Self-Defence as the Just Distribution of Harm

Weijia Zhang

UCL

MPhil Stud Philosophical Studies

Declaration

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

Abstract

This thesis examines the ethics of defensive harm through the lens of distributive justice. I argue that self-defence is best understood not as an exception to ordinary moral constraints, but as a problem of allocating and redistributing unavoidable harms. Chapter 1 addresses the question of the nature of liability. I critically assess rights-based and duty-based theories, including forfeiture, enforcement, and enforceable-duty accounts, and show that they fail to explain proportionality, necessity, and scalar liability. I then advance the discounting account, according to which liability consists in a reduction in the moral disvalue of harms borne by threateners relative to others. Chapter 2 turns to the grounds of liability. I argue that prevailing accounts—culpability, causation, moral status, and enforceable duty—are insufficient. I develop a revised moral responsibility account that treats liability as graded by voluntary initiation of risk, evidence-relative awareness, and causal proximity, and show how this integrates with the discounting framework. Chapter 3 addresses minimally responsible threats. Building on and revising Saba Bazargan's hybrid justification, I propose a hybrid explanatory account that clarifies why intuitions diverge on the permissibility of killing such threats. The thesis as a whole defends the claim that self-defence is justified when it realises, or most closely approximates, the just distribution of harm.

Impact Statement

This thesis advances a distributive account of self-defence, recasting defensive harm as a problem of allocating unavoidable costs rather than as a matter of rights forfeiture or enforceable duties. Academically, it contributes a novel discounting account of liability and a revised hybrid justification, both of which clarify proportionality, necessity, and divergent intuitions in cases of minimally responsible threats. These models provide a fresh foundation for future scholarship in moral and political philosophy, with implications for debates on punishment, risk, and the ethics of war.

Beyond academia, the thesis offers a framework with potential legal and policy relevance. By replacing binary notions of liability with a scalar and distributive model, it supports more nuanced standards for evaluating self-defence, excessive force, and responsibility under duress. More broadly, it enriches public debate by providing a principled vocabulary for thinking about fairness in tragic conflict situations, thereby linking philosophical analysis with practical concerns about justice and responsibility.

Acknowledgement

I am deeply grateful to my supervisor, Joe Horton, for his guidance, encouragement, and intellectual generosity throughout this project. I would also like to thank my cohort for a stimulating and supportive environment during my studies. I owe heartfelt thanks to my parents, whose constant love and support have sustained me throughout my studies and beyond. Finally, I would like to express my deepest gratitude to Yuchen, for her patience and companionship.

Table of Contents

Introduction

This thesis concerns the ethics of defensive violence. Generally, harming others is morally wrong. But one intuitive exception is harming others in self-defence. Imagine that a villain is culpably trying to kill you, and the only way to stop him is to kill him. Most people would agree that killing him is morally permissible. However, although most believe it is sometimes permissible to harm others in self-defence, moral philosophers disagree about when and why it is permissible.

This thesis contributes to debates related to the ethics of self-defence. I intend to explore two important aspects: a) the grounds of permissible defensive harm and b) the limits of the use of force in self-defence. My hypothesis is that the answer to the two questions can be unified by the idea that the ethics of self-defence is a matter of distributive fairness. Self-defence is permissible when, and because, it matches (or approximates) the just distribution of harm. To argue for my hypothesis, I plan to work on three sub-topics: 1) the Nature of Liability, 2) the Grounds of Liability, and 3) the Hybrid Justification of Self-Defence.

Chapter 1: The Nature of Liability. This chapter asks what it is to be liable to defensive harm. It argues that rights-forfeiture, rights-enforcement, and enforceable-duty accounts are inadequate: they face incompleteness/redundancy worries and cannot accommodate proportionality, necessity, or the scalar structure of liability. The chapter develops the discounting account: liability consists in a reduction of the moral disvalue of harms borne by those who, in responsibility-relevant ways, threaten others. On this view, liability is scalar and directly grounds proportionality (permissions expand with responsibility and threatened severity) and necessity (among equally effective means, choose the option with least discounted moral cost). The chapter also addresses familiar objections (e.g., micro-rights strategies; modelling disproportionate defence) and clarifies the distinction between physical magnitude of harm and its moral disvalue.

Chapter 2: The Grounds of Liability. This chapter asks under what conditions, and to what degree, an agent becomes liable. It critically examines culpability, causal, moral-status, and duty-based grounds, arguing that each misfires in characteristic ways (over- or under-inclusiveness; misplaced appeal to practice-level justification; neglect of interpersonal cost-internalisation). The chapter then develops a revised moral-responsibility account on which degrees of liability track three variables: (i) voluntary initiation or maintenance of a risk-imposing activity, (ii) evidence-relative awareness of a non-trivial risk to others, and (iii) causal proximity to the unfolding threat. This graded model yields stable rankings across contested cases (e.g., culpable aggressors, justified-mistaken attackers, careful drivers, coerced agents, pure bystanders) and integrates with the discounting account from Chapter 1 to generate determinate proportionality and necessity verdicts without sharp epistemic thresholds.

Chapter 3: A Hybrid Justification for Self-Defence. This chapter analyses minimally responsible threats and the associated divergence of intuitions about the permissibility of killing such agents. It reconstructs and revises Bazargan’s hybrid strategy. The chapter retains a proportional, responsibility-sensitive account of what an agent is liable to but rejects a rights-forfeiture reading and the need for a second responsibility-grounded “discount” at the lesser-evil stage. On the proposed hybrid explanatory account, liability itself is the responsibility-indexed discount; any further harm is assessed by ordinary lesser-evil reasoning subject to familiar constraints (including the doing/allowing distinction and means-based restrictions). The chapter shows how this framework (a) explains divergent intuitions in minimally responsible threat cases, (b) handles multi-threat scenarios, and (c) clarifies the role of ideal distribution and compensation when justified defensive outcomes remain non-ideal.

The three chapters are cumulative. Chapter 1 provides the conceptual account of liability as discounting. Chapter 2 supplies the graded grounds that determine the degree of discounting in concrete cases. Chapter 3 applies the combined framework to the hardest cases and explains the persistence of reasonable disagreement about permissibility. The overall conclusion is that self-defence is justified when—and because—it realises, or most closely approximates, a fair distribution of discounted moral costs.

Chapter 1: The Nature of Liability

1. Introduction

Seriously harming another person is ordinarily considered one of the gravest moral wrongs. Across both common-sense morality and most philosophical theories, there is widespread agreement that individuals possess robust rights not to be killed, injured, or otherwise harmed. These rights are thought to be stringent—meaning that they cannot easily be overridden or ignored, even in the pursuit of morally good ends.

Classic thought experiments illustrate this. In the Transplant Case (Thomson 1985, p.1396), a surgeon can kill one healthy patient to harvest their organs and save five others. Almost everyone judges this impermissible: the right of the one not to be killed does not yield to considerations of aggregate welfare. By contrast, in the Trolley Case (Thomson 1985, p.1395), where one can divert a runaway trolley to kill one person on the sidetrack rather than allow it to kill five on the main track, many judge the killing permissible. The contrast shows that killing one to save five is not uniformly impermissible or permissible: much depends on the nature of the threat and the way in which harm is distributed.

Against this background, cases of self-defence pose a distinctive challenge. Imagine someone who, unprovoked, charges at you with a knife. You cannot escape. The only way to save your life is to strike first and kill your attacker. Most people, including philosophers, would consider your action permissible. Some would even argue that it is morally required. What makes this case different not only from Transplant, but also from the Trolley Case, where the harm is redirected rather than resisted? In the self-defence case, the attacker seems to lose their ordinary protection against harm. Something about their conduct transforms the moral landscape such that they are no longer shielded in the usual way.

This observation introduces the concept of liability. In moral and legal theory, to be liable to defensive harm is to be someone whom it is permissible to harm for the sake of defending oneself or others. But what does it mean to be liable? What sort of moral status does this describe, and what are the grounds on which one becomes liable?

This chapter explores and evaluates the major theories of the nature of liability. These theories are not concerned primarily with when harming is permissible—that is the subject of subsequent chapters—but rather with what it means to say that a person is "liable to be harmed," and with what kind of moral transformation that claim entails.

The first theory I examine is the Rights Forfeiture Theory, which holds that persons who engage in culpable wrongdoing forfeit the moral rights that would ordinarily protect them from harm. The second is Draper's Rights Enforcement Theory, which tries to retain a rights-based structure without invoking forfeiture. The third is the Enforceable Duty Theory, which explains liability in terms of obligations that others may permissibly enforce. I conclude by defending the Discounting Theory, which treats liability as a matter of reducing the moral weight of harms suffered by those who pose unjust threats.

This last theory reflects a broader theme that I will emphasize throughout: liability is not best understood as an all-or-nothing status. Instead, it is scalar. Individuals can be more or less liable, depending on factors such as culpability, the gravity of the threat, and the fairness of imposing costs. This scalar structure allows us to make sense of proportionality in self-defence, and it will also prove essential in later chapters, where I turn to other problems in the ethics of self-defence.

2. The Rights Forfeiture Theory and Its Objections

2.1 The Rights Forfeiture Theory

The Rights Forfeiture Theory is arguably the most widely accepted and historically entrenched account of liability in moral and political philosophy. According to this theory, individuals typically have rights not to be harmed. These rights impose duties on others not to kill, injure, or interfere with them in certain ways. However, under certain conditions—most prominently, conditions of culpable wrongdoing—these rights can be forfeited. When someone forfeits their right not to be harmed, others are permitted to treat them in ways that would otherwise be morally impermissible (Thomson 1991; McMahan 2005; Lang 2014).

In the context of self-defence, the claim is that a person who culpably poses an unjust threat to another forfeits their moral right not to be harmed. This forfeiture removes the moral constraint that would otherwise make harming them wrong. On this view, the permissibility of defensive harming is not an exception to ordinary moral prohibitions, nor a case of rights being overridden by more important considerations. It is a case of rights no longer applying.

A classic example can help illustrate the theory's appeal. Suppose that Jordan culpably attempts to kill Alex without justification. To save himself, Alex strikes Jordan in the head with a heavy object, killing him. Suppose this is the only way Alex can save himself. Most would agree that Alex's action is morally permissible. The rights forfeiture theorist explains this by saying that Jordan, by attempting an unjustified and culpable attack, forfeited his right not to be killed. Thus, Alex did not violate any right in killing Jordan, and the act was not wrongful.

In this way, the theory preserves the importance of rights in moral theory while allowing for the permissibility of defensive harm in a principled way. The forfeiture of rights is not arbitrary; it is typically understood to require culpability—that is, the person must have knowingly or negligently chosen to engage in wrongful action. Non-culpable threats, such as those posed by children or the mentally ill, are often considered not to trigger forfeiture. This focus on culpability helps explain why it is permissible to harm attackers but not innocent bystanders: the attacker, through their actions, has forfeited moral protections that the bystander retains.

2.2 Objections to the Rights Forfeiture Theory

Despite its intuitive appeal and widespread use, the Rights Forfeiture Theory has attracted significant criticism. One of the most sustained and rigorous critiques comes from Massimo Renzo (2017), who argues that the theory fails to provide a complete or coherent account of liability. His objections focus on what he calls the Incompleteness Problem and the Redundancy Problem, both of which reveal deeper conceptual difficulties in the idea of forfeiture (Renzo 2017: 329–331, 338–341).

The Incompleteness Problem concerns the explanatory power of the theory. Simply saying a person has forfeited their rights does not explain why they have lost those rights. The notion of forfeiture functions more like a label than a justification. It tells us that the person is no longer protected by their right, but not what grounds that moral transformation (Renzo 2017: 329–333).

The Redundancy Problem follows naturally from this. When theorists attempt to provide more substantive accounts of forfeiture—by appealing to culpability, intention, responsibility, or other moral concepts—they end up doing the real justificatory work with those other principles. Once those deeper moral concepts are in place, the language of forfeiture becomes unnecessary. This criticism undermines the central role that forfeiture is supposed to play in explaining liability (Renzo 2017: 340–341).

I now propose another serious objection to the rights forfeiture theory. It concerns its inability to account for degrees of liability, and by extension, its failure to make sense of the proportionality requirement in self-defence. The core of the problem lies in the structure of the rights concept itself. Moral rights, at least as traditionally conceived, are binary in nature: either a person possesses a right, or they do not. There is no obvious room within this framework for saying that someone has “less” of a right, or that they have partially forfeited it. This all-or-nothing structure renders the theory ill-equipped to handle the kinds of proportionality judgments that are central to our moral understanding of defensive harm.

To see this more clearly, consider one of the most common interpretations of proportionality in self-defence: namely, that the harm one inflicts must be proportionate to the harm one is preventing. If someone threatens to break your arm, it is generally considered impermissible to kill them in order to prevent that harm. This is not just a legal norm, but a deeply embedded moral intuition. However, the rights forfeiture model cannot easily explain why this is so. If the aggressor is culpable and poses an unjust threat, then—on the standard model—they forfeit their rights not to be harmed. But that forfeiture alone does not distinguish between the level of harm they may permissibly suffer. Once the right is gone, it is gone; there is no principled way within the theory to say how much harm they are liable to.

To bring out the difficulty, imagine two parallel cases. In the first, a person is threatened with death by a culpable aggressor. In the second, a different person is threatened with a broken arm by another culpable aggressor. According to the forfeiture view, both aggressors have forfeited their right not to be harmed. But intuitively, we believe that lethal force is permissible in the first case and impermissible in the second. The difference between the two cases lies not in the aggressor’s rights—they have both forfeited them—but in the gravity of the threat they pose. Yet the theory gives us no conceptual tools to register this difference. It treats both as equally exposed to defensive harm because both have forfeited the same kind of right.

Someone might try to rescue the theory by claiming that rights are more fine-grained than we initially assumed—that there is a right against being killed, and a separate right against being injured. Under this view, perhaps a culpable aggressor forfeits only certain rights depending on the harm they intend to cause. But this response only delays the problem. Imagine two new cases. In the first, an aggressor threatens to break both of your arms. In the second, a different aggressor threatens to pinch you painfully, but without causing lasting damage. Suppose both aggressors are fully culpable. Do they forfeit the same “right against being harmed”? If so, we would expect them to be liable to the same kinds of defensive harm. But that conclusion is clearly mistaken. The first aggressor may justifiably be met with considerable force, perhaps even serious injury. The second may not be justifiably harmed at all, or at most in a very minimal way. Yet both are said to have forfeited the relevant right. Once again, the model fails to explain the difference in moral exposure between them.

A seemingly promising response is this: Could we posit even more rights—a distinct right against every type or magnitude of harm? On this proposal, the generic claim against being harmed is replaced by a finely partitioned set of micro-rights (a right against being pinched, against minor injury, against serious maiming, against being killed, etc.), and forfeiture is said to occur piecemeal: a culpable aggressor loses protection only with respect to the particular level of harm that may permissibly be imposed.

While this picture might appear to capture fine-grained distinctions, it faces two serious difficulties. First, it is *ad hoc*. The proliferation of micro-rights mirrors our antecedent proportionality judgements rather than explaining them. There is no independent rationale for drawing the boundaries where they are. Second, and more importantly, it fails to account for the distinctive status of disproportionate defensive harm as a *lesser* wrong. Suppose an aggressor is liable to harm of magnitude X. But the defender in fact imposes harm of magnitude Y, which Y is slightly greater than X. We ordinarily judge that the defender acts wrongly, yet less wrongly than someone who imposes Y on an innocent person. The former is a case of *disproportionate defence*, not a wholly unprovoked wrong. By contrast, if we model liability as the forfeiture of a right *precisely to* X and maintain a separate, intact right *not to suffer* Y, then imposing Y necessarily constitutes a *full* violation of the (still-held) “not suffer Y” right—rendering the disproportionate defence morally on a par with imposing Y on an innocent person. This strategy thus collapses the very gradation we seek to explain.

