agent *A* is connected to harm *h* and the extent to which *A* is connected to *h* is determined by the way(s) in which *A* acts or refrains from acting and the relation that action or omission has to the harm that eventuated (or would have eventuated without an intervention) to a Victim. This is the sense in which this chapter seeks to determine the connectedness of agents to harm.

However, although the connectedness of a *type* of agent explains why that type of agent is to a certain degree responsible for (actual or expected) harm to a Victim by virtue of the type of intra-sequence actions that correspond to that type of agent (i.e. an Initiator initiating, a Forbearer forbearing to prevent harm, etc.), some mitigating factors might exist such that a Moral Assessor might decide to reduce the responsibility for the *individual* agent to whom such mitigating factors apply. In other words, although the general level of responsibility is

set for each type of intra-sequence agent (i.e. the level of responsibility is set for Initiators, Forbearers, Sustainers-2, Interveners, and Snowballers) by virtue of the connectedness of their actions or inaction in moral sequencing to harm, individual agents (i.e. individual Initiators, Forbearers, Sustainers-2, Interveners, and Snowballers) might be susceptible to a Moral Assessor imposing a *responsibility-decrease* depending on various mitigating factors.

When talking about 'responsibility' in moral sequencing it is therefore the extent to which an intra-sequence agent is responsible for the harm that resulted (or would have resulted) to a Victim by their action or inaction that is under assessment. In other words, it is the responsibility for the harm that was caused, averted, or mitigated (to/from a Victim) that this chapter seeks to connect to intra-sequence agents. The issue of *moral* responsibility, although inextricably linked to this 'responsibility as connectedness to harm' account, is a separate and subsequent issue that is arguably determined by the account of moral responsibility to which the Moral Assessor subscribes. The responsibility that moral sequencing seeks to determine, namely what harm can be connected to various intra-sequence agents, is therefore arguably explanatory prior to moral responsibility. The two are, however, connected since it seems that the more harm that can be connected to an agent (and thereby the more responsible an agent is for the occurrence of harm) the more likely they are morally responsible. It is in this way that a discussion of the ways in which harm can be connected to the actions or inaction of intra-sequence agents paves the way to the sorts of discussions of moral responsibility in which a Moral Assessor will be engaged. However, and importantly, attributing moral responsibility to an agent relies on comparing the actions of agents against a particular account of moral responsibility, but since it is not the aim of this chapter to state or argue which account of moral responsibility is best, nor suggest to which account of moral responsibility a Moral Assessor should subscribe, the

discussion that follows will not say whether or in what ways certain agents are morally responsible—since, to reiterate, this can only be ascertained with reference to a particular account of moral responsibility[156]. The ways in which an intra-sequence agent is 'responsible' for their intra-sequence actions (or lack thereof) therefore tracks the ways in which, from the perspective of post-sequence agent Moral Assessor, eventuating or projected harm to a Victim can be connected to the actions or inaction of such intra-sequence agents.

That I am using the term 'responsibility' in this way might be unconventional and might be a bone of contention. It is for reasons of transparency, and to avoid possible counter-objections related to the fact that this chapter purports to discuss responsibility ('responsibility' as traditionally understood) but does not so, that I seek to be honest about its usage in moral sequencing and in this chapter. My aim is not to mislead the reader or complicate the issue, but rather to assess from a pre-theoretical perspective[157] the sorts of harm that might be connected to an agent in order to say what sort of harm an agent is responsible for *so that* a discussion of moral responsibility can ensue but importantly *outside of* the discussion of moral sequencing.

The next section will proffer for what harm different intra-sequence agents might be responsible.

---

[156] After gauging the extent to which harm can be connected to intra-sequence agents, a Moral Assessor might further determine the moral responsibility of an intra-sequence agent (in the ways potentially cashed-out by the authors discussed in §5.1.) for the amount of harm that is connected to them, although such a discussion, for reasons explained above, is outside the scope of this chapter.

[157] By this I mean discussing responsibility, and the responsibility of intra-sequence agents for their actions, arising before and without reference to any pre-established theoretical considerations.

5.3. DETERMINING RESPONSIBILITY

This section will discuss the cases of initiating, forbearing, sustaining-2, intervening, and snowballing and discuss what harm can be connected to those intra-sequence agents and whether they should shoulder (at least some of) the responsibility for the harm that befalls (or would have befallen) a Victim. To this end, the notion of *retrospective responsibility* will take centre stage so that a Moral Assessor can make post-sequence judgements about whether the intra-sequence agent(s) are responsible for any harm that occurred or would have occurred to a Victim.

However, this chapter will not discuss cases of enabling since, as a result of the discussion in §2.2.2.3. (particularly in relation to The Avalanche case), enabling a sequence to continue and removing a (dispositional or operative) barrier to harm is a case of initiating, and so related issues will be presented in the section below on initiating. Cases of sustaining-1 will not be discussed either since I argued in §2.2.2.2. and §2.2.2.3. that sustaining-1 (as presented by Woollard) is philosophically incoherent and should not form part of moral sequencing; sustaining-1 is a case of initiating (as in Rock), a case of being an accessory to or complicit in the NPET (as in Rock 2), or is a case of snowballing (as in Rock 3) (discussed in §2.2.2.2.). I argued that the concept of sustaining, if anything, is sustaining-2; sustaining-2 will be discussed in this chapter.

5.3.1. INITIATING

Determining the responsibility of an Initiator is intuitively straight-forward. In §2.2.2.1., I argued that a moral sequence is initiated when an agent (Initiator) brings about a non-pre-existing threat of harm (NPET). To reiterate my comments, initiating necessarily brings

about a potentially harmful state of affairs that did not exist prior to the Initiator's actions, and, since initiating 'is in some relevant sense *dependant* upon the agent' (Woollard, 2008: 226) bringing about a NPET, an Initiator is directly and inextricably connected to the harm that does befall or would have befallen a Victim. *Prima facie*, then, a Moral Assessor would be justified in stating that an Initiator is fully responsible[158] for the threatened harm to a Victim; through their actions, the Initiator changed the *status quo* to bring about, or attempt to bring about, a threat of harm to a Victim.

However, although an Initiator is the *type of* intra-sequence agent that by virtue of their inextricable connectedness to harm is most responsible for harm (eventuating or intervened), the degree of responsibility attributed to *an individual* Initiator would likely be dependent on the circumstances surrounding the initiation of the moral sequence and whether or not a Victim was harmed as a result. The greatest responsibility is reserved for those Initiators whose intentional threat of harm to a Victim eventuates. Here, the Initiator is not only connected to harm that eventuated to a Victim, they intentionally caused that harm. However, there are factors that might decrease the responsibility of an Initiator. What follows is a discussion about the conditions in which a Moral Assessor might impose on an Intervener a *responsibility-decrease* (that is, a reduction in how responsible they are for harm).

---

[158] Stating that an agent is 'fully responsible' does not in the context of responsibility outlined in §5.2. mean that an agent should be blamed, punished, etc., nor that their 'full' responsibility prevents any further apportioning of responsibility to other agents. Indeed, this is not a moral claim, it is not saying that the agent is morally responsible, and it is not saying that this agent is the only responsible intra-sequence agent in a moral sequence —it is simply stating that such an agent is fully connected to the harm to a Victim.

An Initiator that unintentionally brought about a NPET to a Victim would likely have a responsibility-decrease when compared to an Initiator that intentionally brought about a NPET to a Victim; and, moreover, an Initiator that was forced, coerced, or manipulated into bringing about a NPET to a Victim would likely have a further responsibility-decrease for their actions. This is because: in the former, although they are still connected to the harm (to a Victim), they became *accidentally connected* to harm; and, in the latter, although they are still connected to the harm (to a Victim), they became *inadvertently connected* to harm (by virtue of their connection to it being forced, coerced, manipulated, or otherwise non-voluntary).

An Initiator whose NPET was averted by the actions of an Intervener would still be responsible for being the originator of the threat to a Victim that would, without the intervention, have caused the Victim harm, however such an Initiator would likely have a responsibility-decrease by virtue of the fact that the Intervener's (successful) intervention prevented harm from eventuating and thereby *disconnected* the Initiator from actual but not projected harm—the Initiator is still connected to the harm that *would* have eventuated (or was projected to eventuate) even though the Initiator has, as a result of a (successful) intervention, been disconnected from actual harm (since harm did not eventuate).

So how do these overlap? Should, for instance, an Initiator whose unintentionally threatened harm to a Victim eventuates be subjected to a responsibility-decrease when compared to an Initiator who intends to threaten harm to a Victim but which is successfully intervened (and thereby averted)? The *degree* of responsibility of Initiators in such circumstances is dependent on the theoretical considerations of moral responsibility held or employed by the Moral Assessor and, since it is not this chapter's aim to claim *which* theoretical

considerations a Moral Assessor should ascribe to, this will not be offered. Indeed, factors that determine whether an Initiator should be more or less responsible relies on whether, for instance, the Moral Assessor believes that responsibility is best explained as responsibility as attributability, answerability, or accountability, and/or whether they hold consequentialist or non-consequentialist views, and/or indeed any other theoretical considerations that the Moral Assessor believes is relevant to determine the degree of responsibility.

All that is important for our purposes is to recognise that a Moral Assessor would determine that the *type of agent* Initiator would have full responsibility for the threatened harm to a Victim by virtue of the fact that the Initiator is inextricably connected to harm to the Victim in that moral sequence. However, as mentioned, the way in which responsibility is cashed-out on the individual level—that is, to *individual* Initiators—is dependent on what theoretical considerations the Moral Assessor brings to bear in their assessment.

5.3.2. FORBEARING

To recap §2.2.2.4., an agent forbears to prevent harm when that agent, who is aware of the threat of harm to a Victim, knowingly remains inert with respect to intervening, and fails to install an available and appropriate barrier to the harm.

Since a Forbearer (who was previously a Deliberator in the same moral sequence) is aware of the moral sequence and the threat of harm to a Victim, yet chooses not to intervene, a Forbearer is partially responsible for the harm that befell the Victim. A Forbearer could have at least attempted to intervene to prevent the Victim from being harmed—in order to forbear they must have been spatio-temporally able to intervene. But the fact is that a Forbearer, by

definition, chooses to remain inert or act in a way that does not (even attempt to) install a barrier to harm, and it is a Forbearer's position to intervene (even though they do not) that grounds the claim that a Forbearer is spatio-temporal proximate to the (eventuating or projected) harm to the Victim and it is a Forbearer's inertia that makes it apparent that a Moral Assessor would have little trouble in declaring that the Forbearer was partially responsible for the (eventuating or projected) harm to the Victim.

But can a Moral Assessor make a blanket determination that all agents who forbear to prevent harm in a moral sequence hold the same level of responsibility? Are all Forbearers connected to harm in the same way? Indeed, there might be general or personal mitigating factors that decrease an individual Forbearer's responsibility in a particular moral sequence. Like in the case of initiating, there are some Forbear-specific mitigating factors that might result in a Moral Assessor imposing a responsibility-decrease on individual Forbearers. Let us consider what some of these factors might be.

First, a Forbearer might have a reason (or perhaps an excuse) for not intervening. If these reasons or excuses are poor (e.g. if a Forbearer suggested that they could not intervene because doing so would mean that they would miss their favourite soap opera), implausible (e.g. if a Forbearer suggested that they could not intervene because doing so might make them go into cardiac arrest, even though they are healthy and have not until now been worried about such an issue), or patently false (e.g. if a Forbearer suggested that they could not intervene because doing so would jeopardise their travelling to Jupiter for a top-secret mission), then a Moral Assessor would likely not impose a responsibility-decrease since such reasons/excuses do not sufficiently explain why the Forbearer forbore to prevent harm. However, if an intervention would expose a Forbearer to risk of harm then a Moral Assessor

might impose a responsibility-decrease to the Forbearer for the harm the befell a Victim. However, whether and the extent to which varying degrees of risk of harm would excuse an agent from intervening (and thus become a Forbearer) is a grey area. A Moral Assessor would likely not want to impose a responsibility-decrease on a Forbearer who exclaimed that they had forborne to prevent harm to avoid negligible harm (e.g. to avoid receiving a scratch), but would likely do so for a Forbearer who forbore to prevent harm to avoid non-negligible harm (e.g. to avoid losing a limb); although providing an estimate of what sort of harms would warrant certain degrees of responsibility-decrease will not be discussed here. All that is required is an understanding that a Forbearer, by virtue of their connectedness to harm (in that they could have installed a barrier to harm but chose not to), is partially responsible for the harm that befalls or would have befallen a Victim; however there are certain mitigating factors that might decrease a Forbearer's responsibility, although the degree of any responsibility-decrease would be determined by the sort of harm that a Moral Assessor believed a Forbearer would be justified in avoiding (where the avoidance of this harm results in a Forbearer forbearing to prevent harm to a Victim).

Second, regardless of a Forbearer's forbearance, harm to a Victim might not eventuate. If a Forbearer forbears to prevent harm to a Victim, but another Deliberator decides to become an Intervener and successfully intervenes to prevent harm to the Victim, is a Forbearer responsible for the harm that would have eventuated without the Intervener's intervention? In short, yes. A Forbearer is still connected to the harm that would have eventuated without the intervention, and for this they are responsible—although a successful intervention by another agent disconnects a Forbearer from the actual harm to the Victim (since the threat of harm to the Victim was averted), it does not disconnect that Forbearer from the expected harm to the Victim (since the Victim would have sustained that harm if it were not for the

Intervener). However, like in the discussion of Initiators, the degree of any responsibility-decrease would be determined by the theoretical considerations adopted by the Moral Assessor.

5.3.3. SUSTAINING-2

Like in the case of initiating, determining the responsibility of a Sustainer-2 is intuitively straight-forward (and for similar reasons). To recap, an agent sustains-2 a sequence when they knowingly remove a dispositional barrier to harm, but without directly acting on the current NPET.

Since a Sustainer-2 knowingly removes a barrier to harm, they are directly connected to the threatened harm to a Victim since it is because of their actions that harm to a Victim will eventuate (without an intervention). If a Sustainer-2 intentionally removes a dispositional barrier to harm, then they are, like an Initiator, fully responsible for the threatened harm to a Victim; the fact that harm would not have eventuated had they not removed the dispositional barrier to harm highlights that their removal of that barrier is evidence of their connectedness to harm. By choosing to remove a dispositional barrier to harm, a Sustainer-2 knowingly acts in a way that ensures that foreseeable harm to a Victim occurs.

Furthermore, cases in which a Sustainer-2 not only knowingly removes a barrier to harm but also chooses not to install a barrier to the harm (c.f. the Aggrieved Smith case in §2.2.2.3.) and thus simultaneously forbears, illustrates how a Moral Assessor might levy a *responsibility-increase* on that agent since an agent in this position is *both* a Sustainer-2 and a Forbearer. However, since by definition a Sustainer-2 cannot unintentionally remove a

dispositional barrier to harm, a Sustainer-2 would not be the recipient of a responsibility-decrease from a Moral Assessor based on an accidental removal, and equally there are no other mitigating factors that might lead a Moral Assessor to impose a responsibility-decrease on a Sustainer-2.

Like in the cases of the other types of intra-sequence agents, the degree of responsibility that a Moral Assessor would ascribe to Sustainer-2 would have to be decided with reference to certain theoretical considerations; in the case of a Sustainer-2 sustaining, a Moral Assessor may, for instance, fall back to the literature (discussed in chapter 1) on whether there is a moral distinction between doing harm and allowing harm (for Sustainers-2 can be said to have allowed, rather than done, harm). But what is important for current purposes is that a Sustainer-2 is fully responsible for the harm that befell or would have befallen a Victim. Moreover, in cases of sustaining-2 in which that agent simultaneously forbears to prevent harm (by failing to install a barrier), a Moral Assessor would have grounds for increasing that agent's responsibility (compared to if that agent had only sustained-2 the moral sequence); this is because such an agent both removed a barrier to harm *and* failed to install a barrier to prevent harm—in other words, *on not one but on two occasions* that agent could have, but did not, prevent harm to a Victim.

5.3.4. INTERVENING

To recap, an intervention is an agential event consisting of installing a barrier to harm that intrudes into a moral sequence and which originates directly from the actions of an agent; an Intervener knowingly imposes himself on and is motivated to affect the outcome of the moral sequence.

As in the cases of initiating and sustaining-2, attributing responsibility to an Intervener is intuitively straight-forward, but for inverse reasons. Intervening brings about a state of affairs (by way of installing a barrier) that did not exist prior to the Intervener's actions; an Intervener installs a barrier to harm in an attempt to avert the harm that would have befallen a Victim had he not intervened. *Prima facie*, then, a Moral Assessor would not be justified in determining that an Intervener is responsible for the projected harm to a Victim; through their actions, an Intervener acted to avert, or attempt to avert, a threat to a Victim—an Intervener is disconnected from the threat of harm to a Victim.

However, an intervention is not always effective. An intervention can be partially or fully ineffective (see §2.2.2.4.): installing a fully ineffective barrier is a case of a failed intervention and installing a partially ineffective barrier is a case of a sub-optimal intervention. In such cases, how might a Moral Assessor determine an Intervener's responsibility? Although the intervention failed or was sub-optimal, a Moral Assessor would likely not want to attribute even some responsibility to such an Intervener; to do so would fail to recognise that the Intervener attempted to avert the harm. Indeed, even in a failed or sub-optimal intervention, if the Intervener had not acted then (assuming the Intervener acted past the threshold of physical harm) it is likely that harm would have occurred to a Victim. It is therefore unlikely that a Moral Assessor would, in cases of a failed or sub-optimal intervention, determine that an Intervener is responsible; to do so would be to say "I recognise that you acted in a way that, in your mind, was an appropriate intervention that would avert the occurrence of harm to the Victim… but you failed, and by doing so you allowed the Victim to be harmed, and you must shoulder some of the responsibility for that". Such an utterance, however, would be unlikely to occur; I do not think that a Moral Assessor would wish to attribute even some level of responsibility to

Interveners who did all they could to avert the harm to a Victim, but whose efforts failed or were sub-optimal. Indeed, in some cases, perhaps an ineffective barrier is all that is available to an Intervener to install, and in such cases I do not think a Moral Assessor would balk at Initiator for installing such barriers—after all, the Intervener did all they could to avert the harm to a Victim.

Some exceptions to this are if an Intervener (a) knowingly installs a fully or partially ineffective barrier when a potentially more effective barrier was available to them or (b) installs an inappropriate barrier. Both of these scenarios are discussed in §2.2.2.4. There, I argued that (a) is a clear case of forbearing, and likewise, looking to (b), the installation of what I called a blatantly inappropriate barrier (the inappropriate installation of which would be blatantly obvious to any reasonable person) would count as an instance of forbearing, but only so long as an appropriate barrier was available but was not installed. To repeat my comments there, if no appropriate barrier was available, I do not think we would begrudge an agent an attempt to prevent harm, even if any such action would almost certainly fail to prevent the harm—but, to add to this from the discussion in §4.6.2., such a hopeless intervention must be costless. But in cases in which an appropriate barrier was available, yet the agent (unwisely) chose to install an inappropriate barrier, the agent has clearly forborne to prevent harm. We can therefore refer questions of determining the responsibility of agents in cases of (a) and cases of (b) to a Moral Assessor who is considering whether a Forbearer should shoulder any responsibility.

There are, however, two scenarios in which we might want to say that it would have been better had the Intervener not intervened, and in these cases a Moral Assessor may determine that an Intervener is responsible to some extent for the occurrence of harm to a Victim.

The first concerns cases in which another Deliberator (a would-be Intervener) did not get involved in the moral sequence due to the fact that there had already been an intervention—and, had he acted, this would-be Intervener's intervention would have been more effective. This is a problematic case and is a case on which a Moral Assessor may ruminate. In such a case, a Moral Assessor would have to consider the following facts. First, an Intervener's interventions may have precluded the would-be Intervener from acting, although the issue of whether the would-be Intervener should have acted first, talked the Intervener out of installing an ineffective barrier, or otherwise ensured that he installed his more effective barrier would likely ground a Moral Assessor's decision regarding the responsibility of the would-be Intervener—but in cases in which a would-be Intervener could have installed a more effective barrier, a Moral Assessor would likely conclude that the would-be Intervener forbore to prevent harm (bringing with it responsibility for the harm that occurred or would have occurred to a Victim). Second, it might not be possible to know with certainty that the would-be Intervener's intervention would have been more effective, but if it is the case that their intervention would have been more effective than the Intervener's intervention, a Moral Assessor could determine that the Intervener forbore to prevent harm and thus would determine that they are partially responsibility for the harm to the Victim. Importantly, though, in both cases the agent would be a Forbearer, not an Intervener.