The deeper issue is that rights language lacks a scalar structure. It does not allow for talk of degrees of moral protection, or of moral interest being discounted in proportion to the threat posed or the responsibility borne. As such, it cannot ground the nuanced moral assessments that real-life defensive scenarios demand. My alternative, the discounting view, is designed precisely to capture this complexity. Rather than saying someone has lost a right and is now categorically liable, I say that the moral disvalue of harming them is reduced in proportion to factors like their culpability, the seriousness of the threat, and the moral costs at stake. This approach retains our commitment to proportionality while avoiding the rigid binaries of the rights forfeiture framework.

Another serious problem with the rights forfeiture theory is related to the necessity requirement of self-defence. The necessity requirement, as I use it, is the requirement in self-defence to choose the least harmful defensive option among the options that are equally effective. For instance, if a culpable aggressor is trying to kill you. You can either defend yourself by killing him, or you can stun him into unconsciousness without any serious harm. Necessity requires you to stun him and forbids you from killing him. The necessity requirement is intuitive and widely-accepted. But what can the rights forfeiture theory say about it? In a case where killing is the only way to stop the attack, the aggressor is said to have forfeited his right not to be killed. Killing him does not violate any right. If killing does not violate any right in that case, then by the same logic a lesser harm would not violate a right either. So, when a less harmful but equally effective option exists, the forfeiture framework does not explain why we must choose the less harmful one since both options would violate no right.

One might reply by claiming that a person can only be liable to necessary harm, and thus, can only forfeit her rights when facing a necessary defensive action. I believe such a strategy is implausible. It is *ad hoc*. It builds the necessity requirement into the definition of liability without explaining it. I will, in the later section, explain how my account of the nature of liability can adequately explain the necessity requirement.

To summarize, the rights forfeiture theory of liability cannot serve as a satisfying account of the nature of liability. First, it suffers from two major objections offered by Renzo – the incompleteness problem and the redundancy problem. More importantly, I have demonstrated that the binary structure of rights cannot capture the idea of scalar liability and the requirement of proportionality. Furthermore, the rights forfeiture theory cannot offer an adequate explanation of the necessity requirement of self-defence.

3. Draper's Rights Enforcement Theory and Its Objections

3.1 The Rights Enforcement Theory

In response to the challenges faced by traditional rights forfeiture accounts, Kai Draper (2009, 2016) offers an alternative theory of liability, which he calls the rights enforcement account. Draper's goal is to preserve the moral centrality of individual rights, while avoiding what he sees as the conceptual pitfalls of the forfeiture model—especially its metaphysical claims about how rights are "lost" or "forfeited." According to Draper, the justification for defensive harm lies not in the forfeiture of the aggressor's rights, but in the moral permissibility of enforcing the rights of others, in such a way that the aggressor's rights are not themselves infringed.

At the core of this view is the idea that defensive harm can be justified without requiring any prior loss of rights on the part of the aggressor. In Draper's account, an individual who poses a threat of unjust harm may be harmed permissibly because doing so prevents that rights violation, not because the aggressor has forfeited any moral protections. What makes harming the aggressor permissible is not that their rights no longer exist, but that they never had a right not to be stopped, at least not under the circumstances created by their actions.

Draper believes this model better reflects the structure of justifiable defensive harm. He holds that, in standard cases of self-defence—where an attacker is about to harm a victim—it is not that the attacker starts with a right not to be harmed and then loses it. Rather, due to the nature of their threatening conduct, they are not the kind of agent who possesses such a right in that context. Defensive harm thus enforces the victim's rights without violating the aggressor's.

Moreover, the theory avoids the metaphysical language of rights being "lost" or "relinquished." Instead, it offers a conception of rights that is context-sensitive. A right not to be harmed is not an absolute protection, but one that applies only when the individual is not actively threatening the rights of others in a way that justifies intervention.

From Draper's perspective, then, liability to defensive harm is a matter of being non-rights-bearing in that context. This makes the rights enforcement theory structurally distinct from forfeiture accounts. Forfeiture theories posit a change in moral status—from rights-holder to non-rights-holder—brought about by wrongful conduct. Draper, by contrast, denies that such a change needs to occur: some individuals never possess the relevant right in the first place when they engage in conduct that unjustly threatens others.

3.2 Objections to the Rights Enforcement Theory

Despite the conceptual refinements Draper introduces, the rights enforcement theory encounters all the same problems that rights forfeiture theory faces. This is largely because the two views are structurally similar: both are rights-based accounts that aim to explain the permissibility of defensive harming by appeal to the *status of the* *aggressor's moral rights*. The forfeiture theory says the aggressor *loses* the relevant right, while the enforcement theory says the aggressor's right is *inapplicable in this context* and the defender merely *enforces* her right. I now show, point by point, why the enforcement view inherits each difficulty identified for forfeiture.

First, the rights enforcement theory, like rights forfeiture theory, faces the *incompleteness* problem. In the forfeiture case, merely asserting that the aggressor "has forfeited his right" labels the outcome but does not explain *why* the right disappears—what justifies the normative transition from mutual protection to a liberty to harm. By parity, the enforcement view's claim that "the aggressor *has no claim* against harm here and the defender is simply *enforcing* her right" also only *re-describes* the permissibility verdict without explaining *why* the rights of the aggressor becomes inapplicable *in this context*. In both frameworks, the key step—removing the aggressor's protective claim—needs a ground beyond the label ("forfeited" vs "inapplicable"), and until that ground is provided the account remains incomplete.

Now let us turn to the redundancy problem. The rights enforcement theory, like rights forfeiture theory, also runs into redundancy once it supplies the missing ground. According to the forfeiture theory, when one appeals to culpability, responsibility, initiation of unjust threats, or fairness in the distribution of costs to explain permissibility, those deeper principles do the actual justificatory work and "forfeiture" adds nothing. Likewise for enforcement, once we explain *why* the aggressor's claim is inapplicable by invoking the very same principles (culpability, responsibility, fairness), the further gloss—"this is *enforcement* rather than *violation*"—adds no independent reason. The explanatory weight sits with the background principles, not with the *enforcement* label, so the theory becomes redundant in just the same way.

Having explained why the rights enforcement theory still faces the problems proposed by Renzo, I will now explain why it also faces my objections to the forfeiture theory as demonstrated above. First, I will explain why the enforcement theory cannot accommodate scalar liability and the requirement of proportionality. The key reason is that the enforcement framework is still binary at the crucial point. Either the aggressor's right not to be harmed is inapplicable in this context or it is applicable. That binary switch does not by itself yield a graded account of how much harm may permissibly be imposed. As above, consider two otherwise similar cases—one involving a lethal threat, the other a threat to break an arm. On the enforcement view, in both cases the aggressor lacks the right not to be harmed in this context, so both are equally open to "enforcement." But our proportionality judgements distinguish them: lethal force seems permissible in the first and impermissible in the second. The enforcement theory, without further explanation, lacks an internal metric for the *degree* of permissible harm once the claim is deemed inapplicable.

Proponents of the enforcement theory might reply, as proponents of the forfeiture theory might do, in enforcement terms, that there are distinct rights and claims corresponding to different harm-levels, and that only some become inapplicable in the defensive context (e.g., the claim not to be killed is inapplicable, but the claim not to be seriously injured remains). As with the forfeiture analogue, this merely delays the problem. Re-run the examples with an aggressor who threatens to break both your arms versus one who threatens only a painful pinch: if both "lack the relevant claim" at the level of "being harmed," why are the permissible responses so different? If we now say that only the "no-serious-injury" claim is inapplicable in the first case and none is inapplicable in the second, we have simply fitted the claims to our verdicts, not explained them.

Pushing the same idea further, could we posit an even more finely partitioned set of context-inapplicable claims—one for every type or magnitude of harm—so that enforcement targets only the precise level to which the aggressor "lacks a claim" here and now? As above, this is ad hoc. This strategy mirrors the proportionality judgements rather than grounding them. And more importantly, it cannot capture the distinctive moral status of slightly disproportionate defence as a *lesser* wrong. Suppose, on the enforcement picture, the aggressor has *no claim* against harms up to magnitude X, but retains a claim against Y where Y just slightly exceeds X. If the defender imposes Y, she now commits a full violation of the still-applicable “no-Y” claim—morally on a par with imposing Y on an innocent. Yet our ordinary judgements treat slight over-enforcement as less wrongful than a gratuitous Y harming of an innocent. The micro-claims strategy therefore erases precisely the gradation we seek to accommodate.

Furthermore, the enforcement theory, like the forfeiture theory, cannot fully explain the requirement of necessity. As above, necessity requires choosing the least harmful among the equally effective defensive options. Translate this into enforcement language. In a variant where only lethal force would successfully enforce the victim’s right (say, the only way to stop the attack is to kill), the aggressor’s claim against lethal harm is said to be inapplicable, and killing does not violate a right. But then, in a different variant where a non-lethal option would enforce the victim’s rights equally well, the enforcement view still believe that both options do not violate the rights and claims of the aggressor. If the aggressor has no claim against the relevant defensive imposition in this context, *neither* killing *nor* stunning infringes a claim. From *within* enforcement theory, there is therefore no explanation for why we must choose the less harmful means. The usual escape routes repeat the forfeiture pattern. One can only internalise necessity by stipulating that only *necessary* enforcement is permitted (which is ad hoc, because it builds the answer into the definition of permissible enforcement).

To summarize, in each respect, the rights enforcement framework replicates the problems that undermined forfeiture. Its core move—declaring the aggressor’s claim inapplicable and the defensive harm an instance of enforcement—does not itself explain the normative transition (incompleteness); once the real grounds are supplied, the enforcement label adds nothing (redundancy); its binary structure cannot generate scalar liability or capture the requirement of proportionality; and it cannot ground the necessity requirement without *ad hoc* stipulation. Thus, for the same structural reasons as above, rights enforcement theory cannot provide a satisfactory account of the nature of liability.

4. The Enforceable Duty Theory and Its Objections

4.1 The Enforceable Duty Theory

An influential alternative to the rights-based theories is the Enforceable Duty Theory, developed by Victor Tadros (2012; 2016). Unlike the forfeiture model, this account does not depend on the idea that rights are lost or suspended. Instead, it holds that people become liable to defensive harm when they have a moral duty to bear the cost of that harm, and that duty is one that others are permitted to enforce. Liability, in this view, is not about the withdrawal of protections but about the permissibility of imposing burdens on someone to discharge a duty they have.

To illustrate the basic idea, consider a classic rescue case. Imagine a man named Daniel walking along a riverbank. He notices a child drowning but does nothing, even though he could easily jump in and save the child at some personal cost (say, ruining his expensive suit). Most people would say Daniel has a duty to act. Now suppose Daniel refuses, and someone pushes him into the water to compel him to save the child. According to Tadros, Daniel may be harmed—pushed into the river—not because he has forfeited his rights, but because he has a duty to save the child, and that duty is enforceable. That is, others are permitted to impose the relevant cost on him to ensure the duty is fulfilled.

Tadros extends this reasoning to defensive harm. He argues that if a person unjustly threatens another—whether culpably or not—they may have a duty to bear certain costs (even lethal costs) to prevent that harm from occurring. For example, if a soldier fighting in an unjust war poses a threat to civilians, he may have a duty to stand down or withdraw. If he fails to do so, others may justifiably kill him to stop the threat. This is not because he has forfeited his rights, but because he has a duty to cease participating in injustice, and others are permitted to enforce that duty through proportionate harm.

This account avoids some of the moral excesses of forfeiture. On Tadros's view, liability is always linked to a specific duty, and harm must be proportionate and necessary to enforcing that duty. It is not the case that one loses all rights or becomes fair game for any kind of violence. Rather, liability corresponds to the burdens that must be borne to respect the rights of others, and those burdens can sometimes be morally imposed.

Also, Tadros's account aligns with legal and political thinking in some respects. In many legal systems, people are required to bear certain burdens for the public good—such as paying taxes or serving on juries—and these duties are enforceable through coercive mechanisms. Tadros's account, he suggests, is not so different: when someone is morally required to act in a way that prevents harm to others, it may be permissible to compel them to do so—even if compulsion causes harm.

In summary, the enforceable duty theory offers a principled, duty-based explanation of liability. It replaces the language of lost rights with the more grounded notion of enforceable obligations and explains why harming a threatener can be permissible without assuming that their rights no longer exist.

4.2 Objections to the Enforceable Duty Account

Despite its strengths, I believe that the enforceable duty theory also faces a number of serious objections. While it avoids some of the pitfalls of the forfeiture model, I will now argue that it faces its own problems.

One major problem is that the theory might overgenerate liability. If people can be liable whenever they fail to fulfill some enforceable duty, then the class of liable persons could become implausibly large. For example, if I walk past a pond where a stranger is drowning and fail to act, do I thereby become liable to be harmed by someone trying to make me help? Could I be physically dragged into the water, or fined, or imprisoned? While such actions may seem proportionate in some cases, in others they would feel like a moral overreach.

This problem becomes acute in defensive scenarios. Suppose a bystander sees an attack unfolding but does nothing. Under Tadros's theory, if that inaction amounts to a failure to enforce a duty to prevent harm, then the bystander might be liable to defensive harm. This seems too strong. Our intuitions usually distinguish clearly between aggressors and mere bystanders, even passive ones. Liability should not be so easily triggered by mere failure to help.

Tadros, however, is willing to accept the above implications, which I believe are problematic. I will explore this issue in detail in the next chapter in which I discuss the grounds of liability. Even if we leave this problem aside, however, the Enforceable Duty Theory faces other problems.

The second objection I will raise concerns the inflexibility of the enforceable duty theory in accounting for differences in the moral weight or gradability of duties. One of the major advantages this theory claims over rights-based models is that it avoids the all-or-nothing structure of rights. In rights-based frameworks, an individual either has a right against harm or does not. This binary logic makes it difficult to distinguish between degrees of liability: the aggressor who threatens to punch you is treated, structurally, the same as the aggressor who threatens to kill you—both either have or have not forfeited their rights.

The enforceable duty theory attempts to avoid this problem by not invoking rights at all. Instead, it holds that a person becomes liable to defensive harm when they have a specific duty to bear a cost—namely, the cost of being stopped from inflicting unjust harm—and that this duty is enforceable. This seems to allow for contextual sensitivity: for example, a culpable aggressor who attempts to kill you may have a duty to bear lethal harm if that is the only way to stop them. In contrast, a different aggressor who only threatens to break your arm may have a duty to bear proportionately less harm, such as being disarmed or subdued non-lethally. Thus, the content of the enforceable duty appears to depend on the gravity of the threat and the means necessary to avert it. This is an important improvement over the rigid rights-forfeiture model.

However, once we introduce more nuanced examples, a deeper problem emerges. Suppose now that the threat is the same in two cases: someone threatens to break your arm. In the first case, the aggressor is fully culpable. In the second, the aggressor is non-culpable—perhaps they are acting under a genuine but mistaken belief, or under coercion, or due to some excusing condition. Intuitively, we are inclined to treat these two cases differently. In the culpable case, we might judge that the aggressor has a duty to bear more severe defensive harm—perhaps even a serious injury. In the non-culpable case, we might think that you may only impose the minimum necessary harm, and even then, possibly only reluctantly or with moral regret. Yet on Tadros's model, both individuals have an enforceable duty to bear the cost necessary to stop them. What the theory cannot explain is why the scope and strength of that duty should differ based on culpability. If both duties are enforceable, and both aggressors are posing the same level of threat, then both should be liable to the same level of defensive harm. But our intuitions—and I would argue, our moral reasoning—say otherwise.