The second scenario involves cases of snowballing, which will be attended to in the next section.

5.3.5. SNOWBALLING

To recap, snowballing is where an agent acts in a way that increases the severity of harm that will occur as a direct result of the initiation of the NPET.

Snowballing is a curious case in moral sequencing since a Snowballer (an agent who snowballs a moral sequence) does not initiate the moral sequence (he is not an Initiator) but his actions nonetheless result in (increased) harm to a Victim. Although the moral sequencing that I have presented in this thesis has largely been devoid of a discussion of cases of snowballing—so as to focus on the most salient features of moral sequencing and to ensure that a consistent and uncomplicated discussion of the practical and philosophical implications and ramifications of moral sequencing pertaining to deciding if, when, and how to intervene to prevent harm could be presented—to understand the extent to which intra-sequence agents are responsible for harm to a Victim we must consider whether a Snowballer is responsible for the harm that befalls a Victim.

The issue facing a Moral Assessor when deciding whether a Snowballer is responsible for the harm that befalls a Victim (and indeed the degree of responsibility) is that, as a direct result of their actions, a Snowballer increases the harm inflicted on a Victim without initiating a NPET. It is in this way that a Snowballer's action are directly connected to the harm to the Victim in the same way that an Initiator[159] is connected to the harm to the Victim—it is by their actions that harm (or in this case, increased harm) eventuates to the

[159] A Sustainer-2 is also similar in this regard, but since they do not directly act on the threatened harm but rather remove a dispositional barrier to that harm, a Sustainer-2 does have a *prima facie* different connectedness to the harm than in cases of initiating and snowballing—both of which physically act on the threat.

Victim. A Moral Assessor would likely determine that a Snowballer is responsible for the harm that befalls a Victim with respect to the increased harm over and above the harm that would have been sustained by a Victim had the Snowballer not snowballed the moral sequence. That said, shared responsibility is a peculiar thing and a Moral Assessor would likely find it difficult to determine just how much more a Victim was harmed as a result of a Snowballer's actions (e.g. would the Victim have merely been injured and not killed had the moral sequence not been snowballed?), and difficult further still to determine whether a Snowballer and an Initiator are equally responsibility, jointly fully responsibility, or otherwise responsible. However, I think that a Snowballer is at least responsible for the "extra harm" they caused to a Victim by virtue of their connectedness to *that* increased harm, and so a Moral Assessor should approportionate an amount of responsibility to the Snowballer over and above any responsibility ascribed to the Initiator without whom a Snowballer could not have snowballed the moral sequence. I do not think that both the Initiator and Snowballer are equally responsible for the harm caused to a Victim, though, since the Snowballer caused *more* harm than the Initiator threatened—although, as we have seen, this is a peculiar situation since if it were not for the Initiator's actions, the Snowballer would not have had the opportunity to snowball the moral sequence. Indeed, this provides an instance where a Snowballer may appeal to a Moral Assessor for a responsibility-decrease, the mitigating factor here being that they are only responsible for harm to a Victim *because* of the Initiator.

Linking back to the discussion in the previous section, snowballing may also occur because of a Snowballer's (failed) efforts to intervene in a moral sequence to prevent harm befalling a Victim. In such cases, a Snowballer may attempt to intervene in a moral sequence but, in doing so, increases the harm that befalls a Victim. These cases of snowballing are *prima*

facie cases of accidental harm. Here, a Snowballer intervenes in a moral sequence in a way that (accidentally) increases the harm to a Victim. The issue facing a Moral Assessor here is therefore that such a Snowballer intended to act in a way to avert harm befalling a Victim, but whose actions failed and, moreover, actually increased the harm to a Victim. This provides another mitigating factor. Although the fact that an increase in harm to a Victim occurred would provide the justification for a Moral Assessor to assert that a Snowballer (the type of intra-sequence agent) is responsible for the harm caused to a Victim, the same Moral Assessor may impose a responsibility-decrease on an individual Snowballer due to the fact that in this situation they clearly intended to avert or mitigate the harm, but simply failed to do so (and further increased harm). Importantly, though, linking back to the discussion of intervening in the last section, even though a Snowballer attempts to install a barrier to harm, the fact that through their intervention they actually snowball means that such an agent is a Snowballer and not an Intervener.

With the above two cases of snowballing in mind, it is clear that determining the responsibility of a Snowballer would require a Moral Assessor to make a decision with reference to theoretical considerations of moral responsibility; however, since it is not this chapter's aim to claim *which* theoretical considerations a Moral Assessor should subscribe to, the reader is simply asked to recognise that a Snowballer is the type of intra-sequence agent that is responsible for the increased harm to a Victim, although there are some mitigating circumstances that might decrease the responsibility of individual Snowballers depending on the situation, how they acted, their intentions, and so on.

5.4. CONCLUDING REMARKS

This chapter has assessed each type of intra-sequence agent in turn in order to assess whether a Moral Assessor would, post-sequence, have any grounds for attributing responsibility to those types of agents for the harm that befell a Victim or would have befallen a Victim without an intervention. I further offered some possible reasons that a Moral Assessor might have for increasing or reducing the responsibility of *individual* agents based on a number of (potentially mitigating) factors. I argued that cases of initiating, forbearing, sustaining-2, and snowballing are all cases in which those agents have at least some responsibility for (actual or expected) harm to Victim—although the degree of and extent to which responsibility is attributed is dependent on the theoretical considerations of moral responsibility held or employed by the Moral Assessor.

However, there is a potential problem related to responsibility, and in particular whether and to what extent an agent is connected to harm in a moral sequence, and this problem will drive the discussion in chapter 6. This problem concerns an inherent understanding that all intra-sequence agents act in a moral sequence with no inner-agent change. If, however, it is the case that there is an intra-sequence inner-agent change, then that agent may have grounds on which to appeal for diminished responsibility—they might claim, for instance, that they cannot be held responsible for any harm that did befall or would have befallen a Victim since they, say an Initiator post-sequence, are not the same as the Initiator intra-sequence.

The next chapter will address the concept of inner-agent change and will argue that, in order to determine whether a post-sequence agent has at least some grounds for claiming diminished responsibility for their intra-sequence actions, we need only establish that there has been a certain sort of change in personality. This, I will argue, can be determined without

relying on any metaphysical considerations of personal identity; we can consider ethical considerations related to inner-agent changes by taking a *personality approach* to drive ethical considerations related to claims of diminished responsibility.

After this is addressed, chapter 7 will come back to the issue of attributing responsibility to show how the *prima facie* problem outlined in chapter 6 can be accounted for (with the additional benefit of, in some cases, justifiably intervening earlier in a moral sequence than would otherwise have been warranted).

CHAPTER 6:

INNER-AGENT CHANGE

> ‘Why are you not “She”’?
>
> ‘Because “She” does not know the same things that I do.’
>
> ‘But you both have the same arms and legs, haven't you?’
>
> ‘Yes, but arms and legs do not make us the same.’
>
> —Morton Prince, *The Dissociation of a Personality* (1978: 27)

The reader should, at this stage, have a clear understanding of what moral sequencing is and how and why it has been developed, and how this system can be used to underpin a particular moral decision-making process (reactive-calculated moral decisions) applied to a specific type of moral decision (namely deciding to intervene) to yield a system with which one can assess both in real-time and retrospectively if, when, and in what way(s) one can intervene or if, when, and in what way(s) one could have intervened to prevent harm to a Victim. The reader should also have an understanding of how this system of moral sequencing can be employed, by a Moral Assessor, to determine the responsibility of intra-sequence agents post-sequence.

However, one of the most fundamental features of moral sequencing has yet to be identified, discussed, and taken into account: *inner-agent change*. Inner-agent change presents a problem for the system of moral sequencing proposed so far since the system cannot account for inner-agent changes—that is, changes pertaining to alterations in an agent’s personality (e.g. their disposition, mannerisms, etc.). Why is this important? Because an inner-agent

change during a moral sequence can have severe and significant normative ethical consequences. In particular, inner-agent change(s) of intra-sequence agents can have moral repercussions for the extent to which an agent is responsible for their actions, and, in some cases, can justify bringing forward the threshold of physical harm and intervening earlier in a moral sequence. The normative issues of responsibility related to inner-agent changes in moral sequences and the justifiability of intervening earlier will be discussed further in chapter 7.

6.1. THE PROBLEM CASE: INNER-AGENT CHANGE DURING A MORAL SEQUENCE

Consider the following two cases:

> *Car Thief*
>
> Thief steals a car and then sells the car to Casper. Casper does not know that the car was stolen and is not involved in the theft in any way[160].

[160] In order to motivate the problem case and the ensuing discussion, I ask the reader to assume that Casper's purchasing of the stolen car does not make him complicit in the theft or otherwise responsible for any of Thief's actions.

Comatose Thief

> Thief steals a car and drives it for a distance before crashing the car in which he endures a head trauma and goes into a coma. After waking up from his coma, Thief's behaviour, dispositions, idiosyncrasies, and even his memories have changed—jointly, these signify an inner-agent change[161,162].

In Car Thief, one would not likely say that Casper was responsible for stealing the car since he did not steal the car himself (and did not know that it was stolen). The reason that a Moral Assessor would likely not attribute responsibility to Casper is because Casper is a different agent to Thief and is not responsible for Thief's actions. But can the same reasoning be applied to Comatose Thief? *Prima facie*, we should say that in Comatose Thief a Moral Assessor should not hold Thief post-crash responsible for the actions of Thief pre-crash—like Casper in Car Thief, Thief post-crash in Comatose Thief has no knowledge of the car being stolen and appears to have undergone an inner-agent change. The relationship that Thief has to Casper in Car Thief is therefore *prima facie* similar to the relationship Thief pre-crash has to Thief post-crash in Comatose Thief.

---

[161] Of course, the extent to which these have changed and the degree of the severity of the change should be clarified here. But for current purposes and for simplicity I ask the reader to conceive of this change as being radical and as denoting a profound change in Thief's qualities pre-crash.

[162] This scenario is not unrealistic, and reflects real-life cases of head traumas, memory loss, and changes in personality, such as the cases of Donald (see Sacks, 2011: 169–173) and Phineas Gage (see Damasio, 2006).

This problem case highlights that the current picture of moral sequencing cannot make sense of any such *inner-agent change*, *viz.* a change in personality, where such changes are often explained with reference to personal identity and changes thereof.

Inner-agent change is an important component to account for since a Moral Assessor employs moral sequencing (post-sequence) to make normative ethical conclusions, in particular ascertaining the responsibility of intra-sequence agents; it is for this reason that the system of moral sequencing must be revised to account for the sort of inner-agent changes described above.

The rest of this chapter will provide a case for: (a) how we can make sense of the inner-agent changes that are paradigmatic of the case described above (and the empirical cases presented in §6.2.) without relying on an account of personal identity; (b) how inner-agent changes can be expressed as a change in personality; and (c) how changes in an agent's personality can be used to make normative ethical conclusions related to responsibility (including issues related to the concept of diminished responsibility) by incorporating an understanding of personality into the system of moral sequencing.

6.2. BILLY MILLIGAN AND OTHERS: EMPIRICAL CASES EVIDENCING THE NORMATIVE ETHICAL IMPORTANCE OF INNER-AGENT CHANGE

The empirical literature provides numerous examples that illustrate why accounting for inner-agent change has diverse and often serious normative ethical consequences. This section will present some real cases of inner-agent change to justify why this issue is worthy

of ethical assessment; the case of Billy Milligan in particular will be used as a basis for discussion for the remainder of this thesis.

Billy Milligan was arrested and put on trial for kidnapping, robbing, and raping three women in the United States of America in 1977 (see Keyes, 1995). Billy underwent a psychiatric assessment and was diagnosed with Multiple Personality Disorder (MPD), what is now referred to in DSM-5 (American Psychiatric Association, 2013) as Dissociative Identity Disorder (DID). Psychiatrists initially identified 10 personalities, including '[t]he original, or core, personality, later referred to as "the unfused Billy," or "Billy-U"' (Keyes, 1995: xi). The defence argued that, when committing the crimes, two of the other personalities, Ragen and Adalana, were in control of Billy-U's body, and so Billy-U should not be held responsible for the actions of Ragen nor for the actions of Adalana. The jury, accepting the defence's case, found Billy-U not guilty of the crimes. Billy was committed to a psychiatric hospital, the Athens Mental Health Center, and during his time there psychiatrists identified an additional 14 personalities; these had been suppressed by one of the "alters", Arthur, who 'revealed' them to Dr David Caul, a member of staff at the health centre (Keyes, 1995). Billy-U was confined to a psychiatric hospital until 1988, at which time psychiatrists believed that they had 'fused' all personalities together. This fusion came to fruition in 'The Teacher', the 24th personality; The Teacher declared 'I am Billy all in one piece' and referred to the other personalities as 'the androids I made' (Keyes, 1995: xiv).

Although the defence made a plea of "insanity" in the case of Billy Milligan, and that is the ground on which the jury came to their verdict of not guilty—not guilty 'by reason of insanity'—arguably this case demonstrates that issues related to inner-agent change can have significant normative implications. In Billy's case, an inner-agent change (or, more

specifically, changes in his personality) resulted in Billy (or rather Billy-U) having diminished responsibility for the actions of Ragen and Adalana. However, importantly, this case also demonstrates that one cannot escape the fact that Ragen and Adalana acted in the same physical body as Billy-U. As a result, although one might say that Billy-U (the person standing trial) was found not guilty of committing the crimes, because Billy-U is co-somatic with Ragen and Adalana, we might say that Billy (the physical body) must be attributed at least some responsibility—it was, after all, his body that harmed his victims. However, this case raises questions over the extent to which Billy-U (and indeed the other twenty-one "innocent" personalities co-habiting the physical body of Billy) should be held responsible for the actions of co-somatic personalities over which he and the other personalities had no control. This case illustrates how moral sequencing, as it is currently presented, cannot help to ascertain whether Billy should have been held fully, partially, or not at all responsible for the actions of Ragen and Adalana.

Indeed, there are other cases similar to Billy Milligan's that have profound normative implications. What follows is a selection of empirical cases that evidence the normative ethical importance of properly accounting for inner-agent change in moral sequencing. Arthur Abraham was committed to a psychiatric hospital after a jury found him not guilty ('by reason of insanity') of sexually assaulting his wife and later killing her along with their unborn child, and the defence successfully argued that Arthur committed the crimes whilst under the control of the personality of a deceased high school friend (see People v. Abraham, 1997). Juanita Maxwell worked in a hotel in 1979 during which her alternative personality, Wanda Watson, killed an elderly guest and, based on the stark difference in the temperament and dispositions between Juanita and Wanda, the judge committed her to a psychiatric hospital, but, on Juanita's release, Wanda went on to rob two banks (see Slovenko, 1995:

84). Mark Peterson had sex with a woman who he had recently met, but after their encounter Peterson was arrested for sexual assault since one of the woman's alternative personalities said that she did not consent to sexual intercourse, after which Peterson was charged and convicted for having sex with someone who was unable to give consent (see Imrie, 1990). Kenneth Bianchi, known as the "Hillside Strangler", was charged with multiple murders that he did not remember committing, but during hypnotism an alternative personality, Steve, of whom Kenneth was amnesiac, admitted to perpetrating the murders (see Finkel, 1988: 98). And, finally, Mr A was arrested for murdering and dismembering his wife, but Mr A said that he was amnesiac of the events and said that the crime must have been committed by his alternative personality, Billy Ray (see Perr, 1991).

These cases demonstrate the prevalence of the issue in the empirical literature and buttress the claim that inner-agent change, represented in these cases as changes in an agent's personality, is a serious normative ethical issue but one that the current system of moral sequencing cannot currently accommodate.

6.3. AGNOSTICISM ON METAPHYSICAL ISSUES OF PERSONAL IDENTITY

The sorts of inner-agent change presented in the problem case in §6.1. and in the empirical cases in §6.2. are often described and explained with reference to changes in an agent's *personal identity*. However, looking at the philosophical literature on personal identity reveals that theorists and thinkers have, first and foremost, been concerned with considering

metaphysical issues[163] in personal identity (with epistemological concerns[164] either being discussed in tandem or second), with normative concerns following; the implication is that normative concerns[165] can be drawn from or are born from metaphysical/epistemological positions. I will not comment on whether one should or whether it makes philosophical or pragmatic sense to assert that metaphysical and/or epistemological concerns are or should be explanatorily prior to normative concerns. What needs to be discussed are normative ethical concerns related to inner-agent changes, and this can be done *in vacuo* from any metaphysical or epistemological positions on personal identity.

Instead of favouring any particular metaphysical position on personal identity to help explain inner-agent changes, I am *agnostic* about the metaphysical considerations of personal identity. That is to say that whether Billy Milligan pre-crime is one-and-the-same as Billy Milligan post-crime is an issue that will not be attended to—what is of interest is rather, for example, whether post-sequence Billy should be held responsible for the actions of intra-sequence Billy. Simply focussing on metaphysical issues fails to get us into a position in which these ethical concerns can be discussed. I therefore propose that we speed-

---

[163] Metaphysical considerations involve determining what *A* consists in, under what conditions/criteria *x* can be said to be *A*, and under what conditions/criteria *A* can be said to be/remain *A*. The metaphysical issues of identity can be broadly construed as including discussions on: criteria of identity (necessary and sufficient conditions of identity, including considerations on the importance of the body, consciousness, memory, psychology, spatiotemporality, and so on); persistence conditions of identity (what it means to say that *A* persists over time); conditions under which identity changes or ceases to exist; and so on.
[164] Epistemological considerations involve ascertaining whether *x* is *A*. Epistemological concerns are often bound-up in metaphysical considerations; a natural progression from determining the conditions under which *x* can be said to be *A* is to then ascertain whether *x* is indeed *A*.
[165] Normative considerations involve ascertaining and understanding whether and in what ways personal identity has implications for and impacts on the real-world and the many sub-domains of ethics (including responsibility, justice, prudential concern for one's self and future-self, amongst others).

past the metaphysical concerns—remaining agnostic about these in the process—so that we can situate ourselves in a position from which the ethical considerations can be discussed.

One might accuse a proponent of such a stance, a "by-passer" of metaphysical and epistemological personal identity concerns, as drawing conclusions without foundation; "drawing normative ethical conclusions without a metaphysical and/or epistemological position on identity is like walking towards an object, illuminated on a pedestal, positioned at the end of a dark room, without care for what gems one might be blindly walking past and without care for an explanatory understanding of how it got there". However, drawing normative ethical conclusions based on particular metaphysical/epistemological positions in many ways ties those normative conclusions to those positions. And as the lively literature demonstrates, positions are dynamic, open to challenge and revision, and are, above all, plentiful. So, by adding normative ethical blocks to the foundational blocks of metaphysical/epistemological positions, the normative ethical understandings are in jeopardy of falling when the foundational positions are shaken (or even removed) through philosophical discourse. Moreover, in establishing a metaphysical or epistemological position on personal identity, positions are often, at least in part, founded on various intuitions on personal identity, which often give rise to various answers or at least foundational intuitions on a plethora of thought experiments and problem/puzzle cases. Yet if one accepts that these foundational intuitions on the metaphysical concerns on personal identity are normatively important in so far as they are often driven by (or at least partly founded on) ethical intuitions, then one could suggest that normative ethical and metaphysical/epistemological positions are philosophically married. There are of course a number of other methodological possibilities, including that normative ethical positions are explanatorily prior to metaphysical or epistemological considerations. However, the

intention of this chapter is to provide an argument for preserving the autonomy of ethical considerations by discussing ethical issues concerning inner-agent changes, often represented in the literature as a discussion on personal identity, apart from concerns of priority and methodology and apart from metaphysical/epistemological considerations; and one can do so by being agnostic on the metaphysical issues (*viz.* by not subscribing to a particular view or endorsing particular criteria).