The deeper issue is that the enforceable duty model lacks any internal structure for calibrating the degree of the duty. Once someone falls within the category of bearing an enforceable duty, the theory provides no further resources to say whether their duty is stronger, weaker, or morally weightier than someone else's. But the morality of self-defence is not just about whether harming someone is permitted—it is also about how much harm is justified in response to different kinds of threats and different kinds of agents. Without a theory that can track degrees of responsibility, variation in justification, or proportionality grounded in the threatener's moral status, the enforceable duty view remains too coarse-grained. It lacks the conceptual sensitivity to distinguish between the cases that our moral intuitions clearly treat differently.

This is precisely the gap that my own discounting view is designed to fill. By treating liability as a scalar reduction in the moral disvalue of harm, it allows us to register not only the kind of threat posed, but also the moral character of the threatener, and the contextual appropriateness of different levels of defensive response. In that sense, discounting offers not only an alternative to enforceable duty—it offers an explanation of what the enforceable duty theory tries, but ultimately fails, to capture.

5. Discounting Moral Disvalue of Harm

In this section, I develop and defend my own account of liability—one that departs fundamentally from the rights-based and duty-based models discussed earlier. According to the view I propose, liability should be understood as a matter of discounting the moral disvalue of harming someone. In other words, to say that a person is liable to defensive harm is not to say that they have lost or forfeited a right, nor to say that they are under a moral duty to accept being harmed. Rather, it is to say that the harm they suffer, while still real, matters less morally in the overall evaluation of whether it is permissible to impose it. On this account, liability is not primarily a relational status between individuals defined in terms of claims and duties; instead, it is a moral function of how we ought to weigh and distribute harms under conditions of conflict.

5.1 The Structure and Foundation of the Discounting Account

To make my idea of the discounting account more precise, it is helpful to begin by distinguishing two dimensions of harm: its physical magnitude and its moral disvalue. The physical magnitude of harm refers to the concrete, measurable effects of an action—such as the difference between a bruised arm, a broken leg, or a lost life. It is uncontroversial that these physical harms can be ordered by severity. But there is also a second dimension, which I call moral disvalue, which refers to how much a given harm counts morally—how much weight it has when we deliberate about what we may or may not do to others. The same physical harm can vary widely in its moral disvalue depending on how it comes about. For instance, a broken leg caused by an earthquake, by a risky sport someone chose to undertake, or by an unprovoked assault each represents the same physical injury but very different moral situations. Our moral intuitions respond not just to how much harm was caused, but to how that harm was situated in a broader moral narrative of agency, intention, responsibility, and justice.

This insight lies at the heart of the discounting account. Defensive harm is morally permissible, I argue, not because the person harmed has lost their rights or because they are duty-bound to absorb it, but because the harm done to them counts for less—it carries less moral disvalue—than it ordinarily would. Conversely, if the person posing the threat were innocent, or only minimally responsible, the harm done to them would retain more of its moral weight, and harming them would be much harder to justify, even if it served some protective aim.

Let me illustrate how this framework works in practice through a simple case. Imagine that a culpable aggressor is about to stab a victim, and if the victim does not intervene, she will suffer what we might call ten units of physical harm. To prevent this, the only available option is to inflict fifteen units of physical harm on the aggressor. At the level of raw physical comparison, this seems impermissible: the harm inflicted (15) exceeds the harm prevented (10). However, under the discounting account, what matters is not the raw numbers but the moral weight of those harms. Because the aggressor is culpably attempting to wrongfully harm another person, the harm that would befall him if stopped carries far less moral disvalue than the same harm would carry in the case of an innocent person. Perhaps the fifteen units of harm he would suffer have, due to his culpability, a moral disvalue equivalent to only one or two units. Meanwhile, the ten units the victim would suffer retain their full moral disvalue, or maybe even greater disvalue due to the fact that it is a culpable attack that undermines her autonomy. When assessed in terms of moral disvalue rather than physical damage alone, it becomes clear that harming the aggressor is justified: we are preventing a greater moral disvalue at the cost of a lesser one.

This shift in moral perspective enables the discounting account to justify defensive harm without appealing to the machinery of rights—whether retained, forfeited, or overridden. It also avoids relying on the notion of enforceable duties, which, as we have seen, raises problems. Most importantly, it avoids the two key criticisms that plague the forfeiture account as articulated by Renzo: the incompleteness problem, since we do not need to posit any loss of moral status in order to explain permissibility; and the redundancy problem, because we do not repackage deeper moral reasons in the language of rights loss—we simply go directly to those reasons by analyzing the comparative moral weight of harm.

I believe the theoretical foundation of the discounting account is straightforward. Self-defence, at bottom, is a problem of *allocating and reallocating harm/cost* under non-ideal conditions. Setting aside complicating factors, such as responsibility, culpability, agency, and agent-centred restrictions, our most basic, intuitive judgement is simple. We believe it is morally right to *minimise total harm*. A second intuitive judgement follows when total harm cannot be reduced. We believe that, if the overall cost is fixed, *distribute it as evenly as possible* across all agents. Real cases of self-defence are precisely those in which these ideal situiations cannot both be satisfied, because responsibility, culpability, agency, and side-constraints matter. In such cases we must make an all-things-considered distributive judgement. If we cannot minimise *harm* itself, we should minimise *the moral disvalue of harm*. Similarly, if we should not distribute harm *equally*, we should identify, among all the feasible options, the *fairest distribution* of harm once all the morally relevant factors are taken into account. This is exactly what the discounting framework is for. By distinguishing physical magnitude from moral disvalue, and by discounting the latter in light of factors such as responsibility, culpability, causation, and agent-centred limits, the view tells us how to allocate unavoidable costs *fairly* in non-ideal conditions. It preserves the two baseline intuitions (minimise; equalise) as defeasible defaults, then systematically adjusts them to reality: minimise *discounted* moral cost rather than physical harm, and distribute remaining costs *proportionally* rather than uniformly when agents are differently situated. In this way, discounting is the normative machinery that lets us carry intuitive distributive fairness into the hard cases of self-defence. I will return to this theme in later chapters and discuss it in greater depth.

5.2 How the Discounting Account Avoids Problems Faced by Other Accounts

Having set out my discounting account, I now demonstrate how the discounting account avoids those problems faced by other accounts as discussed above. The key claim is that the discounting account does not treat liability as a mere label. Instead, it provides a unified reason why defensive harm can be justified and how such justifications works across cases.

First of all, the discounting account does not face Renzo's incompleteness problem and redundancy problem. Let us first see why the discounting account is complete rather than merely labelling. Rights forfeiture and rights enforcement state a result in the language of status—"the right is lost," or "the claim is inapplicable"—and then look for something else to justify that result. By contrast, discounting builds the justification into the decision procedure itself. The explanatory core is that responsibility for an unjust threat *weakens the moral weight* of the aggressor’s prospective harm. The same physical harm to a culpable attacker counts for *less* in moral consideration than it would to an innocent bystander. That is the mechanism by which we move from mutual protection to permission: as responsibility increases, the *moral disvalue* of harm to the threatener decreases, and the balance of reasons shifts toward defensive action. The “normative transition” is thus grounded directly in a responsibility-sensitive assessment of moral disvalue. Nothing further has to be added to explain why permissibility is obtained.

Also, the discounting account does not become redundant when deeper principles appear. For rights-based theories, once one appeals to culpability, foreseeability, and fairness to justify permissibility, the residual talk of forfeiture or enforcement does no work. On the discounting view, those very considerations are not external add-ons; they are the *inputs* that determine the size of the discount. Culpability, foreseeability, voluntariness, and opportunities to avoid wrongdoing calibrate how much the aggressor’s harm is down-weighted; that calibration, together with the gravity of the threatened harm, *is* the justificatory work. There is no second, surplus layer of vocabulary to repeat the conclusion.

Furthermore, the discounting view can adequately accommodate scalar liability and the requirement of proportionality. Because discounting admits of degrees, permissions emerge in a *smooth* way that tracks morally salient continua. As the aggressor’s culpability grows (intentional > reckless > negligent), as causal contribution rises, or as the threatened harm becomes more severe, the discount increases and the range of permissible defensive harm expands accordingly. Conversely, where the threat is minor or responsibility is low, the discount is small and permissions remain narrow. This directly explains common cases. For example, this explains why lethal force is permissible against a culpable lethal attack but not permissible against a culpable threat to cause a minor injury. The discounting view can also explain why one might impose greater defensive force against a culpable threat than a non-culpable threat of the same magnitude.

Finally, the discounting account offers an satisfying explanation of the necessity requirement. If two defensive options are equally effective at averting the threat, we ought to choose the one that is less harmful. This is because, according to the discounting account, one need to minimize total moral disvalue. Since discounted harm still counts as harm and contains some moral disvalue, the least harmful option (given it is equally effective) would be morally preferable. Hence, if both a taser and a gun would stop the attack with comparable reliability, the taser is required because it lowers the total moral disvalue, not because of an external, free-standing minimisation norm.

In sum, the discounting framework explains *why* defensive harming can be permissible (completeness), avoids becoming a mere rhetorical overlay (non-redundancy), explains scalar liability, proportionality requirement, as well as the requirement of necessity. It therefore provides the unified explanation that other accounts lack.

5.3 Some Worries and My Responses to Them

One might worry, however, that this model lacks the categorical strength of rights-based accounts. In paradigmatic self-defence scenarios, such as a person fighting off a violent attacker, rights theorists can say: “The aggressor has no right not to be harmed, so harming them is plainly permissible.” Can the discounting account offer an equally decisive judgment? I believe that it can. In such clear-cut cases—where the aggressor is fully culpable, the threat is serious, and the victim’s actions are necessary and proportionate—the moral disvalue of harming the aggressor is so substantially diminished that it becomes morally negligible. Even if the physical harm inflicted exceeds the harm averted, the moral calculus still decisively favors the defender. In these cases, the discounting account not only supports the permissibility of defensive harm—it does so with equal confidence and clarity as rights-based accounts, but without their theoretical costs.

Moreover, the discounting model is better equipped to handle complex and borderline cases. As discussed above, where rights-based theories often falter due to their binary structure(either someone has a right, or they do not), the discounting account is intrinsically scalar. The degree of discounting can be fine-tuned in accordance with factors like the aggressor's level of culpability and responsibility, their causal contribution to the threat, whether they assumed certain risks voluntarily, and whether they had fair opportunities to avoid the threatening conduct. This flexibility makes the theory far more capable of capturing the moral nuances that pervade real-world defensive situations. Later chapters will explore these complex cases in greater depth, showing how the discounting account can generate nuanced and principled judgments where other theories remain stuck.

Another potential challenge for the discounting account concerns the question of compensation. Other theories seem to have a straightforward answer: if someone's rights were not violated, or if they were merely fulfilling a duty, then they are not wronged, and thus are not owed compensation. Can the discounting account make sense of the fact that culpable aggressors, even when harmed permissibly, are not entitled to compensation from their victims?

To answer this, we must draw a distinction. First, there are indeed cases where someone who is permissibly harmed might still be owed compensation—something that rights-based and duty-based theories struggle to accommodate. Chapter 3 will explore such cases in detail. Second, in clear-cut cases like that of a culpable aggressor, the discounting account can explain why compensation is not owed. If the aggressor's harm has been heavily discounted, then requiring the victim to compensate them would mean imposing undiscounted harm (on the victim) for the sake of discounted benefit (to the aggressor). This would result in a net increase in the total moral disvalue in the world. Compensation would not restore justice; it would worsen the moral situation. For practical purposes, we can say that the aggressor was not wronged—not because they lacked a right, but because the harm they suffered lacked the kind of moral significance that grounds a claim to be made whole. The notion of wrongfulness here does not depend on rights violation, but on the moral weight of what was done.

In summary, the discounting theory offers a conceptually clear, normatively grounded, and intuitively powerful framework for understanding liability to defensive harm. It allows us to make confident judgments in paradigmatic cases, navigate difficult ones with nuance, and avoid the pitfalls of rights-based and duty-based models. Most importantly, it recasts liability in a way that reflects the complexity of real moral life—not in terms of binary permissions or abstract entitlements, but in terms of the deeper question of how to fairly allocate the moral cost of harm.

6. Conclusion

In this chapter, I have examined the question of the nature of liability in the ethics of defensive harm. I began by considering the two most influential families of accounts: rights-based theories, including rights forfeiture and rights enforcement, and duty-based theories, most notably the enforceable duty account. I argued that none of these approaches can provide a satisfactory explanation of liability. Rights-based theories fail because their binary structure cannot capture the scalar character of liability, and because they cannot adequately account for proportionality or the necessity requirement. Duty-based theories fare no better, as they impose implausible moral obligations on potential victims and also struggle to explain the scalar nature of liability.

I then developed the discounting account. According to this view, liability is best understood not as the forfeiture of rights or the assumption of duties, but as a matter of discounting the moral disvalue of harm. I showed how this account avoids the incompleteness and redundancy problems that undermine rights-based theories, and how it naturally explains proportionality, scalar liability, and necessity without resorting to *ad hoc* stipulations. I also addressed potential concerns for the discounting view. I argued that the account can respond to these challenges while retaining its theoretical clarity and normative power.

Taken together, these arguments establish the discounting view as a more compelling and resilient theory of liability, one that better captures our intuitions about defensive harm and provides a deeper foundation for analysing the complex moral questions explored in the chapters that follow.

Chapter 2: The Grounds of Liability

1. Introduction

In the previous chapter I examined the *nature of liability*, asking what liability to defensive harm *is*. In this chapter my focus shifts from the *nature* of liability to its *grounds*. It concerns under what conditions a person becomes liable in the first place. Although these questions are distinct, they are complementary. To defend a distributive conception of defensive harm, I must not only clarify what liability amounts to but also determine the factors that generate it.

In section 2, I critically examine several influential accounts of the grounds of liability. These include the culpability account, the causal account, the moral-status account, and the enforceable duty account. I argue that each account fails to explain some intuitive cases. They are either too restrictive in some cases, too permissive in other cases, or misdescribe the interpersonal structure of duties and claims. In section 3, I set out the moral responsibility account and develop my own objections to its standard formulation, highlighting its fragile reliance on slight differences in foreseeability and control across structurally similar cases. In section 4, I offer a revision of the moral responsibility account. I use a bystander baseline and three variables—(i) voluntary initiation or maintenance of a risk-imposing activity, (ii) evidence-relative awareness of a non-trivial risk to others, and (iii) causal proximity. I propose a graded model on which degrees of liability track degrees of moral responsibility without sharp thresholds, and I indicate how this account has stronger explanatory power in a wide range of cases.

2. Various Accounts of Liability

The ethics of self-defence requires an account of when it is morally permissible to impose harm on another person. Central to this issue is the concept of liability: under what conditions does an individual forfeit their moral claim not to be harmed? Philosophers have proposed various theories to explain the conditions under which such forfeiture occurs. Among the most influential are the causal account, the culpability account, and the moral status account. Each of these theories attempts to offer principled criteria that track our intuitions about permissible self-defence. However, as I aim to show, each faces serious challenges, particularly in cases where intuitive judgments diverge.

2.1 The Culpability Account

On the culpability account, a person is liable to defensive harm only if she is culpable for posing an unjust threat. To be culpable, in this sense, is to be blameworthy or otherwise appropriately subject to moral criticism for conduct that wrongfully endangers others (Ferzan 2005: 733–739; Ferzan 2012: 669–697). The attraction of the view is straightforward. If liability marks the circumstances under which harming a threatener is permissible in defence, it is natural to connect that status to moral fault: those who intentionally, recklessly, or negligently endanger others seem precisely the sort of agents against whom defensive harm is licensed, whereas the faultless are not. The culpability account also resonates with familiar patterns in criminal and tort law, where culpability underwrites punishment or damages; that analogy helps explain why many find it normatively apt to tie defensive permissions to blameworthiness (Ferzan 2005: 735).

> Murder: An attacker maliciously intends to kill a victim, and the victim can save herself only by using lethal force.

In this paradigm, the culpability account yields the intuitive verdict. The attacker culpably threatens the victim’s life; his wrongful intention grounds his liability to defensive harm, so the victim may permissibly kill him in self-defence. I take this to be both intuitively compelling and theoretically elegant: the ground of liability—culpable wrongdoing—aligns with our ordinary reasons for permitting defensive force.

Once I move beyond these core cases, however, the culpability account appears overly restrictive. A central difficulty for the culpability account is already visible in a case first articulated by McMahan and later endorsed and emphasised by Quong: there is a scenario in which self-defence seems plainly permissible, yet the culpability account cannot explain that permissibility (McMahan 2005: 387; Quong 2020: 23–24).

> Mistaken Attacker: The identical twin brother of a notorious murderer is driving during a stormy night in a remote area when his car breaks down. Unaware that his brother has recently escaped from prison and is thought to be hiding in this same area, he knocks on the door of the nearest house, seeking to phone for help. On opening the door, the armed and frightened resident mistakes the harmless twin for the murderer and lunges at him with a knife. (McMahan 2005: 387).

On Quong's analysis, the resident's attack is evidence-relative permissible and therefore non-culpable. According to the culpability account, the resident must be non-liable to any defensive harm. Quong believes this is counterintuitive. The innocent twin seems entitled to defend himself, even lethally if necessary, as he is the victim of an intentional and objectively wrongful attack. (Quong 2020: 23–24). I agree with Quong's intuitive judgement and diagnosis. The case shows that the permissibility of self-defence does not always depend on the threatener's culpability: liability can arise even when the threatener acts under a fully justifying epistemic position, and the culpability account, by construction, cannot recognise that fact.

The same structure, as Quong stresses, recurs outside epistemic error in cases of coercion: even when the threatener is blameless, defensive force still seems permissible.

> Coerced Assassin: Terrorists have captured Albert's child. They truthfully tell Albert that he must assassinate ten innocent government officials or else the terrorists will torture and murder Albert's child. Albert would normally never wrongly hurt

> anyone, but to protect his child he proceeds with the assassination attempt. (Quong 2020: 24).

Again, Albert is excused from blame, but he wrongfully threatens an innocent life. Intuitively, the officials may defend themselves by killing Albert. However, the culpability account must deny this and maintain that Albert is not liable to any defensive harm since he is not culpable. As with Mistaken Attacker, I agree with Quong that the better explanation of the permissibility of defence is that Albert is liable despite the absence of culpability (Quong 2020: 24).

In my view, the force of these two cases does not depend on any contested theoretical machinery. The intuitive verdicts are exceptionally strong because in each the aggressor poses an intentional and objectively wrongful threat to an innocent person—justified mistake in the one case and coercion in the other do not alter that objective status. We therefore have a powerful intuition that the victim may defend herself and that the aggressor is liable to defensive harm. The culpability account, by design, cannot recognise such cases. For that reason, I take Mistaken Attacker and Coerced Assassin to be decisive objections to the sufficiency condition posited by the culpability account.

A further criticism develops the restrictiveness worry by appealing to cases in which the threatener is faultless, or not even acting, yet many philosophers judge lethal self-defence permissible.

> Conscientious Driver: A person keeps his car well maintained and always drives cautiously and alertly. On one occasion, however, freak circumstances cause the car to go out of control. It has veered in the direction of a pedestrian whom it will kill unless she blows it up by using one of the explosive devices with which pedestrians in philosophical examples are typically equipped (McMahan 2005:

> 393).
>
> Falling Person: A gust of wind blows a person down a well where the victim is trapped; unless he vaporises her with a ray gun, she will crush and kill him; if he does not vaporise her, she will survive the fall (Nozick 1974: 34).

Many philosophers take the intuitive verdict in both cases to be that the threat is *liable to be killed*—for example, McMahan defends this verdict about the driver case, and Thomson explicitly treats the falling-person case as one in which defensive killing is permissible (McMahan 2005: 393–394; Thomson 1991: 290). Read in this way, the cases sharpen the criticism of the culpability account: the intuitive judgement is that lethal defence is permissible because the threat is liable, yet the culpability account cannot explain that judgement, since neither the careful driver nor the wind-blown person is culpable.

My own view is more cautious. I do not accept that the threats in these cases are straightforwardly *liable to be killed*—I set that question aside for the next chapter. What I do accept is that each threatener bears *some degree of liability*, and in later sections of this chapter I develop my own account to explain how that liability arises. Read in this way, the cases still undermine the culpability account: even if lethal measures are not always permitted, there is at least a clear permission to impose non-lethal defensive costs on the driver and the falling person and a corresponding asymmetry in counter-defence (they may not permissibly resist proportionate measures aimed at averting the threat). Those are hallmark features of liability. Because the culpability account ties liability exclusively to blame, it predicts *no liability at all* in these scenarios; that prediction is implausible. Hence these cases remain a genuine challenge to the culpability account even if I ultimately deny that they establish liability to killing.

Defenders of the culpability account can, of course, resist by biting the bullet. One strategy is simply to deny the permissibility of defence in the excused-aggressor cases: perhaps the innocent twin may not kill the resident, and the official may not kill Albert; perhaps our intuitive judgements are unreliable (Ferzan 2012: 672–674). A more moderate strategy accepts that defence is permissible in those cases but treats that permissibility as grounded solely in lesser-evil or agent-centred considerations, rather than in liability. On both strategies the conclusion is that liability still requires culpability, and that what looks like liability without blame is either impermissible or permissibility of a different, non-liability kind. I do not find these replies convincing. The full denial carries an implausibly high cost in intuitive judgments, requiring us to say that innocents must submit to being killed by justified but mistaken or coerced attackers. The lesser-evil strategy is also quite implausible since it struggles to accommodate two features that the language of liability captures more naturally: first, that the defender does not wrong the threatener by using force, and second, that counter-defence by the threatener seems impermissible, as if he lacks the standing to resist.

In sum, the culpability account fails to accommodate an important range of cases in which defence is permissible against excused threateners, and it is, at minimum, under-inclusive in cases where faultless threateners impose immediate lethal danger. I therefore believe that culpability, while morally significant, is not the sole ground of liability to defensive harm. Any adequate account of liability must be able to explain why defensive harm can be permissible even when blame is absent.

2.2 The Causal Account

On the causal account, an individual is liable to defensive harm in virtue of being (about to be) the causal source of an unjust outcome, irrespective of culpability. The view is most prominently associated with Judith Jarvis Thomson’s discussion of self-defence, and is often introduced as a way to explain permissions to use defensive force against both culpable and non-culpable threateners (Thomson 1991). I understand the core mechanism of the causal account as rights-based: if, unless Y uses defensive force against X, X will kill Y, and X has no right to kill Y (i.e., X has a duty not to kill Y), then X is liable to defensive harm. Liability, on this view, tracks this causal posture with respect to people’s rights. Now recall the two cases mentioned above.

> Murder: An attacker maliciously intends to kill a victim, and the victim can save herself only by using lethal force.

Here the causal account explains permissibility by appeal to the attacker’s causal position with respect to the victim’s life. Because the attacker will kill the victim unless the victim uses lethal force, and the attacker has no right to do so, he is liable to the defensive harm necessary to avert the killing. The fact that the attacker is culpable helps to make the case vivid, but, according to the causal account, the causal structure itself is sufficient on its own to ground liability.

> Falling Person: A gust of wind blows a person down a well where the victim is trapped; unless he vaporises her with a ray gun, she will crush and kill him; if he does not vaporise her, she will survive the fall (Nozick 1974: 34).

In this non-culpable case, the same mechanism yields the same verdict. Even though the falling person bears no moral blame and is not even acting voluntarily, she will otherwise kill Victim; since she has no claim to do so, she is liable to defensive harm. The appeal of the causal account lies precisely in this unification: it purports to explain why defensive force can be permissible against both deliberate attackers and innocent threateners without changing its underlying rationale.

Having set out the causal account and its rationale, I now turn to the major objections to the causal account. Michael Otsuka’s critique is the most direct. He grants it is perfectly natural to say that a wind-blown, unconscious human will kill you if she lands on you, just as a wind-blown stone will; what he denies is that it makes sense to add that thereby *violates* your right not to be killed, since neither exercises agency. He argues that the talk of rights-violations goes too far if it implies that a falling stone can violate someone's right (Otsuka 1994: 80–82). Otsuka reinforces the point with a familiar analogy: when an assassin kills with a bullet, it is the assassin—not the bullet—that violates your right (Otsuka 1994: 81–82).

A possible response is to restrict Thomson's picture so that only *moral agents* can be rights-violators, and to relocate any rights-violation in the responsible villain when an innocent body is used as a means. Otsuka anticipates this move and argues that it does not salvage the key claim in the falling-person cases. Even the weaker proposal—that the falling person merely *causes* a rights-violation—is, he contends, misguided. If a villain pushes someone so she will land on you, the rights-violation (if any) is attributable to the villain rather than to the falling person, since the difference tracks the villain's agency, not the falling person's status (Otsuka 1994: 82).

Jeff McMahan sharpens this underlying worry. On his reconstruction of Thomson, to cover non-culpable threats the view must allow that one can "violate a right" *without culpability and even without agency*. But, as McMahan points out, on Thomson's own understanding, having a right consists in others being morally constrained in certain ways—constraints that apply only to responsible agents. Falling stones are not subjects of moral constraints. Similarly, at least some non-responsible attackers cannot be rights-violators. If so, Thomson's view either over-generates (up to stones and bullets) or loses coverage in the very cases of innocent threat for which it was designed (McMahan 1994: 276–77).

My own view is that Otsuka and McMahan are right to press this line of criticism. On Thomson's construction, it is indeed difficult to explain why a wind-blown person and a wind-blown object should be treated differently with respect to liability. That said, I do think there is a morally significant difference between the falling person and the falling stone. But that difference cannot be captured by Thomson’s causal account, nor is it adequately explained by appealing merely to the category of “moral agent.” Rather, the distinction turns on considerations of moral responsibility of the kind that Otsuka and McMahan themselves foreground. I will develop this responsibility-based explanation in later sections. For present purposes, the important point is that Thomson’s causal account does not explain the difference.

Another way of objection to the causal account is offered by Helen Frowe. Consider the following two cases:

> Bridge: Victim is fleeing Murderer, who wants to kill him. Victim’s only escape route is across a narrow bridge that has space for only one person at a time. Pedestrian is out for a walk on the bridge, and cannot get off in time to allow Victim to escape Murderer. Only by knocking Pedestrian off the bridge, killing her, can Victim cross the bridge in time to save his own life (Frowe 2014: 24).

> Malicious Bridge: Everything else is the same, except that Pedestrian could easily get off the bridge in time to allow Victim to escape Murderer, but she dislikes Victim and decides to stay where she is, realizing that this will impede Victim’s escape (Frowe 2014: 25).

Frowe’s point is that, on a causal account, these two cases are the same with respect to the discussion of liability. Both obstructors have the same causal contribution to the threat the victims are facing (i.e., both of them only block the escape route). In each case, the victim will die unless the obstructor is harmed. If causal contribution only—rather than agency or culpability—grounds liability, the view appears to permit the harming of both or neither. However, Thomson wants to allow harming the malicious obstructor while forbidding harm to the innocent one. The causal account, taken on its own terms, cannot mark this difference, and thus fails to explain the intuitive asymmetry (Frowe 2014: 25-26).

I agree with Frowe’s analysis here. I believe the intuitive difference between the two cases is obvious, and the causal account cannot explain this difference. Furthermore, I believe the deeper explanation is that, without reference to moral agency or culpability, we cannot fully explain Pedestrian’s position in the respective scenarios. In Bridge, Pedestrian’s body merely and inadvertently blocks Victim’s escape route; she neither knows nor intends Victim’s death, and her presence is accidental. By contrast, in Malicious Bridge, Pedestrian stands her ground with full awareness that doing so traps Victim and will foreseeably lead to Victim’s death; in that sense, she participates in the killing—she helps to kill by intentionally maintaining the lethal threat.

Causally, these actions are indistinguishable: in both cases Victim will die unless Pedestrian is removed. Morally, however, the actions differ in kind, because the second involves a responsible agential contribution—knowledge and culpable purpose—to the lethal threat. That is precisely why the intuitive asymmetry is so strong. And that asymmetry cannot be captured by a view that fixes liability purely by causal role. It requires recourse to moral responsibility and culpability to discriminate between mere causal presence and responsible participation. A causal account, as such, lacks the resources to draw that distinction.

To sum up, the causal account faces serious difficulties. It cannot explain the liability of some “so-called” non-responsible threats, and it cannot capture important intuitive difference in cases of malicious and innocent obstructors. These problems suggest that causal role alone is insufficient to ground liability. What emerges from this analysis is that moral responsibility and agency are indispensable: without reference to them, we cannot make sense of our intuitive judgments about defensive harm.

2.3 The Moral Status Account

Having seen why a purely causal approach struggles to track our intuitions in paradigm cases, I turn to Jonathan Quong's moral status account. On this view, liability to defensive harm arises when, and only when, the evidence-relative permissibility of one's act depends on assuming that someone else lacks a moral right that she in fact possesses, so that one's act threatens (or reasonably appears to threaten) that person's rights. The core thought is rights-based and relational: to treat others as if they lack rights is to fail to accord them the concern and respect they are due, and doing so can make one liable even absent culpability (Quong 2020: Ch.2). Quong's framework is also explicitly contrastive: some actions are evidence-relative permissible because "the correct moral theory" deems a practice justified despite risk (and so do not diminish anyone's status), whereas other actions are permissible only because the agent's evidence makes it look as if the putative victim has lost standing to press claims (and so do diminish status); the latter—status-diminishing—acts ground liability. Now recall the Mistaken Attacker case:

> Mistaken Attacker: The identical twin brother of a notorious murderer is driving during a stormy night in a remote area when his car breaks down. Unaware that his brother has recently escaped from prison and is thought to be hiding in this same area, he knocks on the door of the nearest house, seeking to phone for help. On opening the door, the armed and frightened resident mistakes the harmless twin for the murderer and lunges at him with a knife. (McMahan 2005: 387).

On Quong's account, the resident is liable because his evidence-relative permission to attack depends on treating the visitor as if he has lost the usual protection against being harmed—i.e., as if he were liable—whereas in fact the visitor retains that protection; the resident therefore wrongfully (though blamelessly) lowers the visitor's moral status and so may be defensively harmed. On the other hand, recall the Conscientious Driver case:

> Conscientious Driver: A person keeps his car well maintained and always drives cautiously and alertly. On one occasion, however, freak circumstances cause the car to go out of control. It has veered in the direction of a pedestrian whom it will kill unless she blows it up by using one of the explosive devices with which pedestrians in philosophical examples are typically equipped (McMahan 2005: 393).

Here the verdict is the opposite. Because the practice of prudent driving is, according to the best moral theory, justified to all in light of its benefits, incidence of risk, and incidental nature of any harms, the pedestrian lacks a right that this driver refrain from driving; the driver's action does not treat anyone as having diminished status and so does not ground liability. For completeness, Quong adds that self-preserving force against non-liable persons may sometimes be agent-relatively permissible (provided they are not used as a means), in which event a non-liable driver might also permissibly counter-defend—but that claim concerns permissions beyond liability and is not essential here.