6.4. THE PERSONALITY APPROACH: A NORMATIVELY RELEVANT UNDERSTANDING OF INNER-AGENT CHANGE

In the last section I argued that we have the conceptual foundations on which to build a normatively relevant understanding of inner-agent change apart from metaphysical considerations on personal identity and in a way that enables us to assess related ethical issues and draw ethical conclusions thereon. Importantly, I argued that this can be accomplished by being agnostic on personal identity; that is to say that ethical discussions can take place *about* personal identity *without reference to* and *without relying on* an account of personal identity. Instead, normative considerations and ethical conclusions can be made based on an assessment of personality, by *taking a personality approach to normative considerations*.

With the above in mind, what follows is a novel approach to discussing inner-agent change in a way that provides the platform for discussing the ethical issue of responsibility related to inner-agent change. This approach is not posited as a metaphysical criterion, nor do I claim that any position on personal identity must, as part of its conceptual analysis, include this requirement; importantly, this approach can be presented without relying on *any*

metaphysical claims. To this end, the rest of this section (§6.4.) will present what I call *the personality approach* to inner-agent change. This approach seeks to understand inner-agent change with reference to changes in an agent's personality instead of with reference to personal identity.

6.4.1. INDIVIDUAL PERSONALITY: A PHILOSOPHICALLY RELEVANT CONCEPT

In the *His Dark Materials* trilogy, Philip Pullman (2011) describes the personalities of characters by attributing to each character a 'daemon', which tracks and represents that character's personality. Daemons take the form of creatures (both actual and fictional), and the form of these daemons alter in children to reflect a change in the personality of the character to whom the daemon belongs; we see, for instance, Lyra's daemon, Pantalaimon, 'shape-shift' throughout the trilogy, changing from a moth to an ermine, eagle, mouse, and dragon, amongst other creatures. When a child matures into adulthood, these daemons 'settle' and take one permanent shape that reflects that character's personality, essence, or spirit. This is an interesting concept and embodies the common idea that a person can undergo many emotional changes that can affect their personality, but that there is a more fundamental personality that persists through these short-term changes. On this view, personality is described as 'more or less stable, internal factors that make one person's behaviour consistent from one time to another, and different from the behaviour other people would manifest in comparable situations' (Child, 1968: 83). This idea of personality persistence is polarized by two parallel debates in which it is often argued that we have the power to change our personality to the extent that we change as a person—as in the case of Saul's transformation to Paul on the road to Damascus (Acts 9:1–19). The idea of "one

body, one personality" has also been challenged due to the counter-evidence of certain mental disorders, most notably Dissociative Identity Disorder (DID), where the same person is said to embody not just one but multiple personalities (c.f. §6.2.).

The next section will assess the nature of personality—i.e. whether personality is largely persistent as in the 'settled' daemons or whether it is subject to change like Saul's transformation to Paul. I will offer an explanation of the nature of personality that will later be embedded into a revised system of moral sequencing by drawing a distinction between *traits* and *personae*: these are the two components of personality that are usually conflated in such discussions and both of which jointly constitute one's personality.

6.4.2. TRAITS AND PERSONAE

Ainsley

> Ainsley likes their freckles, their partner says they're charming and angelic, and Ainsley's partner is also fond of their lips, which they say are sensual. Ainsley is, however, conscious of the small scar on their chin, which they think is ugly and reminds them of the accident in which they received the scar. Ainsley tells their friends of how they dislike their pale skin tone, and is overtly conscious of their spots and blemishes, to the extent that they avoid socialising. Ainsley is told about a new range of cosmetic make-up and decides to try it. Ainsley decides against applying foundation so as not to mask their beloved freckles but uses concealer to hide their blemishes and scar. Ainsley uses bronzer to take the edge off their pale skin tone and applies lipstick to accentuate their lips. After applying the make-up, Ainsley's social anxiety is significantly reduced, and Ainsley feels confident enough to re-engage old friends and apply for a new job.

Just as one can use make-up to present a face to the world that is different to the one without make-up, so individuals can to some extent conceal their traits (inner dispositions to think, feel, etc. in certain ways) behind personae (social 'masks' presented to the world and to others). Ainsley's story is therefore analogous to the way in which *personae* (one's face concealed by different layers of make-up) can be used to conceal, alter, fabricate, or enhance existing *traits* (one's face without make-up); one wears a social mask to hide, distort, or enhance what I will call one's *genuine personality*. This section will detail these two components of personality, illustrate how they are different in kind but rely on each other

for a picture of one's personality, and draw a conclusion concerning the epistemic costs and pragmatic benefits of concealing, fabricating, or enhancing one's traits with personae. This will yield the personality approach to a normatively relevant understanding of inner-agent change.

6.4.2.1. UNDERSTANDING AND DEFINING 'TRAITS'

Many of the ways in which we discuss people's personality have come to be discussions of what we believe to be remarks on one's personality, but which are in fact linguistic guises that remark only on who one *seems* to be. "He has a good character (He's a good person)", "He has a positive disposition (He's got a good attitude)", "He acts kindly (He seems kind)". These are all common expressions, many of which I use when asked about my thoughts of or feelings towards a new acquaintance. But am I, in fact, saying much about my new acquaintance? He might appear to have a good character, but, like in the theatre, just because one plays a hero does not mean that one is a hero off-stage. He might be disposed to positive actions—being generous, being a good conversationalist, etc.—but this might only be because of a feeling of duty, because it is anti-social to act in other ways, or because of other reasons that otherwise conceal his disinclination to be like this, and further conceal his true proclivities. He might act kindly, but, in similar ways, does so only to ensure that others create a certain image of himself, to construct and maintain relationships in ways that further personal ends and that do not directly reveal his true nature, but instead serves as a mask that enables him to be who he wants to be, be who others want him to be, or be who he needs to be in order to fulfil certain goals or desires. Is there any way, then, in which we can talk of who someone *really* is? Can we ever know who someone truly is, *viz.* how can we ever identify someone's *genuine personality*?

I will now argue that there are two distinct components of personality, *traits* and *personae*, which jointly constitute one's personality. These types must be differentiated in order to highlight a difference between one's *genuine* personality and a *manifest/ascribed* personality. Understanding personality in this way enables us to distinguish between what it means to know *of*, that is, what it means to see only one's social mask, and what it means to know *that*, that is, what it means to see behind the social mask. Let us look at these in turn.

Traits are usually taken to be 'broad, enduring, relatively stable characteristics used to assess and explain behaviour' (Hirschberg, 1978: 45). Michael Eysenck (1994: 39) argues that the word 'broad' in this definition is 'crucial':

> 'Smiling, on its own, could not form the basis of a personality trait, because it is too narrow, but smiling, talkativeness, participation in social events, and so on, could together underlie a personality trait such as "sociability"'.

In other words, a trait is taken to be a category under which certain forms and patterns of behaviour resemble or are related to each other in a way such that they can be consistently identified as being representative of a single trait[166]. Those who support this idea of traits are usually referred to as *trait theorists*, in contrast to *type theorists* who argue that people's personalities fit into one of a number of specified categories. Eysenck (1994: 39–40) likens

---

[166] This echoes the definition of traits in the broadest sense possible as described and defined in contemporary literature. This is in line with, for instance, Saul Kassin's (2003) definition of traits as 'habitual patterns of behavior, thought, and emotion' (cited in Ożańska-Ponikwia, 2014: 10).

type theory to the difference between the astrological theory of star signs, where people fall neatly into one of 12 personality types according to which month they were born, or the Ancient Greek idea that 'melancholic', 'sanguine', choleric', and 'phlegmatic' are classes to which people's personalities could be assigned.

This chapter will not discuss in depth whether a particular theory more accurately represents how personalities should be defined and classed; what is of concern to us, philosophically, is an investigation of how to distance ourselves from the trajectory of the psychological literature's preoccupation with methods of assessing personality (i.e. the Minnesota Multiphasic Personality Inventory, Cattell's Sixteen Personality Factor Test, the Eysenck Personality Questionnaire, and so on)—and whose arguments are usually used in support of a particular theory of personality—and instead look at how to define personality in a way that can both make sense of inner-agent change and accurately reflect the kinds of inner-agent changes presented in the empirical cases discussed in §6.2.

There are a number of problems associated with determining both traits and types, but there are two glaring setbacks in current attempts to define and explain them: (a) it is doubtful that any list of personality traits/types can definitively encompass and represent the broad range of personalities that can be ascribed to people; and (b) current understandings of traits/types do not in fact describe the phenomenon of reflecting an individual's personality, but rather seem to define something philosophically distinct, namely personae. The concern outlined in (b) will discussed in §6.4.2.2., and what follows is a discussion of (a). The concern outlined in (a) can be summarised by Eysenck (1994: 53–54): 'Extracting too many factors leads to an unnecessarily complicated picture', for instance Raymond Cattell and colleagues (1970; see also Eysenck and Eysenck, 1985) identifying sixteen traits, 'whereas

extracting too few produces an unrealistically simple solution', for instance H. J. Eysenck (1991) identifying just three: neuroticism-stability, introversion-extraversion, and psychoticism-normality. Others argue for a more conservative number, most notably Norman's (1963) identification of five traits: extraversion (e.g. sociability), agreeableness (e.g. cooperativeness), conscientiousness (e.g. responsibleness), emotional stability (e.g. calmness), and culture (e.g. imaginativeness). This was later developed into the Five-Factor Model by McCrae and Costa (1985), Digman (1990), and Goldberg (1981; 1993), who formulated what has now come to be called the 'Big Five' theory of personality; the five traits identified by McCrae and Costa are: extraversion (*vs*. introversion), agreeableness (*vs.* hostility/jealousy), conscientiousness, neuroticism (*vs.* stability), and openness.

Arguing over the number of traits/types is philosophically moot, especially if one is sceptical of the techniques designed and employed to test whether an agent can be said to have the traits selected by, or that fall under a type proposed by, a particular theory. However, it seems strange, and rather Platonic, to say that there exists an optimal number of traits/types that can describe/represent the personality of every person, and that can accurately explain all kinds of behaviour. To say that there is an optimal number of traits/types is to say that there exists a number of traits/types that adequately encompass and explain all personalities; however, it soon becomes apparent that these categories either turn out to be too narrow (i.e. they fail to recognise the diversity, intricacy, and subtlety of personality) or are too broad (i.e. a large number of traits are assigned to an individual, or an individual cannot be legitimately placed in a particular type). By and large, the number of traits/types that a particular theorist posits as being the optimal number serves only to categorise people in a way that permits them to assess only how one *behaves*. Indeed, one claim, made by Goldberg (1981: 161), is that if one views the identification of traits as

revealing questions that one would wish 'to know the answers to […] about a stranger they were soon to meet', then one can view the classification of traits as an aid to gaining crucial knowledge about other people, and might even assist in determining the extent to which one would like to engage with a person. The traits identified in the Big Five, according to Goldberg (1981: 161), thus ask the following:

(1) Extraversion: 'Is X *active and dominant* or *passive and submissive* (Can I bully X or will X try to bully me)?'
(2) Agreeableness: 'Is X *agreeable* (warm and pleasant) or *disagreeable* (cold and distant)?'
(3) Conscientiousness: 'Can I count on X (Is X responsible and conscientious or undependable and negligent)?'
(4) Neuroticism: 'Is X *crazy* (unpredictable) or sane (stable)?'
(5) Openness: 'Is X *smart* or dumb (How easy will it be for me to teach X)?'

But what do these questions actually reveal about a person? Do they in fact reveal the crucial information about a person that Goldberg thinks they reveal?

Both trait theorists and type theorists are forever chasing their own tail; they theorise various categories of traits, and formulate/use tests to validate and vindicate these categories in an attempt to predict social, functional, or behavioural factors (e.g. how one might behave under certain conditions) and/or describe and characterise someone's personality; but these can only ever assess/describe how a person behaves, *viz.* who a person *appears to be*. In other words, using traits (and types, for that matter) to 'assess and explain behaviour' as symptomatic of a person's 'broad, enduring, [and] relatively stable' personality serves *only*

as an assessment or explanation of that person's behaviour; it does *not* assess or explain who that person *is*. If we look at Goldberg's questions in this light, then the answers to these questions are attained purely by an assessment of a person's behaviour, which does not necessarily represent that person's genuine personality, and reveals only who that person seems to be or who that person would like us to think they are. There is, of course, a separate debate to be had on the extent to which our behaviour reflects and represents our genuine personality, and one of course can, and it is entirely plausible that one might, act in such a way that accurately reflects one's genuine personality; but the fact that one can, and experience tells us that many do, act in ways that conceals or distorts one's genuine personality is testament to the fact that behaviour cannot be, and arguably should not be, used as the sole assessment of one's personality. The very thing that trait and type theorists attempt to assess and explain, *viz.* one's genuine personality, therefore cannot be wholly captured in any classification of traits/types, and cannot in fact be assessed or explained by using one's behaviour as revealing one's personality. This problem is exacerbated if one also views the tests used to assess/explain one's personality, whether they be self-report questionnaires, objective tasks, projective tasks, etc., as only further serving as an assessment of that person's behaviour, and not as an assessment of one's genuine personality.

In light of this, I propose that the term 'trait' should be reserved for a *core feature* of one's personality that can be but is not necessarily revealed by one's (habitual) patterns of behaviour but rather tracks, reflects, and denotes the persistence of forms and patterns of behaviour that resemble or are related to each other in a way such that they can be consistently identified as being representative of a single trait; in other words, traits represent one's *genuine personality*. This is not to say that traits are necessarily hidden or

secret (they are not strictly private), and that one's traits are forever epistemically concealed from other agents (they are not completely inaccessible to others); rather, by having epistemic access to another's traits, one can be said to know something about another's personal, often masked, self.

6.4.2.2. UNDERSTANDING AND DEFINING 'PERSONAE'

The type of personality that is revealed by an assessment of one's behaviour, and that both trait and type theorists seem to describe, are not 'traits' but rather 'personae' (singular: 'persona'), which represent the qualities that one recognises another agent as having, and then judges that agent's personality to be constituted, in part, by those qualities. In other words, unlike traits, personae are *ascribed* to an agent by *another* agent. A persona is formed solely by how an agent's *pattern of behaviour* is perceived and acknowledged by other agents. Agent *A*, for example, might believe that agent *B* is generous if *B* gave *A* some money to help alleviate *A*'s financial stresses. *A* might also consider *B* to be kind-hearted if *A* saw *B* helping a wounded bird. Personae can therefore be formed independently of the host agent's traits: *B* might in fact be selfish and might only be giving *A* money in exchange for *A*'s silence on *B*'s extra-marital affair. One could liken these personae to a *social mask*, where one acts in such a way as to facilitate others in constructing a persona that does not match one's traits[167], and whose roots can be found in Carl Jung's description of a persona as 'a kind of mask, designed on the one hand to make a definite impression upon others, and on the other to conceal the true nature of the individual' (cited in Conger, 2005: 111). On

[167] One might argue that it is the individual, not others, that ultimately creates a social mask; this is certainly true, however I use this terminology in order to differentiate between traits, *viz.* what is true of oneself, and personae, *viz.* what others *believe* is true of oneself.

this view, personae are representative of a Rousseauian social mask, where it is 'in the interest of men to appear what they really [are] not. To be and to seem [become] two totally different things' (cited in Lemmings and Brooks, 2014: 183). Indeed, it is conceivable that one might act in a certain way in order to project a specific persona so that others might come to believe that one is, for example, kind or generous, but where one in fact possesses neither quality. One might reasonably further associate personae to the Greek notion of 'mimesis' (imitation) and 'prosopopoeia' (impersonation). Under the personality approach I propose, personae therefore bridge the gap between concealing aspects of our genuine personality (traits) that we wish to remain private (either in general or in particular relationships/situations) and establishing a "desired image" or social mask of how we would like to be perceived by others in our social environment. Importantly, projecting and establishing personae allows us to interact in the world on "our own terms"—or rather facilitates others in establishing these personae. (Remember, we do not establish our own personae; we act in certain ways so as to *project* personae, in the hope that others recognise this behaviour as representing our personality.) In other words, social mask *x* is created for agent *A* by other agents (say, agent *B*) based on *A*'s patterns of behaviour that are ascribed to *A* by *B*, which subsequently affirms or reaffirms *x*.

6.4.2.3. THE EPISTEMIC AND PRAGMATIC COSTS AND BENEFITS OF CONCEPTUALISING TRAITS AND PERSONAE AS PRESENTED IN THE PERSONALITY APPROACH

Defining and understanding traits in this way has a number of benefits, most notably the benefits of being *independent of behaviour* and *independent of awareness and self-ascription*. First, as we have seen, although many psychologists and sociologists are of the

belief that one's behaviour is indicative of and reveals something about one's personality (indeed, one's behaviour might in some situations or contexts directly reveal one's genuine personality (traits)), as I have argued above one's patterns of behaviour do not necessarily expose one's traits, rather they bare only one's personae. One might act in a way that is inconsistent or that completely contrasts with one's traits. For example, I might have the trait "introvert", but, in order to create new and maintain existing relationships, and further my professional career and personal friendships, I might act in a sociable and extroverted way so that others come to believe that I am outgoing, and therefore ascribe to me a persona of "sociable" (remember, these personae are ascribed not as personae *per se*, but rather as supposed genuine representations of that agent's personality). As such, defining traits in a way such that an agent's behaviour is not considered indicative of or a genuine reflection of one's personality permits us to distance ourselves from the traditional psychological view that behaviour echoes personality (and vice-versa), and allows us to conceive of personality as denoting something deeper, more personal, than an assessment of one's behaviour.

Second, the issue of whether a trait can exist independently of the host agent's acknowledgement of that trait (i.e. whether one can be generous without knowing that fact) is answered. Indeed, it seems entirely plausible that one can be ignorant of one's own traits: despite the fact that I help and give money to others, I might not believe that I am generous or kind, for I might believe that I am merely doing my duty to help others, or I might believe that, in comparison to the generosity of those who donate millions of pounds to help fight poverty or the kindness of those who sacrifice much of their free time to working for charities, my acts do not constitute generosity or kindness. But even in these circumstances, where I am ignorant of or do not self-ascribe certain traits, it seems as though one would be missing something essential to my personality if I was not described as being kind/generous

based solely on the fact that I do not acknowledge or self-ascribe these traits as core features of my personality. As a result, we have firm grounds—at the very least empirical grounds—to declare that traits can exist despite a host agent's ignorance of, or unwillingness to self-ascribe, those traits. Omitting the view that traits must be self-ascribed and that one must be aware of one's traits from our definition of traits has a further benefit: it prevents delusions or confabulations of traits from being considered. It is plausible that one might delude oneself into self-ascribing traits that do not accurately represent one's genuine personality, or one might confabulate traits (due to, for example, feeling ashamed of one's genuine personality or not wanting to believe that one has certain traits); but by viewing traits as being independent of one's awareness of one's traits, we avoid a conflict between one's self-ascribed traits and one's traits being an accurate representation of one's genuine personality. This raises the following two closely related questions:

(1) If I can have a trait without necessarily being aware of it, how can I, or anyone else for that matter, ever correctly identify my traits?

(2) How can I know that I have correctly identified a trait as representing my genuine personality (if I cannot be sure that I have not self-ascribed a delusional or confabulated trait)?

In response, we need to ask: "Does it matter if I cannot identify, am not aware of, or cannot self-ascribe my genuine traits?" An inability to correctly identify one's traits is a serious epistemic burden on self-knowledge: a misidentification or ignorance of traits renders one less self-aware and less self-knowledgeable than one would be if one were aware of, and were able to correctly self-ascribe, one's genuine traits. This also has pragmatic consequences, the consequences of which are a double-edged sword: the misidentification, ignorance, or inability to self-ascribe genuine traits can have serious pragmatic

consequences (i.e. in the case of violent or "dark" traits), but can equally have few, if any, pragmatic consequences in terms of other (non-violent) traits. What is also interesting is that, when the pragmatic costs are low, an epistemic benefit might ensue.