Having set out Quong's moral status account and his contrasting verdicts in the two canonical scenarios, I now turn to Helen Frowe's challenge. Her target is Quong's claim that, because the practice of prudent driving is evidence-relatively permissible, a careful driver who faultlessly threatens a pedestrian is not liable to defensive harm. Frowe argues, first, that practice-level cost–benefit cannot justify imposing risk in a particular interpersonal encounter, and, second, that even when a risky practice is permissible, the prospective victim retains a reasonable claim that the risk-imposer internalise the costs rather than externalising them onto her (Frowe 2022: 511–14, 520–24).

Frowe accepts the set-up of Conscientious Driver but rejects the route from the permissibility of the practice to the non-liability of the driver. Her *type–token* argument is that the permissibility of an activity-type (e.g., prudent driving) cannot, by itself, underwrite the permissibility of *this* token act of imposing a serious risk on *this* person here and now. What justifies imposing the incidental risk must be the prospective benefits attached to the very token act in question; those benefits cannot be “borrowed” from the aggregate payoffs of the practice at large (Frowe 2022: 514–16).

For instance, someone flips a light switch, unknowingly imposing a tiny risk of explosion that would badly burn a nearby person. Whether the flip is permissible must be assessed by the benefits of *this* flip (e.g., illuminating this room now), not by the social benefits of light-switch-flipping *in general* (Frowe 2022: 514–15).

Now see another example offered by Frowe. A technician flips a switch at a small risk to a bystander because doing so enables life-saving surgery. That token act may be justified by its own benefits; but those benefits do not transfer to justify someone else’s trivial light-flipping when those impose equivalent risks on others (Frowe 2022: 516).

From these cases, Frowe extracts a general parity claim: when two actions incidentally impose *equal risks for equal benefits*, they are morally on a par. What matters is the token pairing of risk and benefit, not the label of the wider practice (Frowe 2022: 516–17). In Conscientious Driver, this pushes against the idea that the practice-level value of prudent driving can nullify the pedestrian’s claim in the token confrontation: the driver’s reasons for *this* incident of driving (say, a movie outing) do not automatically license shifting a grave token risk onto *this* pedestrian when the fault manifests (Frowe 2022: 514–16).

Frowe’s second line is the *internalisation model*. Even if people cannot reasonably demand that others never engage in risk-imposing practices, they can reasonably demand that risk-imposers *internalise* at least the foreseeable costs of their risky actions. One may lack a right that others refrain from driving altogether and yet retain a right not to be harmed by that very driving; when bad luck strikes, the pedestrian may permissibly act to prevent the driver from externalising the costs of his self-interested (though usually permissible) activity onto her (Frowe 2022: 520–23). In Frowe's terms: the fact that the practice is permissible does not show that the pedestrian's right not to bear *this* cost has been extinguished; it remains reasonable for her to demand that the driver bear his own costs rather than forcing them upon her (Frowe 2022: 520–24).

Quong replies to Frowe on two fronts. First, he argues that a focus solely on token-level risk–benefit profiles ignores morally crucial "pattern facts." Consider Buzzed Driving: any single episode of mildly intoxicated driving presents a vanishing risk to any particular person and yields some enjoyment to the driver. If we appraise only that token, it may seem that no one can reasonably demand the driver refrain. But once we allow millions of such tokens each year, we know—ex ante—that thousands will be killed and many more injured. That near-certainty of large-scale harm, Quong argues, is a morally relevant fact for assessing the permissibility of each instance; any account that brackets such pattern facts, as he reads Frowe to do, is therefore inadequate (Quong 2022: 562–564). He makes a similar point by contrasting tokens that share the same local risk–benefit ratio but belong to very different activity-types—such as a scientist's one-off satellite launch versus widespread everyday acts—insisting that type-level information about predictable aggregate harms bears on the permissibility of each token (Quong 2022: 563–564).

Second, he resists the internalisation model. First, he insists that where a person lacks a claim that another refrain from Φ-ing, the latter neither violates her claim if harm ensues nor becomes *uniquely* liable to compensate her; any standing claim to be made whole should be addressed by a broader scheme of distributive justice that spreads costs across the population (e.g., taxation) (Quong 2022: 563–564). Second, he argues that such a general compensation scheme is a *more equitable and reliable* way to secure distributive fairness than making non-negligent risk-imposers strictly internalise the costs of their tokens; strict internalisation would treat identical actors differently and is not a dependable basis for permissions (Quong 2022: 564).

I do not think these replies solve Frowe's core challenge. I will start with pattern facts. They are plainly relevant to policy and regulation, but the question in Conscientious Driver is an interpersonal one: whether, here and now, this driver has justification to impose this non-trivial risk on this pedestrian. Benefits that accrue from other people's tokens—or from the practice considered in the aggregate—cannot be "borrowed" to justify the imposition in this dyadic confrontation. Quong's Buzzed Driving shows that pattern facts matter when we decide whether to permit a practice; it does not show that they supply the justification the driver needs in the moment of conflict. If liability is supposed to track how A treats B, then the justification for imposing the lethal risk on B must be found in the reasons attached to A's token act toward B, not in the moral arithmetic of countless third-party acts.

Also, I believe Quong's two clarifications do not unsettle the internalisation claim. First, Quong's suggestion that general taxation can handle fairness does not answer the ex ante point. Defensive and compensatory liabilities can be complementary mechanisms for the same underlying idea—making risk-imposers bear their own costs; victims may reasonably prefer defence to ex post compensation precisely because compensation often cannot restore or deliver the outcome they have reason to prefer (e.g., life, bodily integrity). Second, the appeal to the greater "equity" and "reliability" of population-wide schemes misconceives what internalisation does at the moment of conflict. Internalisation here does not function as a wide programme of distributive justice; it is a local *cost-shifting*. It does not treat identical actors differently; it treats different *token relations* differently—specifically, the relation between the risk-imposer and the person he would otherwise force to pay the ultimate price. Whether a tax fund works well in general is irrelevant to whether this driver may permissibly externalise his costs onto this pedestrian.

To sum up, Quong’s moral status account offers an elegant way of distinguishing Mistaken Attacker from Conscientious Driver, but Frowe’s objections show that its non-liability verdict in the latter case is untenable. Token-level justification and the requirement of internalising costs preserve the pedestrian’s right not to be harmed by this act of driving. Quong’s appeals to pattern facts and to general distributive schemes fail to address that interpersonal demand. Thus the moral status account, at least in its present form, cannot adequately capture liability in cases of faultless but harmful risk-imposition.

2.4 The Enforceable Duty Account

On Victor Tadros’s view, liability to defensive harm arises when a person stands under an enforceable duty to bear some cost for the sake of protecting others; when such a duty obtains, others may permissibly enforce it, and doing so does not wrong the person on whom the duty falls (Tadros 2016: 116–18, 123–26). I understand the structure as stepwise: first, establish that X has a duty to promote an important end even at some cost; second, determine that this duty is enforceable; third, identify Y’s standing to enforce it; when all three are satisfied, X is liable to be forced—even if forcing X harms her (Tadros 2016: 116–18). Duties, on this account, can arise from different sources—culpable wrongdoing, causal involvement, “easy rescue” obligations when one can avert grave harm at modest cost, or fair-procedure burdens—and the breadth of these sources underwrites a correspondingly broad range of liability judgments (Tadros 2016: 119–21).

> Unread Letter: Veronica and Wilma each wish to kill Dan. Independently, each sends a letter to Kev, a hit man. Each letter instructs Kev to kill Dan with a pistol at noon. Kev receives Veronica’s letter. Wilma’s letter gets lost in the mail. Kev immediately acts on Veronica’s instructions, finds Dan, and attempts to kill him at

> noon. Veronica, Wilma, and Irene, an innocent, are standing by. Had Dan received Wilma's letter rather than Veronica's, he would have acted in exactly the same way (Tadros 2016: 119).

Tadros uses this case to compare the liabilities of Veronica, Wilma, and Irene. Veronica and Wilma are "each highly culpable" and perform "identical acts with the same intended consequences," but only Veronica's letter actually causes Kev to mount the noon attack; Irene is neither culpable nor causally involved (Tadros 2016: 119). On Tadros's account, it is uncontroversial that Veronica is liable to be harmed to avert the threat she culpably causes. If necessary, Dan may force Veronica to disarm Kev—even at serious cost to her—and does not wrong her (Tadros 2016: 119).

The more controversial claims concern Irene and Wilma. Tadros argues that Irene—an innocent bystander—can be liable to modest harm when Dan can save his life only by forcing her to avert Kev's attack. The explanation is Irene's enforceable duty of easy rescue: given the gravity of Dan's stake (his life) and the triviality of Irene's cost, Dan's forcing Irene to disarm Kev does not wrong her; indeed, it would be incoherent to say that Irene both has an enforceable duty to rescue and a right not to be forced to rescue. Given (enforceable) duty, liability follows: Irene is liable to some degree of harm to save Dan (Tadros 2016: 120).

Furthermore, Wilma did not in fact contribute causally to Kev's attempt, since her letter was lost in the mail. Nevertheless, Tadros argues that Wilma is not thereby exempt from liability. The reason is her culpability: she intentionally tried to bring about Dan's death, and had her letter arrived rather than Veronica's, Kev would have acted in exactly the same way. Thus, even though Wilma's action did not cause the actual threat that Dan faces, she remains liable to be harmed because her culpable attempt grounds an enforceable duty not to endanger Dan's life. On the duty view, culpability by itself can generate liability, even when causal contribution is absent (Tadros 2016:

121–22).

The most controversial implications of Tadros's duty account emerge in Unread Letter, particularly in his treatment of Irene and Wilma. I agree with Tadros that Irene has a duty to rescue, and I even accept that such a duty can be enforceable. Where I part company with him is in describing Irene's position as one of liability. On any plausible conception, liability to harm entails that the person lacks a claim against that harm and ordinarily does not stand to be compensated if harmed. But intuitively, Irene does retain a prima facie claim not to be harmed—her claim is simply not strong enough to prevail against the more urgent claim of Dan's right to life. Moreover, she seems also to have a claim to compensation if she is harmed. This is the familiar structure of lesser-evil justifications: an innocent person may permissibly be harmed, but only subject to later redress. If Irene is forced to disarm Kev at some costs, and Dan's life is saved, our natural response is not to say "Irene had no claim," but rather to insist that Veronica, or Veronica together with Kev, should compensate Irene. On this picture, Irene's rescue duty is real, but it does not erase her claims. Thus, describing her situation as one of "liability" distorts the normative structure of the case. Tadros does not explain why we should think that Irene's duty entails the absence of claims, and I do not see any theoretical gain in construing it that way. It is more attractive to say that Irene's harm is permissibly imposed but compensable, rather than that she was liable.

The case of Wilma raises a different difficulty. Tadros maintains that Wilma is liable despite her letter not having reached Kev and thus not causally contributing to the attempt on Dan's life. Her liability is said to rest on her culpability alone. But this stretches the idea of liability too far. If culpability alone suffices, then why not say that any culpable wrongdoer—for example, a fugitive guilty of an unrelated crime—has a duty to bear defensive costs in order to save Dan? That seems highly implausible. Tadros has not explained why Wilma's culpability, as opposed to culpability more generally, has the right kind of directionality to ground liability in Dan's case. Perhaps the idea is that Wilma attempted to kill Dan, and that attempt is what matters. But this raises further puzzles: if Wilma merely drafted the letter and abandoned it, or wrote it but never posted it, or resolved firmly in her mind to kill Dan but never acted, would she still be liable? Intuitively, all of these variations leave her in the same position—culpable, certainly, but not causally involved in the threat. If so, then Tadros's claim that Wilma is liable is unstable. Either he must accept that all culpable agents are liable to bear costs in unrelated defensive scenarios, which is untenable, or he must admit that causation does matter after all, in which case Wilma is not liable.

In sum, the treatment of Irene and Wilma shows that the duty account faces serious difficulties. In Irene's case, Tadros collapses duties of rescue into liability, thereby denying claims that she intuitively retains. In Wilma's case, he detaches liability from causation in a way that threatens to make it over-inclusive, extending liability to anyone culpable of wrongdoing, regardless of their relation to the threat. Neither extension seems defensible.

In this section I have critically examined a range of influential approaches to grounding liability to defensive harm. Each encounters serious difficulties. None, I believe, provides a satisfactory foundation. In what follows, I turn to the moral responsibility account to see whether it fares any better.

3. The Moral Responsibility Account and Its Problems

According to the moral responsibility account, an individual is liable for creating an unjust threat only when she is morally responsible for bringing about that threat (McMahan 1994, 2005; Otsuka 1994, 2016). Michael Otsuka proposes that someone counts as morally responsible for a threat when three conditions are satisfied: (1) she is of sound mind, (2) she is exercising control over her behaviour, and (3) she possesses awareness of the danger inherent in her actions (Otsuka 2016: 52).

It is possible for someone to be morally responsible for generating a threat even if she is not culpable. Recall the Mistaken Attacker case:

> Mistaken Attacker: The identical twin brother of a notorious murderer is driving during a stormy night in a remote area when his car breaks down. Unaware that his brother has recently escaped from prison and is thought to be hiding in this same area, he knocks on the door of the nearest house, seeking to phone for help. On opening the door, the armed and frightened resident mistakes the harmless twin for the murderer and lunges at him with a knife. (McMahan 2005: 387).

Here is the moral responsibility account's explanation of this case. In this situation, the twin brother has done absolutely nothing that would lose his rights against being harmed. He presents no actual threat to the resident and has no intention of seeming menacing. Nevertheless, the evidence available to the resident suggests that the twin brother poses a severe danger. To the resident, the twin brother is indistinguishable from a genuine attacker, and thus, the resident is not culpable or blameworthy for issuing the threat. Still, it is plausible to claim that the resident is still morally responsible for producing an objectively unjust threat to the twin brother. Supporters of the moral responsibility account maintain that this makes the resident liable to defensive harm, thereby justifying the twin brother (or a third party) in taking lethal defensive measures against the resident.

A more contentious case is offered by Jeff McMahan, who contends that in the Conscientious Driver scenario, the driver is morally responsible, though not culpable, for threatening the victim. Recall the Conscientious Driver case. The moral responsibility account explains this case by holding that the driver understands that driving involves risk but chooses to proceed regardless. Though the act of driving is permissible based on the driver's evidence—since she has no reason to expect a loss of control—it turns out to be impermissible in fact, as this particular drive endangers the life of an innocent person. McMahan asserts that the driver’s awareness of the general danger associated with driving, combined with the actual wrongness of this instance, renders her liable to defensive harm (McMahan 2005: 394).

Advocates of the moral responsibility account frequently root these conclusions in a specific idea of distributive justice. According to this view, fairness requires that people absorb the costs of the risks they choose to impose on others (McMahan 2005; Otsuka 1994; Draper 2009, 2016; Gordon-Solmon 2018). On such an account, individuals like the driver and the resident are seen as participants in moral gambles. If these gambles result in harm to innocent parties, then justice favours assigning the cost to the person who initiated the gamble. This remains true even if the choice to take the gamble was justified by the agent’s available evidence. In such situations, the agent is said to forfeit her right not to be harmed in order to protect the potential victim. McMahan further argues that in cases where more than one agent is responsible for an unjust threat, it is equitable to impose the full burden of defensive harm on the most responsible agent—provided the burden cannot be divided. Hence, liability is treated as having a comparative aspect, where even minor disparities in responsibility may decisively affect who is liable (McMahan 2011: 551).