When one is ignorant of one's genuine traits, one does not *feel* as though one knows less about oneself, because by being ignorant of *x*, one is ignorant of the *fact* that one is ignorant of *x* (until it is pointed out that one is ignorant of *x*). Even if one thinks that one has trait *v* when one in fact has trait *w*, apart from the epistemic costs for self-knowledge, there can either be huge or few pragmatic consequences of being ignorant of *w*, depending on the kind of trait one has confabulated, is delusional of, is ignorant of, or has incorrectly self-ascribed.

In terms of what I will call "low-impact traits", *viz.* non-violent and non-"dark" traits, one can continue to live in a way that is wholly consistent with who one *thinks* one is without impacting on one's life or others' lives, lifestyle, or health. I can, for instance, think that I am generous—perhaps because I have confabulated that I once set up a monthly direct debit to a charity, when in reality no such direct debit was ever established—without having any long-lasting effects on, or ramifications for, how I live my life, the quality of my life or the life of others (notwithstanding those who aren't benefitting from my supposed generosity), or my health or the health of others. If one of my friends challenged my generosity, and he demonstrated that I could not be generous because I do not donate money to charity, then I might re-evaluate my conviction that I am "generous", and this counter-evidence might even serve to eradicate my confabulation. On the other hand, it is conceivable that the opposite might happen, and I may become so convinced of my supposed trait (perhaps I have falsely ascribed it due to a delusion or confabulation originating from a trauma, brain injury, etc.) that I eventually *adopt* that trait, and I *become* generous. This makes one think of a line in

Orson Scott Card's novel, *Ender's Game* (1985: 231): 'Perhaps it's impossible to wear an identity without becoming what you pretend to be'. If this were possible, there may even be an ensuing epistemic benefit that comes with the adoption of a previously confabulated or delusional low-impact trait: where one might have been ignorant of one's genuine traits, one now has certainty in the fact that one possesses this adopted, previously incorrectly self-ascribed trait. One cannot discount the possible pragmatic benefit of this adoption either: if one was not previously generous, and one became generous through adopting this trait, then, so long as one holds "generous" to be an admirable and positive trait, one could argue that adopting a non-genuine trait as a genuine trait can have the added benefit of making one a "better person", or at the least the benefit of enabling one to adopt a positive trait that one would otherwise not have.

This might be true of low-impact traits, but the same cannot be said for "high-impact traits", *viz.* violent and "dark" traits (which I take to reflect Nicholas Holtzman and Michael Strube's (2012) 'Dark Triad' of Machiavellianism, narcissism, and psychopathy), the presence of which can negatively affect one's life or others' lives, lifestyle, or health. There can be serious, and sometimes devastating, pragmatic consequences associated with high-impact traits, and so, to avoid these pragmatic costs, the adoption of such traits should be avoided and, where possible, deterred. Consider a case where a kind, caring, and altruistic child, Harry, becomes friends with those at school who are members of a local gang, known for terrorising local residents and committing various crimes. Whilst spending time with the gang, Harry likes to think that he, like the other gang members, is violent, as this maintains their position as a formidable gang; but, whilst at home, Harry is a caring brother to his siblings, regularly helps with household chores, and deeply and openly loves his girlfriend. After spending a considerable amount of time with the gang, however, Harry adopts this

previously non-genuine violent trait, and becomes a violent person; he argues relentlessly with his siblings, fights with his father, and physically abuses his girlfriend. Although this case is extreme, it exemplifies the idea that, far from being likened to low-impact traits having few pragmatic consequences, adopting high-impact traits can have serious pragmatic costs. Deluding oneself into self-ascribing or confabulating non-genuine high-impact traits or being ignorant of or misidentifying high-impact genuine traits can have serious pragmatic costs if these high-impact traits are adopted; one becomes a pragmatically worse person in adopting these traits, so long as one holds violent and "dark" traits to be negative and detrimental traits[168].

The epistemic costs of being ignorant of one's genuine traits or deluding oneself into thinking or confabulating non-genuine traits is therefore high and hinders one ever really being able to say anything true of oneself, which is not only epistemically unsatisfactory but philosophically problematic. Moreover, depending on whether these traits are low-impact or high-impact, they can have either few or serious pragmatic consequences respectively. Adopting low-impact traits has few, if any, pragmatic consequences for one's or others' life, lifestyle, or health; whilst adopting a high-impact trait can have serious pragmatic consequences for one's or others' life, lifestyle, or health.

This said, ignorance of one's own genuine traits does not entail that others cannot identify one's traits. But how, then, can others identify my traits (especially if I am unable to identify

---

[168] This is, of course, context-dependent: adopting a high-impact trait if one is in a hostile environment, i.e. warzone, could equally have pragmatic benefits, i.e. making a soldier more efficient in their duties as a soldier. Nonetheless, we can maintain that, on the whole, high-impact traits have detrimental pragmatic consequences due to their generally negative effect on one's life and others' lives, lifestyle, and health.

them myself)? The answer lies in the cosmetic make-up analogy (see §6.4.2.). Ainsley used cosmetic make-up to display their facial assets, accentuate their favourite features, and conceal what they considered to be undesirable features. However, Ainsley wears different make-up for different social contexts and for different people. For instance, Ainsley wears an amount of makeup that conceals all "faults" and "flaws" and highlights/accentuates all "positive features" and "assets" when in work, but wears less make-up—that is, Ainsley conceals less and reveals more of their genuine personality—when with friends, and even wears no make-up whatsoever when with their partner. Person *B* can therefore identify person *A*'s traits in two ways. First, *A* could drop their social mask (remove the make-up that conceals and/or accentuates their genuine features); this is usually the case with close friends, relatives, and/or one's spouse/partner, that is, with people one feels comfortable being "oneself". Second, *B* could spend enough time with *A* such that *A*'s mask "slips" (or, to maintain the analogy, their make-up "rubs off"), and where *B* can start to see more of *A*'s genuine traits and, eventually, perhaps even *A*'s genuine personality; this is usually the case with long-term friendships or relationships. It is, however, conceivable that one might never feel comfortable enough to reveal one's whole genuine personality, might be ashamed to reveal "dark" or "undesirable" traits, or might be on constant guard so that one's mask does not slip, and one's make-up never fades when with others. In such situations, it might be difficult to know of someone's genuine traits. However, one is unlikely to be able to maintain such a façade forever; one's make-up will, if only temporarily, fade; one's mask will, if only for a moment, eventually slip.

Turning attention to personae, one's social mask could have a number of pragmatic benefits, although it also has severe pragmatic costs. Projecting a positive persona might have a number of pragmatic benefits: if, for example, one has sexist thoughts and persistently uses

foul language in one's own home, then making these sexists thoughts known to one's colleagues and using foul language in the workplace could easily jeopardise one's job-security. However, if one acts in a way that projects a positive persona—outwardly behaving as though one endorses sexual equality and refrains from swearing in the workplace—then one could ensure an element of job-security. Projecting a negative persona can be equally beneficial in certain situations: if one is subject to verbal attacks whilst at a pub with one's friends, and if one foresees the situation becoming hostile, then one—who is usually placid and peaceful—might act in an aggressive manner so as to project a persona that intimidates the aggressor, subsequently reducing the likeliness that the antagonist will attack.

There are, however, severe pragmatic costs associated to personae; by projecting a persona that does not accurately reflect one's personality—or, more specifically, one's traits—it is difficult, if not impossible, for others to make an accurate assessment of one's genuine personality. If an agent is only ever assessed on his personae, it is epistemically specious to make an assessment or judgement of his personality on these personae alone. Imagine that you wish to hire a Personal Assistant (PA), and you would like your PA to possess certain personal qualities to ensure that your PA is suited to both your personal and business needs, say, loyalty and efficiency. Now, if during an interview a candidate has suitable qualifications and references, and, based on his references and the way he conducts himself during the interview, then you might reasonably assess that he has the personal qualities you desire—based on his loyalty to previous employers and his efficiency in previous jobs—and so you decide to hire him as your PA. But let us further imagine that these qualities are in fact fabricated, and his behaviour in the workplace is solely aimed at projecting a persona that makes others believe that he is efficient and loyal; perhaps outside the workplace, in personal or family matters, he is completely inefficient and disorganised, and is prone to

being disloyal to his partner and regularly snipes at his friends behind their back. To what extent have you hired a loyal and efficient PA? Granted, the pragmatic costs of hiring a PA that only ever *acts* as though he is loyal and efficient are fairly minor; so long as he consistently acts as though he is loyal and efficient in the workplace, this is arguably enough in this context. Moreover, his diligence in projecting these qualities might even have the pragmatic benefit of making him a good PA. Whether or not your PA *is* loyal and efficient, that is, whether these two qualities are his traits and not just his personae, is fairly epistemically insignificant in the context of a professional (employer-employee) relationship. But the same cannot be said of other relationships; in other contexts, the pragmatic costs can be more severe. Imagine a different scenario where you meet a person at a pub, and after a period of conversing with them you decide that they possess the qualities you admire, say, openness and honesty, and so decide to invite them to meet you again. Further imagine that after a period of time these admirable qualities are persistent and you decide that they would make for a good partner, and so move beyond friendship and enter into a relationship with them—all the while believing that they are open and honest. But what if they were not as open or honest as they made out; perhaps they concealed the fact that were already married, were not open about the fact that they had a sexually transmitted infection, or had withheld the truth about being a parent to an estranged child. In a personal (platonic, sexual, or marital) relationship, the pragmatic costs of personae can therefore be severe. Making a judgement of someone's personality based solely on their personae in certain contexts, specifically personal relationships, can be detrimental to the legitimacy of the relationship, can obscure judgements pertaining to the relationship, and can aid and abet decisions that would not be made if you were informed of the truth. This is not to say that there might not be pragmatic benefits, for if your partner maintained and reinforced their personae then this might, at least in the short-term, assist the continuation of a happy

relationship; but, these pragmatic benefits in personal relationships arguably outweigh the pragmatic costs in a way that is not so in professional relationships.

6.4.2.4. RECONCILING TRAITS WITH PERSONAE

After such a discussion, one might be led to believe that one's personae cannot reveal anything of philosophical value about one's personality. But this is far from the truth. Personality comprises *both* one's traits and one's personae. One's traits reflect core features of one's self without which we would know little about an agent; *traits constitute who an agent, as an individual, is as a person*. On the other hand, *personae tell a story about who an agent, as an individual, is as a character*, and this imparts important information in itself. The wearing of a social mask tells us something about a person; they put on that mask, that is they act in a certain way, for a reason, whether it be for fear of admitting who they are to strangers, concealing home-truths from their employers in order to keep their job, hiding aspects that might damage their relationship with friends or family, or for a number of other reasons. Similarly, the amount of make-up one wears (c.f. the Ainsley analogy in §6.4.2.) tells us something about how that person feels about us as a person, about how they view our relationship and the social situation, and reveals a layer of information about oneself; camouflaging what one considers to be undesirable traits can show one is cautious, anxious, or worried about being open, it might indicate that one wishes to maintain only a professional relationship rather than allowing others to see personal aspects that are only revealed to one's friends, it might show one to be shy, or it could tell a completely different story. Only by viewing personality as being constituted by both traits and personae are we in a position to say anything philosophically, epistemically, and pragmatically meaningful about an agent, discover anything about that agent, make a judgement of how that agent is

likely to act (c.f. chapters 2 and 3 and §6.4.4.), and establish normative conclusions on inner-agent change (see chapter 7).

6.4.3. EXPLAINING CHANGES IN PERSONALITY IN THE PERSONALITY APPROACH

So how *can* the personality approach, focussing on personality as comprising a traits/personae dichotomy, be used to make sense of changes in personality? The issue is essentially a question of how one knows that there has been a change in an agent's personality. When making the distinction between traits and personae, one essentially makes a distinction between one's genuine personality (traits) and one's ascribed/manifest personality (personae). With the definitions and understandings of traits and personae presented in §6.4.2.1. and §6.4.2.2. respectively, the personality approach must rely on explaining inner-agent change by reference to a change in traits and not a change in personae, since traits are one's genuine personality. A change in agent *A*'s personae would only reflect either that agent *B* has changed their ascription of personae to *A* or that *A* has decided to manifest different personae to *B*. The problem, however, is that *A*'s traits are, by virtue of their nature, not directly epistemically accessible to *B* (and, as I mentioned in §6.4.2.3., sometimes even epistemically inaccessible to the person to whom the traits belong), and so *B* would *prima facie* not be able to assert that there has been a change in *A*'s

traits[169]; *B* would *prima facie* only be able to say that *A*'s personae have changed, since *B* only has direct epistemic access to *A*'s personae. The problem can therefore be stated thusly: a change in personality (*viz.* an inner-agent change) requires establishing a change in *A*'s traits, but *A*'s traits are *prima facie* not directly epistemically accessible to *B*. (It is, of course, entirely plausible that *B* can ascribe personae to *A* in a way that accurately reflects *A*'s traits, and likewise *A* can act in ways that enables *B* to ascribe personae to *A* that is representative of *A*'s traits, but the issue is that *B* will never be in a position to say with any epistemic certainty that *A* has trait *x* and that *x* has changed to *y* or that *x* has vanished.)

The waters are muddied further when one considers that the key difference between personae and traits is that personae are *manifested* by an agent and are *ascribed* to an agent by other agents, whilst traits are core features of an agent's *genuine* personality and are *noticed* by others and/or oneself. The following questions now arise: How can I be sure that

---

[169] One might argue that traits *are* directly epistemically accessible if an agent is honest about one's traits. If agent *A* confides in a close friend *B* that *A* has trait *x*, then one might say that *B* has direct access to *A*'s trait. Equally, if *A* behaved in ways that only manifest personae that aligned to *A*'s traits and others' ascriptions of personae to *A* aligned with *A*'s traits, then one might equally say that one has direct epistemic access to *A*'s traits. However, in the former claim, the epistemic limits of testimony prohibit an agent from being able to assert with certainty that *A* has *x*; *B* is left trusting the testimony of *A*. So although one might want to say that in this situation *B* has direct epistemic access to the fact that *A* has *x*, testimony alone cannot arguably ground epistemic truth and should not be taken as firm ground on which *B* can be said to have direct access to *A*'s traits. Indeed, *B* only has indirect epistemic access because *B* cannot *himself* ascribe *x* to *A*; one might therefore say that *B* has direct testimonial access to *A*'s traits but *B* does not have direct epistemic access to *A*'s traits. (One might balk at this suggestion and point their finger at any scientific theory, say that the world is spherical and not flat, and say "You believe that the world is not flat even though you have not seen evidence of this yourself. You trust the testimony of scientists, so why not trust the testimony of others about their traits?" The issue of trusting testimony is an issue that I do not wish to debate, for it would take us off-topic, but one cannot escape the fact that testimony relies on taking as fact that which to one does not have epistemic access, otherwise one could be said to *know that* (*viz.* have direct epistemic access to) *x* is the case.) Similarly, in the latter claim, *B* would only have direct epistemic access to *A*'s personae regardless of whether these personae accurately aligned with *A*'s traits. Essentially, the very nature of traits makes them directly epistemically inaccessible.

I am noticing another's traits and not merely ascribing them personae? How can I be certain that I have correctly identified my own and another's traits? And how can I know someone is not behaving in a way that ensures that I constantly ascribe to them the personae they desire? Well, I can't be categorically certain; this is the ultimate epistemic and pragmatic cost of these two components of personality. I can only ever make warranted assertions of one's traits based on what I consider one's traits to be, and this assessment is made solely on how one behaves. And herein lies the problem: we cannot consider ourselves to know of an agent's genuine personality until we gain an epistemic foothold in accessing one's traits, that is to say that we cannot claim to know one's genuine personality based solely on one's personae (i.e. how one behaves), but, crucially, traits are *prima facie* inaccessible to other agents. However, the personality approach can overcome these issues.

A change in personae, although epistemically inferior to a change in traits, does denote something epistemically meaningful and pragmatically useful. The very fact that *A*'s personae have changed reflects some sort of change in *A*'s personality. Say that I have violent thoughts, intentions, etc. to the extent that I (or an impartial omniscient observer) would ascribe "violent" to me as one of my traits, but I recognise that in order to maintain relationships, hold down a job, etc., I decide to present myself as being non-violent (and so act in ways that ensures that others ascribe to me a non-violent persona). If I then act in a way that is perceived to be violent (whatever that may be, whether it be screaming at someone or physically hurting someone), this change in behaviour reflects something about me as an agent. So long as this change is not persistent or enduring, others will likely not re-ascribe my personae—they will likely not say that I am now violent. Rather they will likely say that I was "acting out of character". Although this is not a change in my traits, it does reflect some sort of change in my personality, a "persona change" rather than a "trait

change" (regardless of whether this "out-of-character-you" is temporary or more enduring), and this change is what is pragmatically useful and what can be used as a basis for drawing normative ethical conclusions—a change in personae provides us with a quick and epistemically available method for attributing a change in personality within that agent. It therefore seems that one has to rely on these persona changes to assess a change in personality.

One might say that this is problematic, for when an agent modifies their behaviour (which forms the basis for persona and trait assessments) all that agent is doing is responding to external stimuli, and this cannot represent anything meaningful about changes in that agent's personality and cannot ground claims of personality (inner-agent) changes. However, the way in which an agent responds to external stimuli does say something about that agent—in this case it says that they have responded to a situation by behaving in a way that is inconsistent with (what others have perceived to be) their normal behaviour and is therefore behaving in a way that is inconsistent with that agent's personae—and this reflects some sort of personality change in that agent.

However, one might be led to argue that the nature of behaviour—including its diversity, subtleties, and intricacies—does not neatly lend itself to being employed by others to notice and ascribe subtle changes in an agent's behaviour to ground claims of a personality change (both trait changes and persona changes). For instance, does raising my voice count as a persona change or should it be reserved for more serious or drastic fluctuations in behaviour? In short, this objection would state: "You can't pinpoint exactly when a change has occurred". However, demanding a definition of what a change *is* is essentially demanding a solution to the problem of vagueness, but such a discussion lies outside of the

scope of this thesis. Appealing to the Sorites Paradox, the issue is that small changes over time don't seem to add up to a new personality (gradually adding grains of sand to a single grain of sand does not facilitate being able to demarcate when the amalgamation of grains becomes a "heap"), but equally working backwards one couldn't say there has been a change (gradually removing one grain of sand from a heap of 1,000,000 grains of sand does not facilitate being able to demarcate when the amalgamation of grains ceases to be a "heap"), even though there does appear to have been some change (both cases of adding grains to the single grain and removing grains from the heap demonstrates that we can say that, in the former, at some point gradually adding grains will yield a heap and, in the latter, gradually removing grains from the heap will result in the heap disappearing, but we cannot pinpoint the grain that tips the scales in either case). The change lies somewhere, vaguely, in the continuum. The personality approach is therefore content with stating that a change has occurred, even though it cannot determine the point at which the change has occurred.

The personality approach can therefore make sense of personae and persona changes, but it has not yet explained how one can make sense of seemingly epistemically inaccessible traits and trait changes. What follows is a discussion of how the personality approach can account for indirect epistemic access to assess traits on which one can make warranted assertions about trait changes.

The personality approach must be able to explain trait changes—a persona change on its own is just a thin veneer of personality change—and, since traits are not directly epistemically accessible to external agents, trait changes (and therefore personality changes) cannot be directly detected. As mentioned above, traits are by their very nature not directly epistemically available to external agents. However, traits are *indirectly* epistemically

accessible—one is warranted in asserting that an agent *A* has trait *x* based on information that is epistemically accessible. If one assesses a trait on a scale of verisimilitude[170], one is warranted in asserting that *A* has *x*; the higher the verisimilitude, the stronger one is warranted in asserting that *A* has *x*. To elucidate this claim, consider the following example.

[170] I use the term "verisimilitude" out of its original context in the philosophy of science, usually attributed to Karl Popper (1976), where it is employed to gauge "truthlikeless". I use the term to refer to a scale on which a higher verisimilitude permits one to warrantly assert that it is more likely the case that *x* than ~*x*. The scale of verisimilitude on this understanding is a continuum on which, at the bottom (lowest) end, there is no verisimilitude and where one can assert that ~*x* is the case, and, on the top (highest) end, there is complete verisimilitude and where one can assert that *x* is the case. On the continuum, the higher one moves, the more verisimilitude there is, where one can warrantly assert that it is more likely that *x* than ~*x* than before.