By contrast, recall the case of Falling Person. According to the moral responsibility account, unlike the resident or the driver, the falling person is not plausibly morally responsible for the threat she poses, because she is being involuntarily hurled down a well by a strong gust of wind. Although she is causally responsible for endangering Victim’s life, the moral responsibility account maintains that she is not liable to be lethally defended against. Michael Otsuka endorses this view, maintaining that the falling person is no more morally responsible for the danger facing Victim than a passive bystander would be. Therefore, her moral status aligns with that of an innocent bystander. Since harming bystanders is impermissible, it follows that harming the falling person is also impermissible (Otsuka 1994). Similarly, the moral responsibility account extends this restrictive judgment to agents who become threats due to delusions or temporary insanity. So long as these threateners are not morally responsible for their conduct, they preserve their ordinary rights not to be harmed—even if failing to harm them would result in the deaths of innocent others. Accordingly, the moral responsibility account endorses killing a narrower range of threateners than the causal account does, but a broader one than that allowed by the culpability account.

Let us now turn to my objections to the moral responsibility account. A particularly troubling problem can be demonstrated by the Cell Phone case:

> Cell Phone: Without Caller's knowledge, a terrorist has secretly modified Caller's phone such that the next call he makes will detonate a bomb and kill Victim (McMahan 2005: 397).

The moral responsibility account explains this case by holding that Caller is not liable to be killed because, in contrast to the driver in Conscientious Driver, Caller has no basis whatsoever to suspect that making a call would endanger anyone. Since Caller lacks any awareness that his action could be harmful, he is not morally responsible for the threat he poses. However, even though McMahan maintains that there is a categorical difference between the driver and Caller, the discrepancy seems more like a matter of degree than of kind (McMahan 2005: 397). It is certainly not beyond the realm of possibility that Caller's phone might be tampered with. The real difference is simply that, based on Caller's information, the probability of such tampering is negligible, whereas the driver has slightly stronger reason—though still a small one—to believe that driving might harm someone. Thus, McMahan's framework depends crucially on the assumption that minute differences in foreseeability are morally decisive. According to this reasoning, the driver is liable to be killed to protect Victim, but Caller is not.

My worry for the moral responsibility account is that it struggles to generate consistent and intuitively acceptable judgments across structurally similar cases. As I see it, my own intuitions about these cases—Conscientious Driver, Cell Phone, and Falling Person—are neutral. I am not committed to the verdicts themselves (e.g., that the driver is liable, Caller is not, and the falling person is not), but rather to the point that the moral responsibility account cannot reliably or coherently derive such judgments from its own principles.

Let us begin with the Cell Phone case. The moral responsibility account explains that Caller is not *morally responsible* for the threat he poses because, on his evidence, there is no reason to believe that using his phone could endanger anyone. According to this view, the threat is entirely unforeseeable from Caller's perspective, and thus, he remains outside the scope of *liability*. However, this assessment becomes increasingly problematic once I offer a slight variation of the circumstances. Imagine a variation of the case in which Caller had previously come across a credible news report warning of terrorists employing remote-detonation devices triggered via unsuspecting civilians' phones. Suppose further that this report described similar cases occurring within Caller's country or region, and that Caller not only read it but briefly considered its implications before dismissing the possibility. Under such conditions, it would seem that Caller had at least some minimal basis for appreciating a non-zero risk associated with making a phone call.

Now, recall the Conscientious Driver case. The driver, too, acts permissibly based on available evidence: she has no reason to think she will lose control of her car. Nonetheless, she knows in a general sense that driving always carries a certain risk of malfunction, unexpected obstacles, or human error. Her decision to drive is therefore made in full recognition of a low-probability but serious potential for harm. The modified Caller likewise acts under some awareness—however minimal—of a catastrophic risk, even if he rationally deems it unlikely. Both agents engage in a routine, socially accepted behaviour (driving and phone use), and both do so while possessing some evidence, however weak, of potential danger. In both cases, the resulting harm stems not from negligence or malice, but from a small probability materialising into an actual threat.

What matters for my objection is that the *actual behaviour* of the agents remains virtually indistinguishable: they perform ordinary actions without immediate signs of recklessness, guided by evidence that leaves room for unexpected harms. Yet the moral responsibility account as it stands offers no clear criterion for why one is deemed *morally responsible and liable*, and the other not. The account hinges on subjective degrees of foreseeability—yet these degrees often differ only marginally and may be shaped by arbitrary factors like what news the agent happens to consume. This fragility calls into question the coherence of the moral responsibility account's judgments. If the mere act of reading a news article can shift an agent from *not liable* to *liable*, then the framework seems overly sensitive to minor epistemic differences that do not track morally significant distinctions in conduct.

A parallel difficulty arises once I re-examine the Falling Person case. Proponents of the moral responsibility account argue that because the person is involuntarily blown into the well by a strong gust of wind, she is *not morally responsible* for the threat and therefore *not liable*. On the face of it, this makes sense: the fall itself is not under the person's control, and thus she seems more akin to a mere projectile than to a culpable or even responsible agent. However, once I broaden the perspective and examine the agent's prior decisions, the case becomes far less straightforward. Suppose that the falling person had earlier chosen to leave her house despite weather warnings of unusually high winds. She was under no coercion; her decision was autonomous and informed. She knew, or ought to have known, that venturing outside in such conditions carried elevated risks—not necessarily of falling into a well, but of losing bodily control, being struck by debris, or endangering others in public space.

This setup, in my view, closely parallels that of Conscientious Driver. The driver also makes an autonomous decision to undertake an action—driving—that carries non-trivial risks to others, even if the danger materialises in an unforeseeable way (e.g., mechanical failure). In both cases, the agents make reasonable, non-culpable choices that initiate a causal sequence involving factors outside their control, ultimately culminating in an unjust threat to an innocent person. Importantly, neither agent is negligent; both act in line with social norms and available evidence. Yet the moral responsibility account assigns liability to the driver, while denying it to the falling person. This inconsistency suggests that the account fails to explain why initiating a risk in one domain (e.g., driving) grounds moral responsibility, whereas initiating a similarly structured risk in another domain (e.g., going outside in dangerous weather) does not.

One might argue that there are differences in the typicality or probability of harm in the two cases. But again, these differences are matters of degree, not kind. Just as the driver cannot predict exactly when or how a malfunction will occur, the falling person cannot anticipate being blown into a well. Yet both consciously enter into risk-laden environments with some awareness of possible adverse consequences. If the moral responsibility account wishes to base liability on voluntary engagement with risky conditions, it must treat these cases equivalently—or else articulate a principled reason why they are morally disanalogous. As it stands, the framework appears ad hoc in its applications, selectively attributing moral responsibility based on factors that may lack genuine normative weight.

My objection, in short, is that the moral responsibility account cannot clearly distinguish when the origin of risk is “close enough” to count as morally responsible. The line it draws between foreseeability and control is thin and unstable. Even if we accept that moral responsibility can exist without culpability, as in the Mistaken Attacker case, it is not at all clear how the moral responsibility account adjudicates the degree of risk-awareness or voluntary agency necessary to ground liability. It seems to rely on fine-grained and perhaps arbitrary distinctions in agents' epistemic access and causal proximity to harm. But moral liability should not hang on such fragile distinctions.

These challenges suggest that the moral responsibility account is underinclusive in problematic ways. If two cases are structurally analogous in terms of voluntary risk-taking and lack of control over the final harmful act, but the moral responsibility account treats them differently, then either the principles are flawed, or their application is unreliable. In either case, the moral responsibility account does not offer a stable or convincing foundation for judgments of liability.

4. A Revised Version of the Moral Responsibility Account

Although I have raised serious concerns about the standard formulation of the moral responsibility account, I believe that its core intuition remains defensible: that moral responsibility for posing an objectively unjustified threat should serve as the fundamental criterion for determining liability. What is needed, therefore, is not a rejection of the moral responsibility account altogether, but a revision that can offer a more consistent and inclusive framework for assessing moral responsibility across a wider range of cases.

On my revised view, we should treat all of the key agents discussed—the resident in Mistaken Attacker, the driver in Conscientious Driver, the Caller in Cell Phone, and the Falling Person—as morally responsible for the threats they pose. This is in line with the earlier analysis, which highlighted a crucial structural similarity among them: each of these individuals, through an autonomous and informed decision, initiates or becomes part of a causal chain that ultimately results in an unjust threat to another person. Importantly, in each case, the chain of events is foreseeable—not in the sense of being predictable in detail, but in the sense that it was not impossible or utterly beyond reasonable anticipation. For example, the driver knows that mechanical failure, while rare, is a known risk of driving; the Caller, especially if he had read reports of phone-triggered attacks, could rationally recognize a non-zero danger; the Falling Person, by choosing to go outside in dangerous weather, voluntarily exposes herself to an elevated risk of bodily instability. Each case involves some degree of awareness and voluntary engagement with risk.

We can also acknowledge, within this broader framework, that agents may differ in the degree to which they are morally responsible. For instance, the resident in Mistaken Attacker clearly bears the highest degree of moral responsibility: she actively chooses to engage in a behavior—aiming a shotgun at a stranger—that is certain to cause harm if her belief is mistaken. By contrast, the driver is less responsible, as she acts within social norms and without specific suspicion, yet still enters into a risk-laden activity. The Caller, especially in the unmodified version of the case, is responsible to an even lesser extent, given the minuscule likelihood of harm. The Falling Person, though often viewed as innocent, still made a voluntary choice to go outside under perilous conditions. These differences are gradational, not categorical. We therefore lack good reason to draw a sharp distinction between those who are deemed morally responsible and those who are not. What we need is an expanded understanding of moral responsibility—one that is capacious enough to encompass all agents who, through informed and voluntary action, contribute foreseeably to a threatening situation.

Crucially, there are cases where we can plausibly say an agent is not morally responsible at all. One such example is a person who has been drugged unconscious and then thrown from a rooftop, landing on and killing another. Here, the individual has taken no voluntary action whatsoever and cannot be said to have entered into any causal chain knowingly or deliberately. This kind of scenario represents the true limit case of moral responsibility—the kind that should mark the outer boundary of liability. But can my account plausibly explain the permissibility of defensive actions taken towards those who are morally responsible? I believe it can. As Victor Tadros insightfully notes in his critique of current approaches to liability:

> “Work on liability to be harmed has been distorted by a focus on death. Death is so awful that we are inclined to restrictive views about liability to be killed. When considering the extent of a person’s liability to be harmed, we are best to compare our candidates with those whose level of liability is very low—mere bystanders. Our question should be whether those who are culpable or who make a causal contribution to a threat are liable to bear any greater cost than these bystanders to avert threats. We then become more sensitive to the considerations affecting liability to be harmed” (Tadros 2016: 130).

This approach helps clarify the plausibility of my revised account. Recall the Conscientious Driver case. Suppose that the only way to save a pedestrian from being run over is to cause the driver to suffer a moderate injury—for instance, a broken arm. The alternative is to inflict that injury on an uninvolved pedestrian. Whom should we harm? Clearly, it is more justifiable to impose the cost on the driver, who has initiated the risk—even if not culpably—rather than on an uninvolved bystander. Most would agree that the driver cannot reasonably object to being harmed in this way.

Recall also the Cell Phone case. Suppose the police discover the imminent threat and must choose between tackling the Caller to the ground—causing a broken rib—or tackling the innocent person standing near the blast radius to protect them. Here again, it is far more plausible to harm the Caller. Even if the Caller had no intent or specific knowledge, he was, in some meaningful sense, morally responsible for the unfolding threat. It would seem morally absurd for him to claim that he, rather than the innocent bystander, should be spared harm.

Adopting Tadros’s *bystander baseline* gives me a more precise test. The question, in each dyadic encounter, is whether the agent has *stronger reasons than a mere* *bystander* to bear the defensive cost, and *how much stronger*. On my revision, the "broader" notion of moral responsibility is fixed by three variables: (i) voluntary initiation or maintenance of a risk-imposing activity; (ii) evidence-relative awareness of a non-trivial risk to others; and (iii) causal proximity to the unfolding threat. Liability then scales with a person's position on (i)–(iii): the higher the combined score, the larger the justified internalisation of defensive costs to that person rather than to the victim.

Applied, this yields determinate, case-level guidance. In Conscientious Driver, (i) and (ii) are satisfied and (iii) is high at the moment the car veers; the driver is therefore liable *at least* to bear *moderate* defensive costs (e.g., forcible stopping that foreseeably injures her) rather than those costs being imposed on an uninvolved pedestrian. In Cell Phone (unmodified), (i) is present, (iii) can be high at the moment of dialling, but (ii) is minimal; liability is therefore lower than the driver's yet above a bystander's. If Caller has minimal risk awareness (the modified variant), (ii) increases and so does liability. In Falling Person, although (iii) is satisfied, (i) and (ii) are minimal. Liability, therefore, is only slightly higher than *the bystander baseline*. By contrast, in Mistaken Attacker (resident), all three variables are high (intentional initiation, salient risk awareness given her evidence, and immediate causal indispensability), placing the resident at the *top* of the liability ranking. In short: *resident (Mistaken Attacker) > driver > caller (modified) > caller (unmodified) > falling person.* This graded structure explains the intuitive moral gradations between agents and why, other things equal, we should place defensive burdens on those who initiate and sustain the causal chains of harm, without relying on implausibly sharp epistemic thresholds.

Thus, by adopting Tadros's framework—comparing our subjects not to the dead, but to mere bystanders—we arrive at a more coherent and defensible account. My revised moral responsibility account, which incorporates a broader notion of moral responsibility grounded in voluntary, foreseeably risky conduct, can explain why liability tracks the degree of moral responsibility without needing to rely on implausibly sharp thresholds. It accounts for both the intuitive moral gradations between agents and the justice of placing burdens on those who initiate causal chains of harm.

This revised moral responsibility account also integrates naturally with the *discounting account* I defended in Chapter 1. On that view, to be liable is not to forfeit or lose one's rights altogether, but to have the moral disvalue of harming one discounted relative to others. The graded framework I develop here supplies the mechanism by which that discounting operates: degrees of voluntary initiation, evidence-relative awareness, and causal proximity determine the extent to which the moral weight of harming an agent is reduced when defensive costs must be distributed. Thus, the revised account not only avoids the fragility of sharp thresholds but also shows how liability can be understood as a matter of *fair harm distribution*, with liability functioning as a way of discounting the claims of those who have made themselves more appropriate bearers of defensive costs.

5. Conclusion

In conclusion, I have critically examined culpability, causal, moral-status, and duty-based accounts. I rejected all these accounts as satisfactory grounds of liability by strengthening existing objections and developing my own objections. I then examined the moral responsibility account. I offered brand-new challenges to the account. Based on my criticism, I have revised the moral responsibility account into a more inclusive account of liability while maintaining its original core idea. I believe the revised moral responsibility account has stronger explanatory power than all other accounts in a wide range of cases.

Chapter 3: The Hybrid Justification for Self-Defence

1. Introduction

In the previous chapter, I developed and defended a revised moral responsibility account of liability. In this chapter, I will revise and defend a hybrid justification for self-defence. First, let us recall three cases I discussed earlier:

> Mistaken Attacker: The identical twin brother of a notorious murderer is driving during a stormy night in a remote area when his car breaks down. Unaware that his brother has recently escaped from prison and is thought to be hiding in this same area, he knocks on the door of the nearest house, seeking to phone for help. On opening the door, the armed and frightened resident mistakes the harmless twin for the murderer and lunges at him with a knife. (McMahan 2005: 387).

> Conscientious Driver: A person keeps his car well maintained and always drives cautiously and alertly. On one occasion, however, freak circumstances cause the car to go out of control. It has veered in the direction of a pedestrian whom it will kill unless she blows it up by using one of the explosive devices with which pedestrians in philosophical examples are typically equipped (McMahan 2005: 393).