Harry

Harry is a coward. When confronted with an undesirable situation, when intimidated, when fearful, and so on, Harry wants to take flight. Harry is, however, aware that being labelled "a coward" is undesirable; it is not a trait that one boasts about, it is not a trait on which dinner party stories are usually based, and it is not a trait that lends itself to the affection of others. Because of this, Harry has manifested a heroic persona, and his behaviour (saving stranded cats in trees, making a citizen's arrest, chasing-down a youth whom he saw steal a man's wallet, and so on) has enabled others to ascribe to Harry a heroic persona. To others, Harry would be described as a hero. One day, walking to work with his partner, Harry walks in close proximity to an area undergoing a terrorist attack. He hears an explosion, gunfire, screams, and cries for help. Harry is aware that he should seek to help those in need—"Help! Help me!" he hears. He is aware that a heroic person would likely rush to their aid, and he is acutely aware that his partner expects him, as the heroic person they believe him to be, to help those in need. But Harry's heroic persona cannot overcome his fear, and his cowardice cannot hide from such terror. Harry grabs his partner's hand and runs.

The point of this example is to introduce the idea that some traits that are concealed by manifest personae cannot remain hidden in every circumstance; situational and circumstantial contexts, if not in totality at least in part, reveal one's trait(s). These situations are likely to be high-pressure or high-stake situations that impose danger, terror, panic,

anxiety, desperation, helplessness, anger, frustration, severe challenge, and so on. Now one might say that Harry acted not out of cowardice but out of self-preservation, or out of love for his partner, or for any number of other reasons. I grant that the example itself is subject to criticism, but the example provides one possible case, and illustrates that there are certain circumstances, in which one's social mask slips and in which others can glimpse one's traits. This is not to say that one would necessarily be able to have direct or full epistemic access to another's traits, but one can indirectly access another's traits when such situations arise that challenge the personae one has ascribed to another agent. The more indirect information one can gather over the length of acquaintance, friendship, marriage, or one's lifetime, the more one is warranted in asserting that agent *A* has trait *x*. It is therefore possible, once one has gathered enough information on an agent's traits, to notice and be warranted in asserting that there has been a trait change. If, in the case of Harry, his partner notices other instances of cowardice (but not manifest by Harry and not ascribed to Harry by his partner), his partner could maintain their ascription of a heroic persona to Harry but also, simultaneously, be warranted in asserting that Harry has the trait "coward". However, if, when confronted with similar situations in which his cowardice was previously revealed, Harry was heroic, one might update their previous warranted assertion to say that Harry has had a trait change and now has the trait "hero".

The personality approach therefore, by virtue of the definitions of the components of "traits" and "personae" which jointly constitute "personality", requires a trait change to determine a personality change. However, as we have seen above, this cannot be categorically asserted since one can only ever gain indirect epistemic access to an agent's traits. One therefore has two methods for assessing "strict" and "weak" personality changes related to trait changes and persona changes respectively. A persona change can be identified simply by an

assessment of an agent's patterns of behaviour because one has direct epistemic access to that agent's behaviour. However, this type of personality change is weak because a change in *A*'s personae would only reflect either that *A* has decided to manifest different personae to *B* or that person *B* has changed their ascription of personae to *A*. Either way, the personae manifested/ascribed do not necessarily reveal the 'broad, enduring, relatively stable characteristics' (Hirschberg, 1978: 45) of *A*; *B* is left without epistemic access to *A*'s core features of *A*'s personality and so cannot be said to have epistemic access to *A*'s genuine personality. To say anything meaningful about *A*'s genuine personality and to assess strict personality changes, *B* requires epistemic access to *A*'s traits (to both ascertain *A*'s traits and then assess if there have been any trait changes in *A*). However, *A*'s traits are by their very nature not directly epistemically accessible to *B*. *B* therefore *prima facie* cannot make any claims about *A*'s traits nor, therefore, any trait changes in *A*; indeed, it seems as though *B* cannot make any claims about strict personality changes in *A*. However, *B* can be warranted in asserting (but cannot not categorically assert) that there has been a strict personality change in *A* by an assessment of *B*'s indirect epistemic access to *A*'s traits.

So, although the personality approach must admit that it cannot account for categorical assertions that there has been a strict personality change in an agent, and this is clearly an epistemological barrier, this does not impact on the epistemic and pragmatic usefulness of the approach for normative considerations and drawing ethical conclusions about inner-agent changes. The personality approach, which side-steps metaphysical issues about personal identity in order to quickly reach a position from which normative assessments about inner-agent change can be made, is therefore a viable option regardless of the epistemic barriers to traits. The personality approach is content with employing directly epistemically accessible personae and personae changes to make claims about weak

personality changes and employing indirectly epistemically accessible traits and trait changes to make a warranted assertion that there have been strict personality changes, whilst recognising that it must remain silent about categorically asserting that there have been strict personality changes.

6.5. REVISING OLD AND ADDING NEW SEQUENCE ARCHETYPES TO ACCOUNT FOR PERSONALITY IN MORAL SEQUENCING

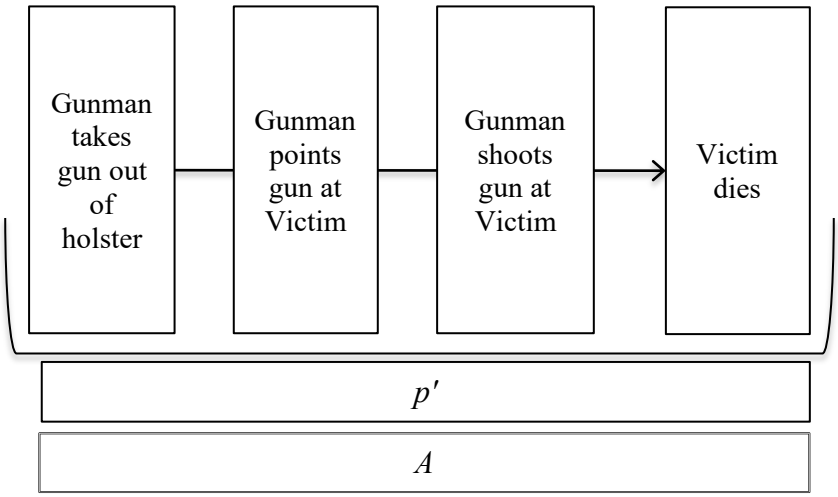
This chapter has established the need to account for personality changes in moral sequencing and has provided an account (the personality approach) to explain what it means for an agent to undergo this change.

In chapter 2, two moral sequence archetypes were presented in §2.2.1. and §2.2.2.: a Moral Sequence without Intervention (MS) and a Moral Sequence with Intervention (MSI), respectively. This section will revise and redress these two sequence archetypes and will add two new archetypes to account for and properly embed personality and changes in an agent's personality into moral sequencing.

6.5.1. (REVISED) MORAL SEQUENCE WITHOUT INTERVENTION

What separates these archetypes from the original archetypes presented in §2.2.1. and §2.2.2. is that this archetype has the concept of personality built into the framework.

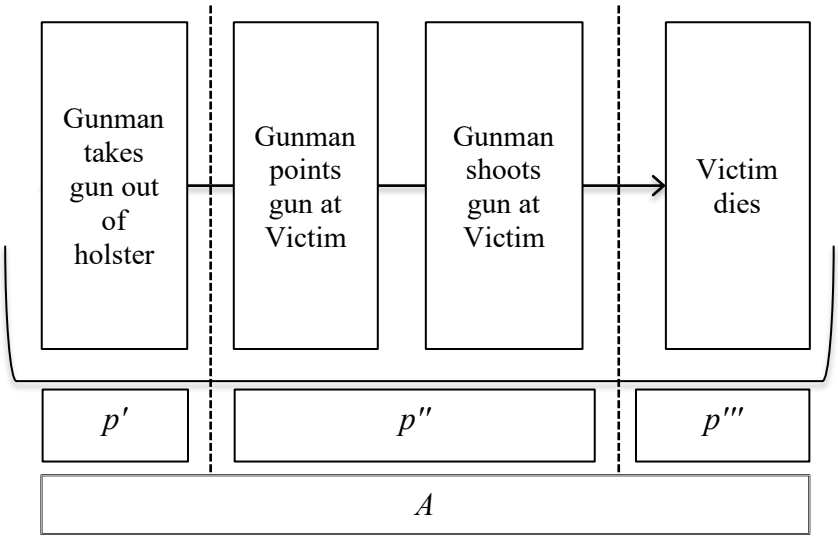
Moral Sequence without Intervention (MS)



This is an archetype of a moral sequence without intervention in which a distinct series of sequence-events occur (the progression of which is represented by each solid box, and the succession of which is indicated by the direction of the solid arrow) but where there is no intervention. Sequence-events are assigned an ordinal number according to the order in which the sequence-event occurs until the final sequence-event, which I will call the sequence-outcome. The initiation of the moral sequence is sequence-event 1, where 'Gunman takes gun out of holster', sequence-event 2 is 'Gunman points gun at Victim', sequence-event 3 is 'Gunman shoots gun at Victim', and sequence-outcome is 'Victim dies'. During the series of sequence-events, Gunman (the agent, represented by the upper-case *A*) initiated the sequence and acts throughout the sequence with a persistent personality (represented by *p′*).

6.5.2. MORAL SEQUENCE WITHOUT INTERVENTION BUT WITH PERSONALITY CHANGE

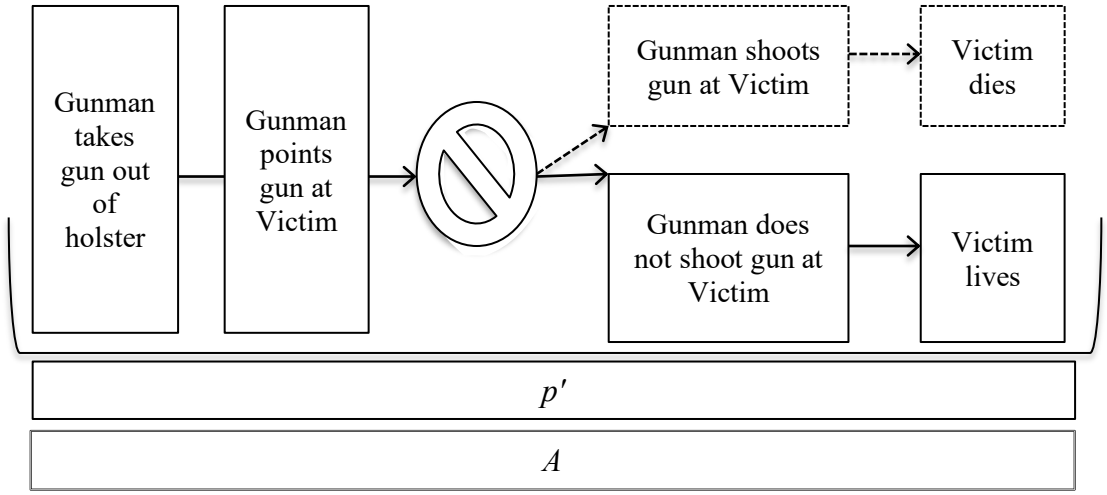
Moral Sequence without Intervention but with Personality Change (MSP)



This is an archetype of a moral sequence without intervention but with personality change. This sequence operates in the same way as MS in §6.5.1., but instead of the agent (Gunman, *A*) acting with one persistent personality throughout the sequence, Gunman acts with three personalities. The dashed lines group certain sequence-events and personalities together. During the series of sequence-events, certain sequence-events are completed by Gunman but with different personalities. In sequence-event 1, Gunman is *p′*; in sequence-events 2 and 3, Gunman is *p″*; and in sequence-outcome, Gunman is *p‴*.

6.5.3. (REVISED) MORAL SEQUENCE WITH INTERVENTION

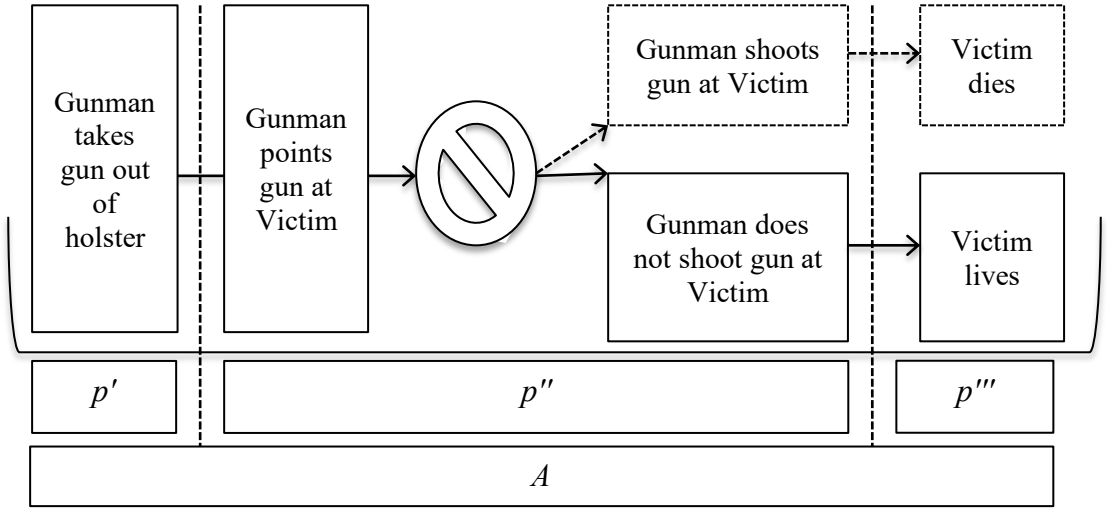
Moral Sequence with Intervention (MSI)



This is an archetype of a moral sequence with an intervention in which a distinct series of sequence-events occur and where there is an intervention. The sequence involves an intervention (represented by the "no entry sign") in which the series of Gunman's sequence-events do not continue past sequence-event 2. After the intervention, the solid arrows represent the continuation of actual sequence-events (sequence-event 3 is "Gunman does not shoot gun at Victim" and the sequence-outcome is "Victim lives"); the dotted arrows represent the projected continuation of the series of sequence-events if the intervention had not occurred ("Gunman shoots gun at Victim" is therefore projected sequence-event 3 and "Victim dies" is projected sequence-outcome). During the series of sequence-events, Gunman (the agent, represented by the upper-case *A*) initiated the sequence and acts throughout the sequence with a persistent personality (represented by *p′*).

6.5.4. MORAL SEQUENCE WITH INTERVENTION AND WITH PERSONALITY CHANGE

Moral Sequence with Intervention and with Personality Change (MSIP)



This is an archetype of a moral sequence with intervention and with personality change. This sequence operates in the same way as MSI in §6.5.3., but instead of the agent (Gunman, *A*) acting with one persistent personality throughout the sequence, Gunman acts with three personalities. The dashed lines group certain sequence-events and personalities together. During the series of sequence-events, certain sequence-events are completed by Gunman but with different personalities. In sequence-event 1, Gunman is *p′*; in sequence-events 2 and 3, Gunman is *p″*; but between sequence-events 2 and 3 an intervention occurs (represented by the "no entry sign"). After the intervention, the solid arrows represent the continuation of actual sequence-events (sequence-event 3 is "Gunman does not shoot gun at Victim" and the sequence-outcome is "Victim lives") during which Gunman is *p‴*; the dotted arrows represent the projected continuation of the series of sequence-events if the intervention had not occurred ("Gunman shoots gun at Victim" is therefore projected sequence-event 3 and "Victim dies" is projected sequence-outcome).

6.6. CONCLUDING REMARKS

This chapter has demonstrated how inner-agent change presented a problem for the system of moral sequencing proposed before this chapter. It made sense of the inner-agent changes that are paradigmatic of the initial problem case described in §6.1. and the empirical cases presented in §6.2. by showing how inner-agent change can be expressed as a change in personality; personality, I argued, is made-up of traits and personae, and these can be used to determine whether there has been a (strict or weak) personality change. The system of moral sequencing has thus been updated; to reflect this change, the two sequence archetypes presented in §2.2.1. and §2.2.2. were updated and two new archetypes were introduced to account for the addition of an agent's personality (or personalities) in moral sequencing.

The next chapter will discuss how, by using the personality approach to understand inner-agent changes as constituting changes in an agent's personality, normative considerations and ethical conclusions on inner-agent changes can be deliberated and drawn. Moreover, chapter 7 will discuss how the revised system of moral sequencing can, in some cases, permit a Deliberator to intervene earlier than was permitted in the old system of §2.2.1. and §2.2.2.

CHAPTER 7:

RESPONSIBILITY REVISITED AND INTERVENING EARLIER

> 'I concede that the murderer is a righty, like me, has the same fingerprints as I do, is clean-shaven like me. He even looks exactly like me in the surveillance camera photographs introduced by the defense. No, I have no twin. In fact, I admit that I remember committing the murder! But that murderer is not the same person as me, for I have changed […] Therefore you cannot punish me, for no one is guilty of a crime committed by someone else' (Sider, 2005: 7).

Chapter 6 introduced a problem for moral sequencing that questioned whether it is possible to attribute responsibility to an agent who has an inner-agent change. I argued that this inner-agent change can be explained with reference to a change or changes in an agent's personality. This chapter will continue this discussion of personality change.

The purpose of this chapter is twofold: first, to finish the discussion of responsibility started in chapter 5 to ascertain whether, through assessing if an agent undergoes inner-agent change, a Moral Assessor has at least some grounds for diminishing[171] the responsibility of that agent for their actions; and second, to ascertain whether an inner-agent change can

---

[171] Although diminished responsibility is, especially in UK law, predominantly reserved for assessing cases of murder—and whether this act of killing should be regarded as murder or manslaughter—I use this term more broadly, and in reference to whether an intra-sequence agent can have any responsibility attributed to them for their intra-sequence actions.

provide a justification for intervening earlier in a moral sequence. To achieve this, we must ensure that moral sequencing can properly account for inner-agent change.

Inner-agent change presents a problem for the system of moral sequencing that I have proposed so far since the system cannot account for apparent strict personality changes. In other words, in the current presentation of moral sequencing there are no inner-agent changes in agents (or more particularly, as I will shortly argue, moral sequencing only currently accounts for a persistent agent-personality relationship in all agents).

This chapter will argue that responsibility and the level thereof can be attributed to an agent depending on the persistence or lack thereof of what I will call *the agent-personality relationship*. To this end, the next section will explain this agent-personality relationship so that, in §7.2., we understand why a Moral Assessor would have at least some grounds for diminishing the responsibility attributed to intra-sequence agents for whom there has been an agent-personality relationship change.

7.1. THE AGENT-PERSONALITY RELATIONSHIP

The agent-personality relationship can be understood as the relationship an intra-sequence agent has to their personality. This relies on understanding that there is a difference between a physical *agent* and the *personality* that the agent has. To help understand the agent-personality relationship, we can view a moral sequence as containing the set "agent" which has as its members that agent's "personality" (or "personalities" depending on whether there

is more than one personality)[172]. The relevant period of time—which for a Deliberator includes only the moral sequence itself, but for a Moral Assessor includes both the duration of a moral sequence and any post-sequence time-frame relevant to normative considerations (e.g. when an agent is in police custody, on trial in court, and so on)—can be represented as t^{m}. During a moral sequence, intra-sequence agent A has personality p, where the superscript prime symbol accompanying p denotes a distinct personality (which is/are member(s) of A).

If the personality of an intra-sequence agent A that was present during their first intra-sequence action (e.g. for an Initiator this would be when they initiated the moral sequence) persists (during t^{m}) then we can say that there is a persistent agent-personality relationship. In other words, we can say that an agent-personality relationship persists when an intra-sequence agent does not undergo a personality change. Conversely, if the personality of an intra-sequence agent A changes from the personality (p') of A that was present during their first intra-sequence action, then we can say that the agent-personality relationship has changed (e.g. an Initiator's personality when they initiated the moral sequence (p') changing to another personality (p'')). In other words, we can say that an agent-personality relationship changes when an intra-sequence agent has a personality change.