> Falling Person: A gust of wind blows a person down a well where the victim is trapped; unless he vaporises her with a ray gun, she will crush and kill him; if he does not vaporise her, she will survive the fall (Nozick 1974: 34).

According to the original version of the moral responsibility account, the resident and the driver are minimally responsible threats. Each of them poses a threat of objectively wrongful harm to an innocent victim. However, each is epistemically justified in believing that what she is doing is morally permissible. They are engaging in acts which, as they recognize, impose a small but non-trivial risk of wrongful harm on other people. According to my revised version of the moral responsibility account, the falling person is also considered as a minimally responsible threat. She also voluntarily engages in an act which imposes a small but non-trivial risk of wrongful harm on others.

The problem is, then, whether it is morally permissible to kill MRTs in self-defence? The defenders of the original moral responsibility account have argued that it is intuitively permissible to kill the resident and the driver, and a satisfactory theory of self-defence should be able to explain why they are liable to be killed (McMahan 2005; Otsuka 2016; Gordon-Solmon 2018). Some have argued it is intuitively permissible to kill the resident, the driver, and the falling person (e.g. Thomoson 1991; Frowe 2014). On the other hand, some have argued it should be sometimes impermissible to kill MRTs. Quong have argued it is intuitively permissible to kill the resident, but intuitively impermissible to kill the driver and the falling person (Quong 2020: Ch.2). Some would suggest that it is intuitively impermissible to kill any MRT and a theory which permits the killing of MRTs is too permissive (Lazar 2009; Ferzan 2012).

I believe the difference in intuition regarding whether it is permissible to kill MRTs in self-defence is something we should take seriously. We should not let significantly disputed intuition be the criterion for evaluating different theories. Instead, I believe a satisfactory theory of self-defence should explain why people's intuitions diverge in these cases. To do so, I argue that we should accept a hybrid justification of self-defence (Bazargan 2014). The hybrid justification can explain why people's intuitions diverge in cases of MRTs. It also avoids many problems faced by other theories of self-defence.

I proceed as follows. In section 2, I introduce the simple account of liability, which holds that, if one believes MRTs are liable, they are liable to bear the full cost of the defensive harm. This means they are liable to lethal harm if it is necessary. In section 3, I introduce the structure of Bazargan's hybrid justification. In section 4, I explain how my version of the hybrid justification is different from Bazargan's. In Section 5, I discuss the advantages of adopting a hybrid justification, including both those identified by Bazargan and further advantages that emerge from my revised account. In Section 6, I defend the hybrid justification against a series of objections.

2. The Simple Account of Liability

A natural and influential starting point for thinking about the permissibility of killing MRTs is what Saba Bazargan (2014) calls the *simple responsibility-based account of liability* (henceforth, the simple account of liability). This view is widely accepted and supported by the proponents of the original moral responsibility account (e.g. McMahan 2005; Otsuka 2016; Gordon-Solmon 2018). On this view, self-defence is essentially a problem of distributing an unavoidable harm between two parties. When only one of two people must suffer, the morally relevant comparator is *relative responsibility*. Whoever is more responsible for creating the situation in which harm is unavoidable should bear the cost, since that is fairer than imposing it on the less responsible party.

The principle of the simple account of liablity can be summarised as follows:

> Simple account of liability: The party who is more morally responsible for a wrongful threat is liable to suffer whatever harm is necessary to prevent that threat from being imposed on her victim.

This simple account thus treats liability as an *all-or-nothing* matter. Once the balance of responsibility moves to favour one party, that party becomes liable to the full defensive harm.

There is, indeed, some intuitive appeal of this simple account. Defensive harm arises in tragic, forced-choice contexts. In these contexts, someone must bear a cost. The victim has done nothing to warrant losing her life. On the other hand, the threatener has, through her own agency, created the danger. Even if the threatener acted permissibly given her evidence, fairness dictates that she—not the wholly innocent victim—should absorb the loss if the risk materialises.

Recall the earlier cases. In Mistaken Attacker, the resident mistakenly but reasonably believes she faces a villainous aggressor. On the simple account, she is liable to be killed, since her victim is entirely innocent and bears no responsibility at all. In Conscientious Driver, the driver is minimally responsible for imposing the risk since she chooses to drive, knowing there was a small but non-negligible chance of harming others. When that risk materialises, she is more responsible than the pedestrian. Therefore, she too is liable to be killed if that is the only way to save the victim.

According to the original moral responsibility account, the falling person is not morally responsible for creating the threat. Since the threat issues entirely from external forces (a gust of wind that blows her from a height), she has played no morally relevant role in generating the danger. According to my revised account, however, matters look different. The falling person is at least *minimally morally responsible*. She is fully aware that her body has weight, and falling from a great height would almost certainly cause lethal harm to others. Yet she voluntarily chooses to go out despite this background risk. When an unlucky gust of wind does in fact blow her off the ledge, the risk she knowingly carried materialises and threatens an innocent person.

If we now combine my revised moral responsibility account with the simple account of liability, the implication is that the falling person, no less than the mistaken resident or the conscientious driver, is *liable to be killed*. For she, rather than her wholly innocent victim, is the more responsible party for the wrongful threat that must inevitably fall on someone.

The simple account’s strength lies in its clarity and determinacy. It provides a decisive answer in hard cases: identify who is more responsible, and allocate the full defensive burden to that party. It also captures a powerful fairness intuition: it is better that the threatener suffers than the wholly innocent victim.

However, the account’s simplicity makes it problematic. First, it delivers harsh results for MRTs. Even those who are only minimally responsible, like the conscientious driver and the falling person who have done everything morality can reasonably demand, are, on this account, fully liable to be killed. As discussed earlier, many believe this is counterintuitive. Second, because liability is comparative and all-or-nothing, the simple account might imply that multiple MRTs threatening the same person are each fully liable to be killed. This result conflicts with the widespread intuition that it cannot be permissible to kill a large number of MRTs in defence of one victim. The account seems to have no internal resources to forbid such killing.

These difficulties suggest that while the simple account has some intuitive appeal, it is ultimately unsatisfactory. It sets the stage for more nuanced alternatives. I will now turn to Bazargan’s alternative solution, the hybrid justification.

3. Bazargan’s Hybrid Justification

Bazargan’s hybrid justification of defensive harm offers a two-pronged framework that departs from standard liability-based approaches. His account is motivated by a desire to explain how it can be permissible to defensively kill someone like an MRT, despite the fact that she *is not liable to be killed*. At its core, the hybrid justification incorporates two components: a refined theory of liability that limits what an MRT is liable to, and a separate lesser-evil rationale that can justify defensive killing even when it exceeds that limit.

3.1 The Complex Responsibility-Based Account of Liability

The first component is a proportional theory of liability, which Bazargan labels the "complex account." This view challenges the "simple" liability model as I have discussed earlier. It holds that once a person is deemed morally responsible for a wrongful threat, she can be harmed up to the full amount needed to avert the threat. Bazargan also finds this implausible, especially when applied to minimally responsible agents.

Instead, he proposes that liability should be a function of both the severity of the threat a person poses and the degree of moral responsibility she bears for it. This gives us a proportional model: if a person is *n*% responsible for a threat of harm *H*, then she is liable to defensive harm equivalent to those who fully responsibly creates a threat of *n*% of *H*. Importantly, this model assumes that both harm and moral responsibility can be at least roughly quantified or ranked along a scalar scale (Bazargan 2014: 121-122).

To illustrate, suppose that death constitutes 100 units of harm. One can say that a typical MRT, such as the driver, is 5 to 10 percent responsible for a threat of harm of 100 units. Bazargan stipulates that the driver is 5 percent responsible for the threat of death she imposes. On the complex account, the driver's liability to defensive harm would be equivalent to the liability of a fully responsible and culpable aggressor who threatens to impose a harm of 5 units (that is the 5 percent of the harm of death). Bazargan further stipulates that, for the sake of argument, fully responsible and culpable aggressor is liable to suffer a harm *ten times greater* than the harm which she threatens to impose. This is intuitive as it is usually morally permissible to inflict greater harm to avert a culpable threat. For example, one might permissibly kill the aggressor if the aggressor is culpably threatening some severe bodily injuries or rape (which are considered less harmful than death). So, the driver's liability to defensive harm, as stipulated, is calculated as the 5 percent of moral responsibility for threatening the harm of death (100 units) times ten, which is 50 units of harm

(Bazargan 2014: 122-123, 126).

This proportional account preserves the verdict that MRTs are not completely innocent since they are morally responsible for imposing non-trivial risks. Also, it avoids the troubling result that MRTs are liable to suffer *the full cost* simply because they are more responsible (arguably, slightly more responsible) than the potential victims.

3.2 The Lesser-Evil Justification and the Lesser-Evil Discounting View

If MRTs are not liable to be killed, when is it permissible to kill them? According to Bazargan's illustration of the Conscientious Driver case, the driver is liable to only 50 units of harm. But killing the driver would be imposing 100 units of harm on the driver. How can such killing be justified? Bazargan's answer is to add a lesser-evil justification. If the harm exceeds what the aggressor is liable to, it can nonetheless be justified if it is the lesser evil.

However, this is not straightforward. It is widely accepted that, in order to have a lesser-evil justification for inflicting harm on a person who is not liable to that harm, the harm imposed must be sufficiently small with respect to the harm it wishes to prevent. The key modification Bazargan made to the lesser-evil justification is the lesser-evil discounting view. It holds that, when comparing the two exclusive options: (a) impose a harm on the MRT *beyond* what she is liable to, or (b) allow that harm to fall on her wholly innocent victim, the moral disvalue of imposing the extra harm on the MRT should be *discounted* in light of her partial responsibility. Other things equal, it is *fairer* for the unavoidable loss to fall on the person who bears *some* responsibility for the predicament. The lesser-evil discounting view, according to Bazargan, permits the victim to impose greater harm beyond the aggressor's liability to avert the threat, by treating the harm imposed on the MRT beyond her liability as morally less bad. (Bazargan 2014, pp. 127–29).

3.3 How the Two Components Work Together: An Illustration

Having set out Bazargan's two components of the hybrid justification: the complex account of liability and the lesser-evil discounting view, let us see how these two principles combine. As Bazargan summarizes:

> "Though MRTs are not morally liable to be killed, there is an agent-neutral permission to kill them defensively. This is because imposing a lethal harm on someone who bears some moral responsibility for an unjust threat is (ceteris paribus) the lesser evil relative to the alternative of allowing the MRT to kill her non responsible victim. The lesser evil is substantial enough to provide a justification for killing the MRT since the disvalue of the harm imposed on the MRT is discounted relative to the weight of the harm that the MRT would otherwise impose on her victim." (Bazargan 2014: 129).

To illustrate, consider again the case of the Conscientious Driver. The driver is 5 percent morally responsible for posing a threat of harm of 100 units (death). According to the complex account of liability demonstrated above, her liability is equivalent to an aggressor who fully responsibly and culpably threatens harm of 5 units. If we hold that a fully culpable aggressor is liable to defensive harm ten times greater than the harm she threatens, the drive is liable to 50 units of harm.

To kill driver, the pedestrian would impose 100 units of harm on the driver. This would exceed driver's liability by 50 units of harm. According to Bazargan's hybrid justification, the 50 units of harm beyond the driver's liability can be justified as a lesser evil since the driver is partial morally responsible. The moral disvalue of harming her beyond her liability is discounted, and thus, the 50 units of harm can be a lesser evil.

A similar structure applies to another main example, the Mistaken Attacker. The resident, thinking she is under attack by a dangerous criminal, shoots an innocent person who happens to resemble the criminal. Her belief is epistemically justified, but she is still engaged in a morally risky action: using lethal force in self-defence. If a third party could stop her only by killing her, we face a dilemma.

On the complex account, the resident may again be liable to some defensive harm (perhaps more than the driver, since she acts intentionally), but the resident is likely not liable to be killed. Still, the act of killing her might be justified under the lesser-evil principle: she imposed the risk, even if justifiably, and so the moral disvalue of killing her is reduced relative to killing her victim.

To summarize, Bazargan's hybrid justification thus consists of two interlocking parts. The first is the complex account of liability. A person is liable only to harm proportional to her degree of moral responsibility for an objectively unjust threat. MRTs, by definition, fall low on this scale. The second is the lesser-evil discounting view. Even if a person is not liable to the full harm necessary to prevent her threat, it may still be permissible to inflict that harm if the overall outcome is the lesser evil—especially if her partial responsibility warrants discounting the moral disvalue of harming her.

4. Distinguishing My View from Bazargan's

While I adopt Bazargan's hybrid structure as a promising foundation for thinking about defensive harm, and I certainly agree with Bazargan on the complex account of liability. My version of the hybrid justification diverges from his in two important respects: first, in its normative aim, and second, in its rejection of the lesser-evil discounting view. These two differences are closely related, and together they result in a significantly different theoretical orientation and practical application.

4.1 A Difference in Aim

Bazargan's hybrid account is intended as a *justification for killing MRTs*. His goal is to show how, despite the fact that MRTs are not fully liable to lethal harm, we can still permissibly kill them if doing so is the lesser evil. His account is therefore justificatory in nature. It aims to vindicate intuitively permissible acts of defensive killing.

By contrast, my version of the hybrid account is not committed to showing that it is permissible to kill MRTs in these cases. Rather than offering a justification, my aim is explanatory. I intend to offer a framework for morally evaluating MRT cases, one that helps make sense of the ambivalent and divergent intuitions we often have in response to them. I do not assume that we must resolve MRT cases with a binary verdict of permissible or impermissible. Instead, I aim to preserve the complexity of the moral intuitions and judgements by showing why people have such divergence in the first place.

4.2 Rejecting the Lesser-Evil Discounting View

A second key difference concerns the mechanism by which killing MRTs is justified when it goes beyond their liability. Bazargan proposes that, although an MRT is only liable to a fraction of the harm required to avert her threat, the remaining harm can be justified by appeal to lesser-evil reasoning, provided we discount the moral disvalue of harming her in light of her partial responsibility. This is his lesser-evil discounting view as explained above

I do not accept this view. The key reason is that Bazargan and I have fundamentally different view on the nature of liability. On Bazargan's picture, liability marks the boundary of *forfeited rights*. Within that boundary, harming the threatener does not wrong her. On the other hand, beyond liability, any further harm is a *rights-infringing* imposition that must be justified by a separate lesser-evil principle. Such rights-infringing harm can be further *discounted* in light of the threatener's partial responsibility (Bazargan 2014: 123, 129-131). On my view, as defended in Chapter 1, this understanding is misguided. The best interpretation of liability is itself *discount-based*. To say that an aggressor is liable is precisely to say that the *moral disvalue* of harming her is already *discounted* by her responsibility for the predicament. Liability, on my account, just is responsibility-indexed *discounting* of the harm's disvalue. Once liability is understood this way, Bazargan's lesser-evil discounting is not a new justificatory consideration at all; it simply *repeats* the same responsibility-grounded discount. This is why the *double-counting* worry arises (which Bazargan acknowledges and answers by distinguishing first-order fairness from second-order distributional fairness). But that reply presupposes a forfeiture model I reject: if liability is already a responsibility-sensitive discount on the badness of harming the threatener, re-invoking responsibility to *further* discount the cost of a rights-infringing imposition counts the same normative factor *twice* (cf. Bazargan 2014, 129–31).