Where A has only one personality during t^{m} the agent-personality relationship persists, and we can say that A during $t^{m} = \{p'\}$ or, put simply, $t^{m}\ \{p'\} \in A$. However, where A has more than one personality during t^{m} —where there has been an inner-agent change—the agent-personality relationship has changed, and we can say that $t^{m}\ \{p', \ldots, p^{n}\} \in A$. More

[172] In set theory, the symbol ∈ is used to denote that p^{n} is a member of A.

specifically, where A has two personalities during t^m we can say that A during $t^m = \{p', p''\}$ or simply that t^m $\{p', p''\} \in A$, where A has three personalities during t^m we can say that A during $t^m = \{p', p'', p'''\}$ or simply that t^m $\{p', p'', p'''\} \in A$, and so on.

A change in the agent-personality relationship therefore necessarily involves a distancing of that (physical) agent from the personality that agent has at the start of a moral sequence to another personality or personalities. To put this another way, an agent-personality relationship change requires an agent to have undergone at least one personality change.

To ensure that the concept of agent-personality relationship can be properly applied to moral sequencing, the next section will discuss how an agent-personality relationship change can be determined.

7.1.1. DETERMINING WHETHER THERE HAS BEEN AN AGENT-PERSONALITY RELATIONSHIP CHANGE

I now refer the reader back to §6.4.3. to elucidate and help understand when we can say that there has been a personality change (required to state whether there has or has not been an agent-personality relationship change). To briefly recap this discussion, I argued that one can make a warranted, but not categorical, assertion that there has been a strict personality change; even though an agent's traits are not directly epistemically accessible (and a change in traits is what is required to categorically assert that there has been a personality change), one does have indirect epistemic access to an agent's traits (c.f. §6.4.3.) and can thus say something meaningful about the trait changes required to determine whether an agent has undergone a personality change.

Determining whether there has been an agent-personality relationship change therefore requires a Moral Assessor to make a warranted assertion that one or more strict personality change(s) occurred in an agent. For reasons outlined in §6.4.3., requiring a Moral Assessor to categorically assert that there has been a strict personality change is too restrictive; indeed, as I have previously argued, categorically asserting that a strict personality change has occurred is deeply epistemically problematic since a Moral Assessor can only ever have indirect epistemic access to an agent's traits. But determining a personality change by reference to a weak personality change seems, in the case of a Moral Assessor, to be too epistemically lenient. This is because the task of a Moral Assessor is to draw normative ethical conclusions that potentially have serious pragmatic consequences. For example, in the case of determining whether the responsibility attributed to an agent should be diminished, establishing only a weak personality change—which would be based solely on an assessment of an agent's personae—does not help to establish with any level of philosophical certainty that a personality change has occurred; indeed, all that a Moral Assessor could be sure of is that an agent's persona has changed and this, as discussed in §6.4.2.3., is open to deception—an agent could act in a way that makes it seem as though they have had a personality change in an attempt to diminish their responsibility for their actions (this perhaps bringing with it a softer punishment). It is for this reason that I think we have good reason to require an agent-personality relationship change to be ascertained with reference to whether one can make a warranted assertion that an agent has undergone one or more strict personality changes.

Before moving on to discuss how a Moral Assessor can employ the concept of agent-personality relationship to attribute (or reduce the attribution of) responsibility, I will first clarify the parameters of t^m (the "relevant period of time").

7.1.2. CLARIFYING THE "RELEVANT PERIOD OF TIME"

I anticipate a potential objection to the way in which I have framed the agent-personality relationship, particularly with regards to how we decide what falls under t^m. I have stated that t^m denotes the "relevant period of time" which I framed as being dependent on whether the considerations pertain to a Deliberator or a Moral Assessor; for a Deliberator, I said that only the moral sequence itself constitutes the "relevant period of time", but for a Moral Assessor the "relevant period of time" includes both the duration of a moral sequence and any post-sequence time-frame relevant to normative considerations. Such relevant time-frames include (but are not limited to) when an agent is in police custody or on trial in a court of law. However, what constitutes a post-sequence time-frame relevant to normative considerations is arguably open-ended and open to philosophical debate. It is not clear, for instance, whether t^m should be limited to intra-sequence agent-personality relationship changes and only such changes occurring immediately following, or shortly after, the sequence. But such a stipulation seems arbitrary. If, say, a Moral Assessor would justify diminishing any attributed responsibility to an agent based on agent-personality relationship changes during or shortly after a moral sequence, then it would seem arbitrary to deny diminishing any attributed responsibility to an agent for such changes that occur months, years, or even decades later. It is for this reason that I am willing to bite the bullet and say that any such distant agent-personality relationship change requires the attention of a Moral Assessor in order to determine whether this change should constitute a retrospective diminishing of any responsibility attributed to that agent (with respect to a particular moral sequence). However, for current purposes, I ask the reader to consider the "relevant time period", t^{m}, to denote the moral sequence itself and any post-sequence time-frame that is *immediately relevant* to the normative consideration—relevant in the sense that it pertains

to whether an agent should be detained, put on trial, imprisoned as a result of their actions, and so on.

Now that we have a clear understanding of the nature of the agent-personality relationship, the next section will move on to a discussion on attributing responsibility to agents for whom an agent-personality relationship change has occurred. This will motive a further discussion, alluded to at the end of chapter 5, on whether a Moral Assessor should diminish the responsibility of agents who undergo an agent-personality change.

7.2. RESPONSIBILITY IN MORAL SEQUENCING

This section will ascertain how a Moral Assessor might determine an agent's responsibility for their actions in a moral sequence, and whether a Moral Assessor has at least some grounds for diminishing the responsibility they attribute to that agent who has undergone an agent-personality relationship change.

Before continuing, it is important to note that any attribution of responsibility to an agent, or indeed whether the responsibility should be diminished, is separate from and does not necessarily directly inform, nor can it necessarily be employed to proportion, the sort of post-sequence consequences to which an agent should be liable. In other words, the discussion of responsibility started in chapter 5 and in this chapter thus far does not necessarily have any bearing on, for instance, whether that agent should be punished for their actions (including, but not limited to, imprisonment).

7.2.1. ATTRIBUTING RESPONSIBILITY IN CASES OF A PERSISTENT AGENT-PERSONALITY RELATIONSHIP

Before moving on to a discussion of attributing responsibility (or diminishing responsibility) for agents who undergo an agent-personality relationship change, let us return to the discussion in chapter 5 which examined why a Moral Assessor would be justified in attributing at least some responsibility to certain intra-sequence agents—namely Initiators, Forbearers, Sustainers-2, and Snowballers—for the harm that did befall a Victim or would have befallen a Victim if an Intervener had not (successfully) intervened. There I argued that because these intra-sequence agents are necessarily involved in directly or indirectly facilitating the occurrence of harm to a Victim, a Moral Assessor would have at least some grounds for attributing responsibility to those agents. As a result of the discussion in chapter 6 and the introduction and clarification of agent-personality changes and the role such changes play in attributing responsibility (see §7.1.), we are now in a position to further justify the claims made in chapter 5 about the how such agents should be attributed at least some responsibility for their actions. This will be accomplished by reinforcing the distinction between a persistent and a changed agent-personality relationship, and how the attribution of responsibility (or lack thereof) can be mapped onto this distinction.

In short, the fact that Initiators, Forbearers, Sustainers-2, and Snowballers have a persistent agent-personality relationship—in addition to the fact that they are necessarily involved in directly or indirectly facilitating the occurrence of harm to a Victim—further grounds the claim that these agents are at least partially responsible for the (actual or expected) harm to a Victim.

In a moral sequence in which an agent *A* (here referring only to those agents who initiate, forbear, sustain-2, or snowball) acts, and where *A*'s agent-personality relationship persists, the reason that a Moral Assessor has at least some grounds for holding *A* responsible for the harm that befalls a Victim is *precisely because A*'s agent-personality relationship persisted. *A* acted in a way that demonstrated a persistence of their personality throughout the moral sequence. The same agent, with the same persisting personality, acted throughout the moral sequence.

The next section will pick-up where we left off and will explain why a Moral Assessor has at least some grounds for diminishing the responsibility attributed to agents in cases where they undergo an agent-personality relationship change.

7.2.2. ATTRIBUTING RESPONSIBILITY IN CASES OF AGENT-PERSONALITY RELATIONSHIP CHANGE

Let us return to the problem case introduced in §6.1. There, I provided the two cases of Car Thief and Comatose Thief to show how the relationship that Thief has to Casper in Car Thief is *prima facie* similar to the relationship pre-crash Thief has to post-crash Thief in Comatose Thief. A Moral Assessor would likely not attribute responsibility to Casper in Car Thief because Casper is a different agent to Thief and is not responsible for Thief's actions (nor, I ask the reader to assume, complicit or otherwise involved in the theft). For similar reasons that we would say that Casper is not responsible for Thief stealing the car, we want to say that a Moral Assessor should not hold post-crash Thief responsible for the actions of pre-crash Thief. After all, post-crash Thief would say that *he* did not steal the car—his

personality, behaviour, dispositions, idiosyncrasies, and even his memories have changed, all of which signify a change in his agent-personality relationship.

The question that now requires an answer is: does a Moral Assessor have at least some grounds on which to diminish the responsibility attributed to an agent who undergoes an agent-personality change? In other words, does a Moral Assessor have at least some grounds on which to diminish the responsibility attributed to post-crash Thief for stealing the car?

This question can be answered without reference to whether responsibility should be attributed to post-crash Thief because he has the same physical body (we might say the same physical agency) as pre-crash Thief. Such a consideration would rely on a Moral Assessor ascertaining whether an agent who undergoes a personality change (post-crash Thief) should be considered the same agent or a different agent to the agent before such changes occurred (pre-crash Thief). One might, for instance, argue that the fact that (the physical agent) Thief *did* steal the car justifies why post-crash Thief should be attributed at least some responsibility for stealing the car. Indeed, at face value and devoid of metaphysical considerations of whether an inner-agent change constitutes a change in agency, pre-crash Thief and post-crash Thief are still the same physical agent (this partially reflects the position of proponents of the Bodily Criterion of Personal Identity[173]). But equally one might say that, because post-crash Thief cannot remember stealing the car (nor, we assume,

---

[173] Very briefly, proponents of the bodily criterion claim that personal identity is constituted by sameness of body. In other words, the bodily criterion states that *P*1 at *t*1 is *P*2 at *t*2 if and only if *P*1 and *P*2 have one-and-the-same body. Importantly, on this view the persistence of identity over time does not rely on persistence of physical matter but rather on persistence of form, or more specifically by the gradual replacement of *P*1's constituent parts in a way such that the body of *P*1 at *t*1 is identical to the body of *P*2 at *t*2.

the lead-up to this), post-crash Thief should not be attributed responsibility for stealing the car (this partially reflects the position of proponents of the Memory Criterion of Personal Identity[174]). The issue is that any such determination concerning the attribution of responsibility (or lack thereof) to post-crash Thief for pre-crash Thief's actions can only be decided with reference to a particular metaphysical account of personal identity. However, as discussed in §6.3., such considerations can be by-passed; we can make normative conclusions about the attribution of responsibility, including those related to diminishing any such responsibility, without reference to such accounts and without engaging in a discussion of which metaphysical account should be adopted.

All that needs to be recognised is that post-crash Thief's personality change described in Comatose Thief provides a Moral Assessor with at least some ground for diminishing the responsibility attributed to his actions. The reason that a Moral Assessor would have at least some grounds for saying that post-crash Thief's responsibility for his actions should be reduced (in cases where an adopted metaphysical position attributes responsibility to Thief) or that no responsibility should be attributed (in cases where an adopted metaphysical position does not attribute responsibility to Thief) is because the relationship that post-crash Thief has to pre-crash Thief changed in so far as Thief underwent an agent-personality relationship change.

---

[174] Very briefly, the memory criterion can be loosely formulated as: *P*1 at *t*1 is one-and-the-same as *P*2 at *t*2 if and only if *P*1 at *t*1 and *P*2 at *t*2 are connected by a continuity of memories. Under this definition, I am the same person I was ten years ago, even though I can't remember what I was doing or did ten years ago, because, so long as my stream of memories overlap and are thus connected, I can be said to be the same person.

The personality approach has therefore provided a way to understand the role of personality in moral sequencing in a way that can justify why a Moral Assessor could at least reduce the responsibility attributed to an agent who has undergone an agent-personality relationship change.

7.3. UPDATING THE SECONDARY NARRATIVE TO INTERVENE EARLIER

In §7.2. I argued that when a Moral Assessor seeks to determine whether any responsibility should be attributed to an agent for their actions in a moral sequence, his decision should be based on whether there has been an agent-personality relationship change, which itself relies on being able to make a warranted assertion that an agent has undergone a strict personality change. By employing this (via the personality approach outlined in chapter 6), moral sequencing can be used to explain why a Moral Assessor has at least some grounds for diminishing the responsibility attributed to an agent due to the fact that there is a disconnect between an agent before and after an agent-personality relationship change. This discussion provided a way to answer the problem case and empirical cases outlined in §6.1. and §6.2. respectively. However, there is another upshot of accounting for personality changes in moral sequencing.

The personality approach, properly embedded into moral sequencing, allows a Deliberator to potentially intervene in a moral sequence earlier than would otherwise have been permitted under the system of moral sequencing presented prior to chapter 6. By accounting for personality changes in moral sequencing, or rather accounting for those moral sequences in which an agent undergoes a personality change, a Deliberator has access to some potentially vital information that can be used in his decision-making process (concerning

deciding if and when to intervene to prevent the occurrence of harm to a Victim); a Deliberator can use information pertaining to changes in an agent's personality and related issues to *update the secondary narrative* (see §4.2.1.). In other words, a Deliberator can use information related to an agent's personality—including whether that agent has undergone personality changes in the past, which may be indicative of a related psychiatric issue or which may reflect a psychiatric diagnosis such as Dissociative Identity Disorder (DID), or simply that the agent may be prone to unannounced changes in their behaviour—to justify intervening in a moral sequence earlier than would otherwise (without such information) have been warranted. A Deliberator may be justified in intervening earlier in moral sequences where such a secondary narrative is available due to the fact that this information may alter the position of the threshold of physical harm (c.f. §4.3.); such a secondary narrative (in conjunction with other narratives) may bring forward the threshold of physical harm (if such information increases the probability that harm will occur to a Victim, for example), thus justifying an earlier intervention than would have been warranted without such information.

Importantly, in determining if and when to intervene in a moral sequence, a Deliberator should not be subjected to the same standard of epistemic rigour that a Moral Assessor is required to abide by when determining whether or not to diminish any responsibility attributed to an agent. Whilst a Moral Assessor can only diminish responsibility attributed to an agent in cases where the Moral Assessor can make a warranted assertion that there has been a strict personality change, a Deliberator should not necessarily be required to do the same. After all, a Deliberator is not in the business of deciding whether an agent's agent-personality relationship has changed; a Deliberator is only concerned with deciding if and when to intervene in a moral sequence (in order to prevent harm befalling a Victim)—a

Deliberator's task is purely pragmatic, whereas a Moral Assessor's task has an added epistemic element. A Deliberator's task can therefore be accomplished by simply assessing weak personality change(s)—that is, updating the secondary narrative with information pertaining to an agent's personality based on that agent's persona, *viz.* based on that agent's behaviour.

The reason that a lower epistemic standard is permitted of a Deliberator is because the pragmatic costs of a Deliberator mistakenly attributing a personality change to an agent that has not undergone such a change are arguably low. Although mistakenly updating the secondary narrative with information relating to an agent's personality (where such information is false) could lead a Deliberator to intervening *too early* in a moral sequence (by mistakenly identifying the threshold of physical harm), this arguably has few pragmatic consequences (apart from the obvious fact that the Deliberator has potentially intervened in a situation that did not require an intervention).

7.4. CONCLUDING REMARKS

This chapter has accounted for the issue presented in chapter 6, namely that the system of moral sequencing (as presented in the chapters preceding chapter 6) was not able to account for the sorts of inner-agent change outlined in the problem case of Car Thief / Comatose Thief (§6.1.) and the practical cases of Billy Milligan, Juanita Maxwell, Mark Peterson, Kenneth Bianchi, and Mr A (§6.2.).

To overcome this problem, the concepts of agent-personality relationship change and persistence was introduced (§7.1.). Doing so equipped us with the means to, firstly, explain

why a Moral Assessor would have at least some grounds for diminishing any responsibility attributed to an agent for their intra-sequence actions (§7.2.) and, secondly, why a Deliberator could use information related to an agent's personality (and changes thereof) to update the secondary narrative and thus justify intervening earlier in a moral sequence than would otherwise have been permitted (§7.3.).

CONCLUSION

The motivation for this thesis was twofold: to provide a novel system of moral sequencing that can be applied to general moral problems to decide if, when, and in what way(s) an agent can intervene to prevent harm from occurring to another agent; and, off the back of this, discuss the responsibility that can be attributed to certain intra-sequence agents for their actions. Both of these motivations have been at the centre of the discussions in this thesis and these aims have been fulfilled by providing a novel system of moral sequencing. The upshot of moral sequencing is that it can also make sense of practical cases of inner-agent changes.

Chapter 1 began by presenting an overview of the most pertinent theoretical literature that underpins the system of moral sequencing I introduced in chapter 2. The concepts of agency, sequencing, initiating, sustaining, enabling, forbearing to prevent, intervention, interposition, barriers, responsibility within a sequence, and negative rights were all introduced and discussed—and it is the introduction of this terminology and how they form part of the larger literature on moral sequencing (discussed in terms of the distinction between doing harm and allowing harm) that provided the foundations for constructing a novel system of moral sequencing that can be applied to general moral issues.

The moral sequencing presented in chapter 2 added the "missing ingredients" required to extend the success of the literature on the doing/allowing distinction—most notably building on the successes of Fiona Woollard's work in which the precursors to moral sequencing can best be seen—to encompass a fuller understanding of moral sequences that can be employed outside of the doing/allowing distinction. This new system of moral sequencing was

established in a way that enables it to be employed to assess if, when, and in what way(s) a Deliberator can justifiably intervene to prevent the occurrence of harm to a Victim and to ascertain whether responsibility can be attributed to intra-sequence agents for the harm that befalls a Victim or would have befallen a Victim without an intervention. To help explicate this new system of moral sequencing, two archetypes of moral sequencing were presented in §2.2.1. and §2.2.2. (with their parameters set in §2.1.3.): a moral sequence without intervention and a moral sequence with intervention respectively.

Chapter 3 started by honing-in on the most philosophically (and ethically) relevant type of moral decision (§3.1.), it then discussed (§3.2.) decision theory (in particular, Bayesian Decision Theory) to oust solely formal systems of decision-making as impractical for moral decision-making (normatively speaking) (§3.2.1.), and then continued to argue that reactive-calculated moral decisions are (normatively) the best method for moral decision-making, and, by focussing on a specific type of moral decision (namely deciding to intervene), it can be used to justify a decision to intervene in a moral sequence by appealing to the role of emotions in decision-making (§3.2.2.). Chapter 3 culminated in the claim that the most practical way of making moral decisions is to employ Damasio's somatic-marker hypothesis in the broader framework of the formal decision-making process to reduce this process to a basic, almost intuitive, calculation of the probability of harm occurring, but in a way that avoids the impracticalities of formal decision-making processes (namely that they are too complex and time-consuming to be considered a realistic moral decision-making process and they require an element of reflection that is impractical in many moral situations) by appealing to and cashing-in on the "gut feelings" experienced in moral situations that denote the presence of a somatic marker.

Chapter 4 built on my claim that reactive-calculated moral decisions are normatively favourable above the other four kinds of moral decisions and applied this to a particular moral decision, namely deciding if and when to intervene to prevent harm, by arguing for a basic probability-driven formal decision-making process that is underpinned by a reactive element linked, first and foremost, to the progression of the primary narrative of the sequence (*viz.* how the sequence-events unfold) and, if available, an assessment of the secondary narrative of the same sequence (*viz.* one's prior knowledge of the beliefs, dispositions, etc. of the intra-sequence agent that the Deliberator is evaluating) (§4.2. to §4.5.) to establish the probability of harm eventuating to a Victim. These primary and secondary narratives help a Deliberator to ascertain the risk of harm, but I argued that to ensure that an intervention is philosophically justifiable a Deliberator needs to consider the tertiary narrative (*viz.* how to intervene), namely: whether an intervention is necessary (§4.6.4.1.), whether the intervention is proportionate to the threatened harm to a Victim (§4.6.4.2.), and whether any agent harmed by an intervention is liable to that harm (§4.6.4.3.). I argued that intervention is only justifiable when a moral sequence passes the threshold of physical harm, itself requiring that a Deliberator satisfies a number of conditions within the primary, secondary, and tertiary narratives. This added to the picture of moral sequencing introduced in chapter 2 by ultimately establishing that the moral decision-making process outlined and discussed in chapter 3 is a reliable method for ensuring interventions are justified.

Chapter 5 completed the picture of moral sequencing presented thus far by turning attention to posing and answering questions related to determining the responsibility of intra-sequence agents for the harm that occurred to a Victim or would have occurred without an intervention. Section 5.1. considered some salient philosophical literature on responsibility

to ground the discussion and act as a benchmark against which moral sequencing's employment of 'responsibility' as a term of art could be defined and best understood. In §5.2. each intra-sequence agent was considered—namely Initiators, Forbearers, Sustainers-2, Interveners, and Snowballers—to ascertain whether any responsibility could be attributed to them for their actions and to outline that and why a Moral Assessor might wish to account for various agent-specific mitigating factors that might reduce that agent's responsibility. I argued that these considerations can be made *in vacuo* from any theoretical accounts of responsibility.

Chapter 6 outlined a potential problem for the system of moral sequencing that I had presented, namely that it couldn't *prima facie* account for a motivation to diminish the responsibility attributed to agents that underwent an inner-agent change. The problem (as presented in §6.1. and reiterated with practical examples in §6.2.) was that, for the same reason that we would say that Casper was not responsible for Thief stealing the car, we want to say that a Moral Assessor should not hold post-crash Thief responsible for the actions of pre-crash Thief. In §6.3., I argued that we can be agnostic on personal identity; we do not need to rely on an account of personal identity to have ethical discussions about personal identity. I argued that normative considerations and ethical conclusions can be made based on an assessment of personality by taking a personality approach to normative considerations. Section 6.4. then presented this personality approach as a way of understanding inner-agent change with reference to changes in an agent's personality. This led to a discussion in §6.5. about how traits and personae are the two components of personality. The epistemic and pragmatic costs and benefits of conceptualising traits and personae as presented in the personality approach was discussed in §6.4.2.3. in order to support the claim of §6.4.2.4. that only by viewing personality as being constituted by both

traits and personae are we in a position to say anything philosophically, epistemically, and pragmatically meaningful about, and establish normative conclusions on, inner-agent change. This discussion was concluded in §6.4.3. by arguing that both Deliberators and Moral Assessors have two methods for assessing strict and weak personality changes, related to trait changes and persona changes respectively. I argued that: it is possible to have direct epistemic access to personae and personae changes and so it is possible to make claims about weak personality changes; that it is possible to have indirect epistemic access to traits and trait changes and so it is possible to make a warranted assertion that there has been a strict personality change; but that is it not possible to categorically assert that there has been a strict personality change. Section 6.5. finished the chapter by revising the two sequence archetypes introduced in §2.2.1. and §2.2.2. and added two new archetypes to account for and properly embed personality and changes in an agent's personality into moral sequencing. Sections 6.5.1. to 6.5.4. presented a moral sequence without intervention, a moral sequence without intervention but with personality change, a moral sequence with intervention, and a moral sequence with intervention and with personality change respectively.

Chapter 7 then turned attention back to the issue of attributing responsibility to show how the *prima facie* problem (of moral sequencing not being able to account for inner-agent change) outlined in chapter 6 could be accounted for by revising the system of moral sequencing and introducing the concepts of agent-personality relationship change and persistence (§7.1.). I argued that an agent-personality relationship change can only be ascertained with reference to whether one can make a warranted assertion that an agent has undergone one or more strict personality changes (§7.1.1.). The rest of this chapter was then divided into two parts. The first part (§7.2.) revisited the issue of attributing responsibility

to an agent for their actions in a moral sequence. Section 7.2.1. argued that the fact that Initiators, Forbearers, Sustainers-2, and Snowballers have a persistent agent-personality relationship further grounds the claims made in chapter 5 that at least some level of responsibility should be attributed to these agents for their intra-sequence actions. Section 7.2.2. argued that a Moral Assessor should at least reduce the responsibility attributed to an agent who has undergone an agent-personality relationship change due to the fact that there is a disconnect between an agent before and after an agent-personality relationship change. The second part (§7.3.) demonstrated how the personality approach (now embedded into the system of moral sequencing) allows a Deliberator to potentially intervene in a moral sequence earlier than would otherwise have been permitted under the system of moral sequencing presented prior to chapter 6; by accounting for personality changes in moral sequencing, a Deliberator can use information pertaining to changes in an agent's personality and related issues to update the secondary narrative.

REFERENCES

Aboodi, R., Borer, A. and Enoch, D. (2008). 'Deontology, individualism, and uncertainty'. *Journal of Philosophy, 105*(5): 259–272.

Aksoy, S. (2014). *Bayesian Decision Theory*. Lecture notes for CS 551. Bilkent University, Department of Computer Engineering. Available at: http://www.cs.bilkent.edu.tr/~saksoy/courses/cs551/slides/cs551_bayesian.pdf [Accessed 02 April 2014].

Alexander, L. (2016). 'Recipe for a Theory of Self-Defense: The Ingredients, and Some Cooking Suggestions'. *In*: C. Loons and M. Weber (eds.), *The Ethics of Self-Defense*. New York: Oxford University Press, pp. 20–50.

Alexander, L. and Ferzan, K.K. (2009). 'Culpable Acts of Risk Creation'. *Ohio State Journal of Criminal Law*, *5*(2): 375–405.

Allhoff, Fritz (2019). 'Self-Defense Without Imminence'. *American Criminal Law Review*, *56*: 1527–1552.

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.

Appiah, A. (2007). *The Ethics of Identity*. Princeton, N.J., Princeton University Press.

Baron, Marcia (2011). 'Self-Defense: The Imminence Requirement'. *In*: Leslie Green and Brian Leiter (eds.), *Oxford Studies in Philosophy of Law: Volume 1*. Oxford: Oxford University Press.

Bayes, T. and Price, R. (1763). 'An Essay Towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S.'. *Philosophical Transactions (1683–1775), 53*: 370–418.

Beauchamp, T. L. (1979). A reply to Rachels on active and passive euthanasia. *In:*W. L. Robinson and N. J. Clifton (eds.), *Medical Responsibility.* New York: Humana Press, pp. 182–195.

Ben-Haim, Yakov (2006). Info-Gap Decision Theory: Decisions Under Severe Uncertainty. Oxford: Academic Press.

Benbaji, Yitzhak (2007). 'The Responsibility of Soliders and the Ethics of Killing in War'. *Philosophical Quarterly*, *57*(229): 558–572.

Bennett, Jonathan (1966). Whatever the Consequences. *Analysis*, *26*(3): 83–102.

Bennett, Jonathan (1967). Acting and Refraining. *Analysis, 28*(1): 30–31.

Bennett, Jonathan (1980). 'Accountability'. *In*: Z. van Stratten (ed.), *Philosophical Subjects: Essays Presented to P.F. Strawson*. New York: Oxford University Press.

Bennett, Jonathan (1981). 'Morality and Consequences'. *In:* Sterling McMurrin (ed.), *The Tanner Lectures on Human Values.* Salt Lake City: University of Utah Press.

Bennett, Jonathan (1993). 'Negation and abstention: Two theories of allowing'. *Ethics, 104*(1): 75–96.

Bennett, Jonathan (1995). *The Act Itself*. Oxford: Oxford University Press.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Second edition. New York: Springer.

Bernardo, J. M. and Smith, A. F. M. (2000). *Bayesian Theory.* Chichester: Wiley-Blackwell.

Bortolotti, Lisa (2015). *Irrationality*. Cambridge: Polity Press.

Bradley, R. (2007). 'A unified Bayesian decision theory'. *Theory and Decision, 63*(3): 233–263.

Bradley, Richard (2017). *Decision Theory with a Human Face*. Cambridge: Cambridge University Press.

Brown, K.F., Rumgay, H., Dunlop, C., Ryan, M., Quartly, F., Cox, A., Deas, A., Elliss-Brookes, L., Gavin, A., Hounsome, L., Huws, D., Ormiston-Smith, N., Shelton, J., White, C., and Parkin, D.M. (2018). 'The fraction of cancer attributable to known risk factors in England, Wales, Scotland, Northern Ireland, and the UK overall in 2015'. *British Journal of Cancer*, *118*: 1130–1141.

Burnside, Jonathan (2001). 'Bash the Burglar? Reflections on the Tony Martin case'. *Law & Justice*, *147*: 122–131.

Butterfield, Kenneth D., Klebe Trevin, Linda and Weaver, Gary R. (2000). 'Moral Awareness in Business Organizations: Influences of Issue-Related and Social Context Factors'. *Human Relations*, *53*: 981–1018.

Cancer Research UK (2018). 'Tobacco statistics'. Available at: https://www.cancerresearchuk.org/health-professional/cancer-statistics/risk/tobacco [Accessed 15 June 2019].

Card, O. S. (1985). *Ender's game*. New York: A Tom Doherty Associates, Inc.

Cattell, R. B., Eber, H. W. and Tatsuoka, M. M. (1970). 'Handbook for the sixteen personality factor questionnaire (16 PF)'. *In*: clinical, educational, industrial, and research psychology, for use with all forms of the test. *Institute for Personality and Ability Testing*.

Child, I. L. (1968). 'Personality in culture'. *In:* E. F. Borgatta and W. W. Lambert (eds.), *Handbook of personality theory and research.* Chicago: Rand McNally.

Clarkeburn, Henriikka (2002). 'A Test for Ethical Sensitivity in Science'. *Journal of Moral Education*, *31*(4): 439–453.

Coelho, Helder, da Rocha Costa, António Carlos and Trigo, Paulo (2014). 'On Agent Interactions Governed by Morality'. *Interdisciplinary Applications of Agent-Based Social Simulation and Modeling,* IGI Global, 20–35.

Conger, John P. (2005). *Jung and Reich: The Body as Shadow*. California: North Atlantic Books.

Cook, K. (2014). Kitty Genovese: The Murder, the Bystanders, the Crime That Changed America. New York: W. W. Norton & Company.

Cushman, F., Young, L. and Hauser, M. (2006). 'The Role of Conscious Reasoning and Intuition in Moral Judgment: Testing Three Principles of Harm'. *Psychological Science*, *17*(12): 1082–1089.

Damasio, A. R. (2006). *Descartes' Error. Emotion, Reason and the Human Brain.* Revised edition. London: Vintage.

Darwall, Stephen (2006). *The Second-Person Standpoint: Morality, Respect, and Accountability*. Cambridge, MA: Harvard University Press.

de Finetti, B. (1937). 'Foresight: Its Logical Laws, Its Subjective Sources'. *In:* H. E. Kyburg and H. E. K. Smokler (eds.), *Studies in Subjective Probability.* Huntington, NY: Robert E. Kreiger Publishing Co.

Digman, J. M. (1990). Personality Structure: Emergence of the Five-Factor Model. *Annual Review of Psychology*, *41*: 417–440.

Dinello, Daniel (1971). 'On Killing and Letting Die'. *Analysis*, *31*(3): 83–86.

Donagan, Alan (1977). *The Theory of Morality*. Chicago: University of Chicago Press.

Dougherty, T. (2013). 'Vague value'. *Philosophy and Phenomenological Research, 89*(2): 352–372.

Draper, Kai (2005). 'Rights and the Doctrine of Doing and Allowing'. *Philosophy and Public Affairs, 33*(3): 253–280.

Draper, Kai (2009). 'Defense'. *Philosophical Studies*, *145*(1): 69–88.

Dunn, Barnaby D., Dalgeish, Tim and Lawrence, Andrew D. (2006). 'The somatic marker hypothesis: A critical evaluation'. *Neuroscience and Biobehavioral Reviews*, *30*: 239–271.

E7 v Holland (2014). EWHC 452.

Einstein, Albert and Infeld, Leopold (1938). The Evolution of Physics: The growth of ideas from early concepts to relativity and quanta. New York: Simon and Schuster, Inc.

Ekstrom, Laura Waddell (2000). *Free Will: A Philosophical Study*. Boulder, CO: Westview Press.

Eshleman, Andrew (2014). 'Moral Responsibility'. *Stanford Encyclopedia of Philosophy*. Available at: https://plato.stanford.edu/entries/moral-responsibility [Accessed 23 March 2019].

Everitt, B. (2002). *The Cambridge Dictionary of Statistics*. Cambridge: Cambridge.

Eysenck, H. J. (1991). 'Dimensions of personality: 16, 5, or 3? Criteria for a taxonomic paradigm'. *Personality and Individual Differences*, *12*: 773–790.

Eysenck, H. J. and Eysenck, M. W. (1985). *Personality and individual differences: a natural science approach*. New York: Plenum Press.

Eysenck, Michael W. (1994). *Perspectives on Psychology*. Hove: Psychology Press Ltd.

Fabre, Cécile (2009). 'Permissible Rescue Killings'. *Proceedings of the Aristotelian Society*, *109*: 149–164.

Fabre, Cécile (2012). *Cosmopolitan War*. Oxford: Oxford University Press.

Fabre, Cécile (2014). 'Cosmopolitanism and Wars of Self-Defence'. *In*: Cécile Fabre and Seth Lazar (eds.), *The Morality of Defensive War*. Oxford: Oxford University Press, pp. 90–115.

Fabre, Cécile (2016). *Cosmopolitan Pease*. Oxford: Oxford University Press.

Feinberg, Joel (1970). *Doing and Deserving: Essays in the Theory of Responsibility*. Princeton: Princeton University Press.

Feller, W. (1971). An introduction to probability theory and its applications. New York: John Wiley.

Ferzan, Kimberly Kessler (2004). 'Defending Imminence: From Battered Women to Iraq'. *Arizona Law Review*, *46*: 213–262.

Ferzan, Kimberly Kessler (2005). 'Justifying self-defense'. *Law and Philosophy, 24*(6): 711–749.

Ferzan, Kimberly Kessler (2012). 'Culpable Aggression: The Basis for Moral Liability to Defensive Killing'. *Ohio State Journal of Criminal Law*, *9*(2): 669–697.

Ferzan, Kimberly Kessler (2016). 'Self-Defense: Tell Me Moore'. *In*: K. K. Ferzan and S. J. Moore (eds.), *Legal, Moral, and Metaphysical Truths: The Philosophy of Michael S. Moore*. Oxford: Oxford University Press, pp. 219–232.

Finkel, Norman J. (1988). *Insanity on Trial*. New York: Plenum Press.

Firth, Joanna Mary and Quong, Jonathan (2012). 'Necessity, Moral Liability, and Defensive Harm'. *Law and Philosophy*, *31*: 673–701.

Fischer, John Martin and Ravizza, Mark (1993). *Perspectives on Moral Responsibility*. Ithaca: Cornell University Press.

Fischer, John Martin and Ravizza, Mark (1998). *Responsibility and Control: A Theory of Moral Responsibility*. New York: Cambridge University Press.

Fischer, John Martin and Ravizza, Mark (eds.) (1993). *Perspectives on Moral Responsibility*. Ithaca: Cornell University Press.

Fitzgerald, P. (1967). 'Acting and Refraining'. *Analysis, 27*(4): 133–139.

Fletcher, George (1973). 'Proportionality and the psychotic aggressor'. *Israel Law Review*, *8*: 367–90.

Flew, A. (1979). *A Dictionary of Philosophy*. Indiana: Pan Books.

Foot, Philippa (1984). 'Killing and Letting Die'. *In*: Jay L. Garfield and Patricia Hennessey (eds.), *Abortion and Legal Perspectives*. Amherst: University of Massachusetts Press.

Foot, Philippa (1985). 'Utilitarianism and the virtues'. *Mind, 94*(374): 196–209.

Foot, Philippa (1994). ‘The problem of abortion and the doctrine of double effect’. *In*: B. Steinbock and A. Norcross (eds.), *Killing and Letting Die*. New York: Fordham University Press, pp. 266–279.

Fraser, Andrew M. (2009). *Hidden Markov Models and Dynamical Systems*. Philadelphia: Society for Industrial and Applied Mathematics.

Frowe, H. (2010). ‘A practical account of self-defence’. *Law and Philosophy, 29*: 245–272.

Frowe, Helen (2014). *Defensive killing*. Oxford: Oxford University Press

Frowe, Helen (2016). *The Ethics of War and Peace: An introduction*. Second edition. New York: Routledge.

Glover, Jonathan (1970). *Responsibility*. New York: Humanities Press.

Goldberg, L. R. (1981). ‘Language and individual differences: The search for universals in personality lexicons’. *Review of personality and social psychology*, *2*(1): 141–165.

Goldberg, L. R. (1993). ‘The structure of phenotypic personality traits’. *American Psychologist*, *48*(1): 26.

Haidt, J. (2001). ‘The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment’. *Psychological Review, 108*(4): 814–834.

Hansson, S. O. (1994). *Decision Theory: A Brief Introduction*. Revised edition. Stockholm: Royal Institute of Technology. Available at: http://home.abe.kth.se/~soh/decisiontheory.pdf [Accessed 6 January 2015].

Haque, A. A. (2012). ‘Killing in the fog of war’. *Southern California Law Review, 86*(1): 63–116.

Haque, A.H. (2009). ‘Rights and Liabilities at War’. *In*: P. H. Robinson, S. P. Garvey, and K. K. Ferzan (eds.), *Criminal Law Conversations*. New York: Oxford University Press, pp. 395–396.

Harris, John (1975). ‘The Survival Lottery’. *Philosophy. 50*(191): 81–87.

Hawley, P. (2008). ‘Moral absolutism defended’. *Journal of Philosophy, 105*(5): 273–275.

Hirschberg, N. (1978). ‘A correct treatment of traits’. *In*: H. L'ndo (ed.), *Personality: A new look at metatheories*. New York: Macmillan.

Holtzman, Nicholas S. and Strube, Michael J. (2012). ‘People With Dark Personalities Tend to Create a Physically Attractive Veneer’. *Social Psychological and Personality Science*, *4*(4): 461–467.

Horder, Jeremy (2002). ‘Killing the Passive Abuser: A Theoretical Defence’. *In*: Stephen Shute and Andrew Simester (eds.), *Criminal Law Theory: Doctrines of the General Part*. Oxford: Oxford University Press.

Howard-Snyder, F., Howard-Snyder, D. and Wasserman, R. (2009). *The Power of Logic.* Fourth edition. New York: McGraw-Hill.

Howard-Snyder, Frances (2011). ‘Doing vs. Allowing Harm’. *Stanford Encyclopedia of Philosophy*. Available at: https://plato.stanford.edu/entries/doing-allowing/ [Accessed 10 May 2014].

Howard, R. A. (1988). ‘Decision Analysis: Practice and Promise’. *Management Science, 34*(6): 679–695.

Huemer, M. (2010). ‘Lexical priority and the problem of risk’. *Pacific Philosophical Quarterly, 91*(3): 332–351.

Hunt, S. D. and Vitell, S. (1986). ‘A general theory of marketing ethics’. *Journal of Macromarketing*, *6*(1): 5–16.

Hurka, T. (2007). ‘Liability and Just Cause’. *Ethics & International Affairs*, *21*(2): 199–218.

I-Robot (2004). [Film] USA: Alex Proyas.

Imrie, Robert (1990). ‘Multiple personality disorder clouds rape case’. *The Times*. 17 August 1990. Available at: https://news.google.com/newspapers?nid=1665&dat=19900817&id=kVYaAAAAIBAJ&sjid=tSQEAAAAIBAJ&pg=5300,4708639&hl=en [Accessed 12 April 2017].

Isen, A. M. and Patrick, R. (1983). ‘The effect of positive feelings on risk taking: When the chips are down’. *Organizational Behavior and Human Performance, 31*(2): 194–202.

Jackson, F. and Smith, M. (2006). ‘Absolutist moral theories and uncertainty’. *Journal of Philosophy, 103*(6): 267–283.

Jackson, F. and Smith, M. (2015). ‘The implementation problem for deontology’. *In*: B. Maguire and E. Lord (eds.), *Weighing reasons*. Oxford: Oxford University Press, pp. 279–291.

Jones, T. M. (1991). ‘Ethical decision making by individuals in organizations: An issue-contingent model’. *Academy of Management Review*, *16*: 366–395.

Kaelbling, Leslie Pack, Littman, Michael L. and Cassandra, Anthony R. (1998). ‘Planning and Acting in Partially Observable Stochastic Domains’. *Artificial Intelligence*, *101*: 99–134.

Kagan, Shelly (1988). The additive fallacy. *Ethics, 99*(1): 5–31.

Kagan, Shelly (1989). *The Limits of Morality*. Oxford: Oxford University Press.

Kamm, Frances Myrna (1983). ‘Killing and Letting Die: Methodological and Substantive Issues’. *Pacific Philosophical Quarterly, 64*(4): 297–312.

Kamm, Frances Myrna (1986). ‘Harming, not aiding, and positive rights’. *Philosophy and Public Affairs*, *15*(1): 3–32.

Kamm, Frances Myrna (2007). *Intricate Ethics*. New York: Oxford University Press.

Kassin, S. M. (2003). *Psychology*. Upper Saddle River, NJ: Pearson/Prentice Hall.

Kaufman, W. R. P. (2009). *Justified Killing: The Paradox of Self-Defence*. Plymouth: Lexington Books.

Keyes, Daniel (1995). *The Minds of Billy Milligan*. New York: Bantam.

Klein, Colin (2017). 'Precaution, proportionality and proper commitments: Commentary on Birch on Precautionary Principle'. Available at: https://animalstudiesrepository.org/cgi/viewcontent.cgi?article=1232&context=animsent [Accessed 22 May 2019].

Klir, George J. (1995). *Fuzzy Sets and Fuzzy Logic: Theory and Applications*. New Jersey: Prentice Hall.

Kramer, Matthew H. (2014). *Torture and Moral Integrity: A Philosophical Enquiry*. Oxford: Oxford University Press.

Lave, J. (1988). Cognition in Practice: Mind, Mathematics and Culture in Everyday Life. Cambridge: Cambridge University Press.

Lave, J. and Wenger, E. (1991). *Situated Learning: Legitimate Peripheral Participation.* Cambridge: Cambridge University Press.

Lave, J., Murtaugh, M. and de la Rocha, O. (1984). 'The Dialectic of Arithmetic in Grocery Shopping'. *In:* B. Rogoff and J. Lave (eds.), *Everyday Cognition: Its Development in Social Context* Cambridge. Masssachusetts, MA: Harvard University Press, pp. 67–94.

Lazar, Seth (2012). 'Necessity in Self-Defence and War'. *Philosophy & Public Affairs*, *40*(1): 3–44.

Lazar, Seth (2013). 'Associative Duties and the Ethics of Killing in War'. *Journal of Practical Ethics*, *1*(1): 3–48.

Lazar, Seth (2016). 'War'. *Stanford Encyclopedia of Philosophy*. Available at: https://plato.stanford.edu/entries/war/ [Accessed on 11 January 2019].

Lazar, Seth (2018). 'In dubious battle: uncertainty and the ethics of killing'. *Philosophical Studies*, *175*(4): 859–883.

Lefkowitz, D. (2009). ‘Partiality and Weighing Harm to Non-Combatants’. *Journal of Moral Philosophy*, *6*(3): 298–316.

Lemmings, David and Brooks, Ann (eds.) (2014). *Emotions and Social Change: Historical and Sociological Perspectives*. New York: Routledge.

Leverick, F. (2006). Killing in Self-Defence. Oxford Monographs on Criminal Law and Justice. Oxford: Oxford University Press.

Lichtenberg, Judith (1982). ‘The Moral Equivalence of Action and Omission’. *Canadian Journal of Philosophy, 12*(sup1): 19–36.

Littman, Michael L. (2009). ‘A tutorial on partially observable Markov decision processes’. *Journal of Mathematical Psychology*, *53*(3): 119–125.

Locke, Don (1982). ‘The Choice Between Lives’. *Philosophy, 57*(222): 453–475.

Luce, R. D. and Raiffa, H. (1957). *Games and Decisions.* New York: Wiley.

Magill, Kevin (1997). Freedom and Experience: Self-Determination without Illusions. New York: St. Martin’s Press.

Magill, Kevin (2000). ‘Blaming, Understanding, and Justification’. *In*: T. van den Beld, *Moral Responsibility and Ontology*. Dordrecht: Kluwer.

Malm, Heidi (1992). ‘In Defense of the Contrast Strategy’. *In:* Fischer and Ravizza (eds.), *The Oxford Handbook of Contemporary Philosophy*. Oxford: Oxford University Press, pp. 272–277.

May, Douglas R. and Pauli, Kevin P. (2002). ‘The Role of Moral Intensity in Ethical Decision Making’. *Business & Society*, *41*: 84–117.

McCrae, R .R. and Costa Jr., P. T. (1985). ‘Comparison of EPI and psychoticism scales with measures of the five-factor model of personality’. *Personality and Individual Differences*, *6*(5): 587–597.

McKenna, Michael (1998). ‘The Limits of Evil and the Role of Moral Address: A Defense of Strawsonian Compatibilism’. *Journal of Ethics*, *2*: 123–142.

McKenna, Michael (2012). *Conversation and Responsibility*. New York: Oxford University Press.

McKenna, Michael and Russell, Paul (eds.) (2008). Free Will and Reactive Attitudes: Perspectives on P.F. Strawson's "Freedom and Resentment". Burlington, VT: Ashgate Publishing.

McMahan, Jeff (1988). 'Death and the value of life'. *Ethics, 99*(1): 32–61.

McMahan, Jeff (1993). 'Killing, Letting Die, and Withdrawing Aid'. *Ethics*, *103*(2): 250–279.

McMahan, Jeff (1994). 'Self-Defense and the Problem of the Innocent Attacker'. *Ethics*, *104*: 252–290.

McMahan, Jeff (2002). *The Ethics of Killing: Problems at the Margins of Life*. New York: Oxford University Press.

McMahan, Jeff (2005). 'The Basis of Moral Liability to Defensive Killing'. *Philosophical Issues*, *15*: 286–394.

McMahan, Jeff (2009a). *Killing in War*. Oxford: Oxford University Press.

McMahan, Jeff (2009b). 'Self-Defence Against Morally Innocent Threats'. *In*: P.H. Robinson, S.P. Garvey, and K.K. Ferzan (eds.), *Criminal Law Conversations*. New York: Oxford University Press, pp. 385–394.

McMahan, Jeff (2011). 'Who is morally liable to be killed in war?' *Analysis, 71*(3): 544–559.

McMahan, Jeff (2011). 'Who Is Morally Liable to Be Killed in War'. *Analysis,* 71: 548

McMahan, Jeff (2014). 'Self-Defense Against Justified Threateners'. *In*: H. Frowe and G. Lang (eds.), *How We Fight: Ethics in War*. Oxford: Oxford University Press, pp. 104–138.

McMahan, Jeff (2015). 'Proportionality and Time'. *Ethics*, *125*: 1–24.

McMahan, Jeff (2017a). ‘Proportionate Defense’. *In*: Jens David Ohlin, Larry May, and Claire Finkelstein (eds.), *Weighing Lives in War*. Oxford: Oxford University Press.

McMahan, Jeff (2017b). ‘Liability, Proportionality, and the Number of Aggressors’. *In*: Saba Bazargan-Forward and Samuel C Rickless (eds.), *The Ethics of War: Essays*. Oxford: Oxford University Press.

National Safety Council (2019). ‘Odds of Dying’. Available at: https://injuryfacts.nsc.org/all-injuries/preventable-death-overview/odds-of-dying/ [Accessed on 15 June 2019].

Norcross, A. (1994). ‘Introduction to the Second Edition’. *In:* B. Steinbock and A. Norcross (eds.), *Killing and Letting Die*. New York: Fordham University Press, pp. 1–23.

Norman, W. T. (1963). ‘Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings’. *The Journal of Abnormal and Social Psychology*, *66*(6): 574.

Norris, J. R. (1998). *Markov chains*. Cambridge: Cambridge University Press.

Nozick, Robert (1974). *Anarchy, State and Utopia*. Malden, MA: Basic Books.

Office, W. H. P. (2013). Fact sheet: U.S. policy standards and procedures for the use of force in counterterrorism operations outside the United States and areas of active hostilities.

Otsuka, M. (1994). ‘Killing the innocent in self-defense’. *Philosophy and Public* Affairs, *23*(1): 74–94.

Ożańska-Ponikwia, Katarzyna (2014). Emotions from a Bilingual Point of View: Personality and Emotional Intelligence in Relation to Perception and Expression of Emotions in the L1 and L2. Newcastle-upon-Tyne: Cambridge Scholars Publishing.

Parzen, E. (1962). *Stochastic processes*. San Francisco: Holden-Day.

Payne, Thomas H. (2000). ‘Computer Decision Support Systems’. *CHEST*, *118*(2): 47S-52S.

People v. Abraham (1997). A074868.

Perr, Irwin N. (1991). ‘Crime and Multiple Personality Disorder: A Case History and Discussion’. *Bulletin of the American Academy of Psychiatry and the Law*, *19*(2): 203–214.

Peterson, M. (2009). *An Introduction to Decision Theory*. Cambridge: Cambridge University Press.

Pidd, H. (2012). ‘Anders Behring Breivik spent years training and plotting for massacre’. *The Guardian*. Available at: http://www.theguardian.com/world/2012/aug/24/anders-behring-breivik-profile-oslo [Accessed 11 January 2014].

Popper, K. (1976). ‘A Note on Verisimilitude’. *The British Journal for the Philosophy of Science, 27*(2): 147–159.

Prince, Morton (1978). *The Dissociation of a Personality*. Oxford: Oxford University Press.

Pullman, Philip (2011). *Northern Lights: His Dark Materials*. London: Scholastic Ltd.

Quinn, Warren S. (1989). ‘Actions, intentions, and consequences: The doctrine of doing and allowing’. *Philosophical Review, 98*(3): 287–312.

Quong, Jonathan (2012). ‘Liability to Defensive Harm’. *Philosophy & Public Affairs*, *40*(1): 45–77.

Quong, Jonathan (2015). ‘Proportionality, Liability, and Defensive Harm’. *Philosophy & Public Affairs*, *43*(2): 144–173.

R v Grant (2014). EWCA Crim 143.

R v Moloney (1985). 1 AC 905.

Rachels, James (1975). ‘Active and passive euthanasia’. *The New England Journal of Medicine*, *292*: 78–80.

Rachels, James (1989). ‘More impertinent distinctions’. *In:* Robert M. Baird and Stuart E. Rosenbaum (eds.), *Euthanasia: The Moral Issues*. Buffalo, NY: Prometheus Books, pp. 61–8.

Raghunathan, R. and Tuan Pham, M. (1999). ‘All negative moods are not equal: Motivational influences of anxiety and sadness on decision making’. *Organizational Behavior and Human Decision Processes, 79*(1): 56–77.

Ramsey, P. F. (1926). Truth and Probability. *In:* H. E. Kyburg and H. E. K. Smokler (eds.), *Studies in Subjective Probability.* Huntington, NY: Robert E. Kreiger Publishing Co.

Resnik, M. D. (1987). *Choices: An Introduction to Decision Theory.* Minneapolis: University of Minnesota Press.

Rest, J. R. (1986). Moral Development: Advances in Research and Theory. New York: Praeger.

Richard Peto, Alan D Lopez, Hongchao Pan Jillian Boreham and Michael Thun (2015). ‘Mortality from smoking in developed countries 1950–2010’. Available at: http://gas.ctsu.ox.ac.uk/tobacco/C4308.pdf [Accessed 15 June 2019].

Robinson, P.H. and Kurzban, R. (2007). ‘Concordance and Conflict in Intuitions of Justice’. *Minnesota Law Review*, *91*: 1829–1907.

Robinson, Paul H. (1984). *Criminal Law Defenses*. St Paul: West Publishing Co., §171(c)(1).

Rodin, David (2002). *War and Self-Defense*. New York: Oxford University Press.

Rodin, David (2011). ‘Justifying Harm’. *Ethics*, *122*: 79.

Russell, Paul (1992). ‘Strawson's Way of Naturalizing Responsibility’. *Ethics*, *102*: 287–302.

Russell, Paul (1995). Freedom and Moral Sentiment: Hume's Way of Naturalizing Responsibility. New York: Oxford University Press.

Russell, Paul (2013). 'Responsibility, Naturalism, and 'The Morality System''. *In*: Daviv Shoemaker (ed.), *Oxford Studies in Agency and Responsibility* (Volume 1). New York: Oxford University Press.

Sachs, Jeffrey D. (2006). The End of Poverty: Economic Possibilities for Our Time. New York: Penguin.

Sacks, Oliver (2011). *The Man Who Mistook His Wife For A Hat*. London: Picador.

Scanlon, Thomas (1998). *What We Owe Each Other*. Cambridge, MA: Belknap Press of Harvard University Press.

Scanlon, Thomas (2008). *Moral Dimensions: Permissibility, Meaning, Blame*. Cambridge, MA: Belknap Press of Harvard University Press.

Scanlon, Thomas (2015). Forms and Conditions of Responsibility. *In*: R. Clarke, M. McKenna, and A.M. Smith (eds.), *The Nature of Moral Responsibility: New Essays*. Oxford: Oxford University Press.

Schoemaker, P. J. H. (1982). 'The Expected Utility Model: Its Variants, Purposes, Evidence and Limitations'. *Journal of Economic Literature, 20*(2): 529–563.

Sebok, Anthony (1996). 'Does an Objective Theory of Self-Defense Demand Too Much?'. *University of Pittsburgh Law Review*, *57*: 725–755.

Shoemaker, David (2007). 'Moral Address, Moral Responsibility, and the Boundaries of the Moral Community'. *Ethics*, *118*: 70–108.

Shoemaker, David (2011). 'Attributability, Answerability, and Accountability: Toward a Wider Theory of Moral Responsibility'. *Ethics*, *121*(3): 602–632.

Shoemaker, David and Tognazzini, Neal (eds.) (2015). *Oxford Studies in Agency and Responsibility* (Volume 2). New York: Oxford University Press.

Sider, Theodore (2005). 'Personal Identity'. *In:* E. Connee and T. Sider (eds.), *Riddles of Existence: A Guided Tour of Metaphysics*. Oxford: Oxford University Press.

Singhapakdi, Anusorn, Vitell, Scott J., and Kraft, Kenneth L. (1996). 'Moral Intensity and Ethical Decision-Making of Marketing Professionals'. *Journal of Business Research*, *36*(3): 245–255.

Skyrms, B. (1986). *Choice and Chance.* Belmont, CA: Wadsworth.

Slovenko, Ralph (1995). *Psychiatry and Criminal Culpability*. New York: John Wiley and Sons, Inc.

Smith, Angela (2005). 'Responsibility for Attitudes: Activity and Passivity in Mental Life'. *Ethics*, *115*: 236–71.

Smith, Angela (2012). 'Attributability, Answerability, and Accountability: In Defense of a Unified Account'. *Ethics*, *122*(3): 575–589.

Squires, P. (2006). 'Beyond July 4th?: Critical Reflections on the Self-Defence Debate from a British Perspective'. *Journal of Law, Economics and Policy*, *2*(2): 221–264.

Steinbock, B. (1994). 'Introduction'. *In:* B. Steinbock and A. Norcross (eds.), *Killing and Letting Die*. New York: Fordham University Press, pp. 24–47.

Steinhoff, U. (2008). 'Jeff McMahan on the Moral Inequality of Combatants'. *Journal of Political Philosophy*, *16*(2): 220–26.

Stern, Lawrence (1974). 'Freedom, Blame, and the Moral Community'. *Journal of Philosophy*, *71*: 72–84.

Strawson, P. F. (1962). 'Freedom and Resentment'. *Proceedings of the British Academy*, *48*: 1–25.

Sullivan, T. D. (1977). 'Active and passive euthanasia: an impertinent distinction?'. *The Human Life Review*, *3*(3): 40–46.

Swift, L. J. (1970). 'St. Ambrose on violence and war'. Transactions and Proceedings of the American Philological Association, 101: 533–43.

Tadros, Victor (2012). 'Duty and Liability'. *Utilitas*, *24*(2): 259–277.

Taurek, John M. (1977). ‘Should the numbers count?’ *Philosophy and Public Affairs*, *6*(4): 293–316.

Taylor, M. (2011). ‘Norway gunman claims he had nine-year plan to finance attacks’. *The Guardian*. Available at: http://www.theguardian.com/world/2011/jul/25/norway-gunman-attack-funding-claim [Accessed 3 March 2015].

Tenbrunsel, A. E. and Messick, D. M. (2004). ‘Ethical Fading: The Role of Self-Deception in Unethical Behavior’. *Social Justice Research, 17*(2): 223–236.

Tenbrunsel, Ann E. and Messick, David M. (1999). ‘Sanctioning Systems, Decision Frames, and Cooperation’. *Administrative Science Quarterly*, *44*(4): 684–707.

Tenbrunsel, Ann E. and Smith-Crowe, Kristin (2008). ‘Chapter 13: Ethical Decision Making: Where We've Been and Where We're Going’. *The Academy of Management Annals*, *2*(1): 545–607.

Thomson, Judith Jarvis (1991). ‘Self-defense’, *Philosophy and Public Affairs*, *20*(4): 283–310.

Thomson, Judith Jarvis (1996). ‘Critical Study of the Act Itself’. *Nous*, *30*: 545–57.

Tijms, H. C. (2003). *A first course in stochastic models*. New York: Wiley.

Tooley, Michael (1994). ‘An Irrelevant Consideration: Killing versus Letting Die’. *In*: B. Steinbock and A. Norcross (eds.), *Killing and Letting Die*. New York: Fordham University Press.

Valentine, S. and Fleischman, G. (2003). ‘Ethical Reasoning in an Equitable Relief Innocent Spouse Context’. *Journal of Business Ethics*, *45*(4): 325–339.

Vallier, Kevin (2010). ‘Thomas Scanlon, Moral Dimensions: Permissibility, Meaning, Blame’. Journal of Value Enquiry, 44: 561–565.

van Fraassen, B. (1998). *Laws and Symmetry.* Oxford: Oxford University Press.

Walen, A. (2015). ‘Proof beyond a reasonable doubt: A balanced retributive account’. *Louisiana Law Review, 76*(2): 355–446.

Wallace, R. J. (1994). *Responsibility and the Moral Sentiments*. Cambridge, MA: Harvard University Press.

Watson, Gary (1987). 'Responsibility and the Limits of Evil'. *In*: Ferdinand Schoeman (ed.), *Responsibility, Character, and the Emotions*. New York: Cambridge University Press.

Watson, Gary (1996). 'Two Faces of Responsibility'. *Philosophical Topics*, *24*: 227–248.

White, D. J. (1993). *Markov Decision Processes*. Chichester: John Wiley and Sons Ltd.

Woollard, Fiona (2008). 'Doing and Allowing, Threats and Sequences'. *Pacific Philosophical Quarterly, 89*: 261–277.

Woollard, Fiona (2012). 'The Doctrine of Doing and Allowing II: The Moral Relevance of the Doing/Allowing Distinction'. *Philosophy Compass*, *7*: 459–469.

Woollard, Fiona (2013). 'A Defence of the Doctrine of Doing and Allowing'. *Pacific Philosophical Quarterly*, *94*: 315–341.

Woollard, Fiona (2015). *Doing and Allowing Harm*. Oxford: Oxford University Press.

Yetmar, Scott A. and Eastman, Kenneth K. (2000). 'Tax Practitioners' Ethical Sensitivity: A Model and Empirical Examination'. *Journal of Business Ethics*, *26*(4): 271–288.

Zimmerman, Michael (1988). *An Essay on Moral Responsibility*. Totowa, NJ: Roman and Littlefield.

Zimmerman, M. J. (2008). *Living with uncertainty: The moral significance of ignorance*. Cambridge: Cambridge University Press.