Second, my hybrid justification does not need a discounting principle at the lesser-evil stage. As demonstrated above, Bazargan's aim is to vindicate the *permissibility of killing MRTs*. Thus, the additional discounting is introduced to bridge the gap between ordinary lesser-evil constraints and the permissibility verdict he seeks. My aim is different. I do not assume that killing MRTs must come out permissible. I can accept that in some cases killing MRTs is impermissible, and I take seriously the fact that our intuitions diverge across cases. Precisely because I do not build in a second discount, my version better explains that divergence: liability (as discount) can speak in favour of harming MRTs up to a point, while any further harm must be examined whether it is a lesser evil with respect to ordinary lesser-evil constraints. These include considerations such as the doing and allowing distinction and the means principle, which I will discuss in more detail in later sections.

5. The Advantages of the Hybrid Justification

I have explained how my version of the hybrid justification is different from Bazargan's account. I now turn to the advantages of the hybrid justification. I will demonstrate both the inherited advantages from Bazargan's view and new advantages I propose.

5.1 Multiple threats

Here is an advantage of the hybrid justification which Bazargan and I are in agreement. Bazargan argues that a significant upshot of the hybrid justification over the simple account of liability is its ability to explain why it is wrong to kill multiple MRTs (Bazargan 2014: 132-133). Consider a scenario where there are multiple conscientious drivers. Each driver is, on her own, causing a threat that is both necessary and sufficient for the harm to occur unless stopped. Even though there may be debate about whether it is permissible to kill the driver, nearly everyone—including supporters of the Simple Account—agrees that it would be impermissible to kill multiple drivers each threatening the same victim with death.

The problem for the simple account lies in its inability to account for the number of MRTs. Since each MRT is individually liable for the harm she causes, proponents of this view must claim that killing any number of MRTs does not wrong them or infringe their rights. Thus, proponents of the Simple Account cannot explain why it feels intuitively wrong to kill multiple MRTs threatening the same innocent person.

The hybrid justification, however, resolves this issue effectively. According to this account, the reason why killing an MRT might be permissible is because it is the lesser evil. Importantly, this justification acknowledges that killing an MRT involves harming her beyond her liability, giving her a basis for complaint. It becomes morally significant when aggregated: killing several MRTs eventually ceases to qualify as the lesser evil. Consequently, the hybrid justification reaches the intuitive conclusion that it is impermissible to kill multiple MRTs.

5.2 Compensation and Ideal distribution of harm

In this subsection, I will propose a new advantage of the hybrid justification. It is not identified by Bazargan, but it is partly related to another advantage offered by Bazargan. To help see my point, let us first consider Bazargan's point.

According to Bazargan, the hybrid justification offers another advantage (2014, p.133). Intuitively, an MRT is owed compensation for any harm imposed on her that exceeds what she would be liable for if the harm could be distributed. In such cases, we have the opportunity, ex post, to address this injustice by fairly redistributing the costs of the defensive harm. However, under the simple account, if the MRT—or her estate—is owed compensation while she remains liable to be killed, this undermines the principle that no one can be entitled to compensation for the harm to which they are liable. Rejecting this principle comes at a significant theoretical cost.

The hybrid justification avoids this issue. On this account, the MRT is not liable for harm that exceeds what she would bear if it were distributable. Instead, killing her is justified as the lesser evil. However, because the harm imposed exceeds her actual liability, her rights are infringed. This rights infringement provides a clear basis for compensation: we justified harming the MRT, but we also treated her in a way to which she was not morally liable. Therefore, she—or her estate—is owed compensation for the harm that exceeds her liability.

The advantage of the hybrid justification I propose is related to the explanation of why MRTs are owed compensation. To me, Bazargan merely claims it is *intuitive* that MRTs are owed compensation without offering deeper explanations. I agree with Bazargan's conclusion that an MRT is owed compensation if she is killed, and the hybrid justification can adequately account for that. I wish to explain why this is the case. To explain this, we need to refer to the ideal distribution of harm. I believe this is also a new upshot of hybrid justification.

Recall the case of Conscientious Driver. Assume that we can freely distribute the harm between the driver and the pedestrian while holding the total amount of harm constant. For instance, we can inflict *n* units of harm on the driver and let the pedestrian suffer 100-*n* units of harm. Now it seems intuitively impermissible to kill the driver (i.e.

inflicting 100 units of harm on the driver and letting the pedestrian suffer no harm at all) since there are morally better options available. What, then, is the ideal distribution of harm? Indeed, it is difficult to determine the exact values. However, the upshot of the hybrid justification is that it recognizes that there exists an ideal distribution of harm, despite the fact that it is difficult to determine. But we can imagine that, once we have determined the exact amount of harm to which the driver is liable and the moral significance of the constraints of the lesser-evil justification (such as the doing and allowing distinction), it is likely to result in an ideal distribution. Let us assume that, for the sake of argument, the ideal distribution of harm is inflicting 60 units of harm on the driver and letting the pedestrian suffer 40 units of harm. Then, if it is possible to achieve the ideal distribution, it is the only justified option. Even if it is impossible, it remains the ideal distribution. The only difference is that one of the options in the real world now becomes the justified option.

Before I elaborate on the moral significance of the ideal distribution, I will first explain why this is not available to the simple account of liability. The reason is that the simple account views liability as a comparative matter. It holds that individuals who are more morally responsible for a wrongful threat are liable to the full harm necessary to prevent that threat. According to the simple account, the amount of harm to which the aggressor is liable is sensitive to real-world factors. If the harm cannot be divided and distributed, then the most responsible person is liable to the full amount of harm. On the other hand, the hybrid account holds liability as a non-comparative matter. The amount of harm a person is liable to is solely determined by her degree of responsibility for a certain threat. Even if the harm cannot be divided and distributed, the harm to which the person is liable is not changed.

I believe that recognizing the ideal distribution of harm is significant. Let us consider an analogy of distributive justice more generally. Many believe in some ideal principles of distributive justice. An ideal society should implement policies that follow these principles when distributing goods and benefits. In the real world, it might be impossible or extremely difficult for a society to do so.

Assume that the political leaders must choose between Policy Set A, which follows 90 percent of the ideal principles, and Policy Set B, which follows 50 percent of the principles. Obviously, they must choose Policy Set A. But the fact that Policy Set A is the best option available does not affect our beliefs in an ideal Policy Set X which perfectly follows all the principles. We can still say that, although Policy Set A is the current best option, it is not ideal. We can still urge the political leaders to improve the policy. The ideal principles set the direction we should be working towards.

Now back to the discussion of defensive harm. The central idea of the hybrid justification is that liability determines a person’s “fair share” of harm. The fair share of harm is solely determined by her choices and actions and not affected by contingent factors such as how available defensive options can distribute the harm. The simple account, in its most plausible version, merely tells us what the best available defensive option is. The mistake of the simple account is to imply the best available defensive option is the ideal one since it claims that the most responsible person is *liable* to the *full* harm. On the other hand, the hybrid account holds there is an ideal distribution of harm which states the fair share of harm for each individual. A defensive option might be justified when it approximates the ideal distribution of harm. However, a defensive option can only be ideal when it perfectly matches the ideal distribution of harm. For example, if a villain is fully culpable and trying to kill you, it is justified to kill her. It is also the ideal distribution of harm since, in this case, the villain should be liable to all the harm (i.e., her fair share of harm should be the total harm). But in cases of MRTs, even if it is justified to kill an MRT, she is *not* liable to the full harm. Killing might be a justified defensive option, but not an ideal one.

Now let us go back to consider why MRTs are owed compensation. It is true that compensating the MRT is a remedy to offset the wrongful harm she has suffered. But the moral significance of compensating the MRT lies in the fact that it is an ex-post effort to achieve a better approximation of the ideal distribution of harm. As stated earlier, let us assume that the ideal distribution of harm is inflicting 60 units of harm on Driver and allowing 40 units of harm on Victim. Let us compare a) killing the driver and compensating her estate afterwards and b) killing the driver without compensating her estate afterwards. Clearly, option a) is a better approximation of the ideal distribution since the pedestrian bears more costs by compensating, and compensation to Driver can be considered as a reduction to the total amount of harm she has suffered.

6. Defending the Hybrid Justification

In previous sections, I have developed my own version of the hybrid justification and demonstrated its advantages. I now defend the hybrid justification against two objections proposed by McMahan.

6.1 The Double Counting Problem

The first objection to the hybrid justification is the double counting problem. McMahan argues that the hybrid justification involves an objectionable form of double counting, where the prevention of harm to the victim is counted twice. Take Conscientious Driver as an example. The driver is an MRT, and she is liable to some degree of defensive harm. Suppose death is 100 units of harm and the driver is liable to 50 units of harm. The driver is liable to suffer 50 units of harm as a means of preventing the pedestrian from suffering 100 units of harm. The infliction of the other 50 units of harm, as Bazargan claims, is the lesser evil, compared to allowing the pedestrian to suffer 100 units of harm. In this case, according to McMahan, the prevention of the 100 units of harm to Victim is counted twice. But the killing of the pedestrian is not prevented twice. McMahan argues that if preventing the driver from killing the pedestrian justifies inflicting 50 units of harm on the driver, it exhausts its power to justify further harming on the driver (McMahan 2017: 21).

McMahan offers a possible solution to the double counting problem but remains sceptical about whether it actually solves the problem (2017: 21-22). Suppose that a person can save himself by inflicting some harm on an MRT (the amount of harm inflicted is equal to the amount of harm to which the MRT is liable) and unavoidably inflicting a small amount of harm on an innocent bystander as a side-effect. In this case, it seems that the harm inflicted on the bystander is justifiable as a lesser evil. Whether the harm inflicted on the bystander is the lesser evil is determined by comparing the amount of harm on the bystander and the prevention of harm on the victim. If the harm inflicted on the bystander is substantially less than the harm prevented, it is justified to harm the bystander as a lesser evil. However, this case also seems to involve the kind of objectionable double counting. After all, the prevention of harm to the victim is counted twice (once in determining the liability of the MRT and the other in determining whether harming the bystander is a lesser evil).

The reason that McMahan is sceptical about the solution is that McMahan claims there are two morally significant differences between cases involving bystanders and cases of MRTs. I reject McMahan's scepticism and argue that cases involving bystanders indeed help us understand why there is no double counting in both kinds of cases.

One reason for scepticism given by McMahan is that, in the bystander case, the harm to the bystander is a side effect. In MRT cases, the harm beyond the MRT's liability is an intended means. I agree that this is a morally significant difference, but I do not think that this is relevant to the double counting objection. It is likely that harming someone intentionally as a means is morally worse than harming someone as a side-effect. It is much more difficult to justify the former than the latter. But McMahan fails to show why this difference is relevant to the double counting objection. The issue here is not which way of harming is easier to justify. Instead, according to the double counting objection, the issue here is whether the prevention of harm can be considered twice. If the double counting objection is true and the prevention of harm on the victim can only be counted once, it means that any further harm cannot be justified, regardless of whether it is a side-effect or an intended means. If it is acceptable to double-count the prevention of harm in bystander cases and unacceptable to double-count the prevention of harm in MRT cases, it would follow that the double-counting objection only rises when the harm inflicted is an intended means. I am sceptical that an argument in support of this distinction can be provided.

The other reason for scepticism offered by McMahan is that, in the bystander case, the harm imposed on the aggressor is completely effective on its own. In MRT cases, the harm to which the MRT is liable is less than death. But only death is effective in preventing the MRT from killing the victim. Inflicting the amount of harm to which the MRT is liable is wholly ineffective (McMahan 2017: 22). I disagree with McMahan that this is a morally significant difference between two kinds of cases. In fact, if we consider the cases realistically, we can find that they are rather symmetrical. Recall Conscientious Driver. The only two options for the pedestrian are 1) doing nothing and letting himself be killed and 2) killing the driver with a grenade. In this case, it is *impossible* to inflict the exact amount of harm to which the driver is liable. Thus, the question of whether such harm is effective does not rise at all. It is the *act* of killing the driver with a grenade that is effective. The *act* consists of inflicting 100 units of harm, and the harm can be broken down into harm to which the driver is liable and harm beyond the driver's liability. The point of breaking down the harm is to morally evaluate the *act*. I fail to see the point of discussing whether some amount of harm itself is effective.

Similar ideas can be applied to cases involving bystanders. Suppose that the victim can only save herself by shooting the aggressor with her gun, but the bullet will penetrate and injure an innocent bystander. The *act* of taking the shot is effective. The act will harm both the aggressor and the bystander. There is no possible act which only inflicts harm on the aggressor. If we want to claim that the harm inflicted on the aggressor is effective on its own, we must say something such as that the movement of the bullet from the gun to the aggressor's body is effective, and what happens afterwards is irrelevant. But this is absurd.

I will now explain why there is no double counting in both bystander and MRT cases. Let us take Conscientious Driver as an example. The *act* of killing the driver with a grenade consists of 100 units of harm inflicted on the driver. Since the driver is morally responsible for posing the threat, she is liable to 50 units of harm to prevent the pedestrian from being killed. It would be double counting if we try to justify the other 50 units of harm to the driver as a means to prevent the 100 units of harm to the pedestrian since, in this way, the prevention of the harm to the pedestrian is counted twice. But we do not have to do that. We can say that the 50 units of harm beyond the driver's liability is unjustified *yet*. It is true that, according to the driver's liability, she is wronged by the 50 units of harm. So, the *act* of killing the driver consists of 50 units of wrongful harm. Moreover, the *act* of killing the driver is *intentionally inflicting* 50 units of wrongful harm. But killing the driver is not the only act available to the pedestrian. He can also choose to do nothing and let himself be killed. The *act* of doing nothing consists of 100 units of wrongful harm since Victim is wholly innocent. Doing nothing is *allowing* 100 units of wrongful harm. The following table demonstrates the comparison:

	Wrongful Harm Inflicted	Wrongful Harm Prevented	Wrongful Harm Allowed
Killing the driver	50	100	0
Doing Nothing	0	0	100

In this case, the possibility that killing the driver may be justifiable is not because the 50 units of wrongful harm inflicted is a lesser evil compared to 100 units of wrongful harm prevented. The prevention of the 100 units of harm is why driver is liable to 50 units of harm. The reason why the 50 units of harm might be the lesser evil is that it might be a lesser evil compared to the 100 units of wrongful harm allowed if the pedestrian chooses to do nothing.

This explanation also yields the correct result in the bystander case. We can say that, in a bystander case, the victim has two options. One is to allow herself to be killed by the aggressor, and the other option is to kill the aggressor who is fully liable and to inflict a small amount of wrongful harm on the bystander. The bystander is innocent, and the harm inflicted on her would wrong her. However, harming the bystander can still be justifiable, all things considered. The reason is that harming the bystander might be the lesser evil, compared to allowing much greater harm to the victim.

6.2 Issues with the Lesser-Evil Justification

The second objection to the Hybrid Account is that, according to McMahan, the lesser-evil justification is not powerful enough to justify the killing of MRTs (2017: 22-23). For there to be a lesser evil justification for the infliction of harm on a person who is not liable to that harm, the harm caused must be sufficiently small, or that prevented sufficiently great, to justify overriding the constraint against harming. McMahan argues that, even if we hold that an MRT is liable to 90 units of harm, there cannot be a lesser evil justification for killing the MRT. 10 units of harm is ten percent of the harm of death. Suppose that in both the bystander case and the MRT case, both the victim and the aggressor would lose fifty years of good life in being killed. On that assumption, ten percent of the harm of death is equivalent to the loss of five years of good life. In the bystander case, it would not be justifiable as the lesser evil to cause an innocent bystander a loss equivalent to the loss of five years of good life as a side effect of preventing the innocent victim from losing fifty years of good life. That would be disproportionate. In the MRT case, effective defence requires that the same harm be inflicted not as a side effect but as an intended means, which is morally worse and much more difficult to justify.

McMahan claims that one might argue that because the harm that must be justified as the lesser evil in the MRT case would be inflicted on a person who is already liable to a substantial amount of harm, it can be greater, other things being equal, than the maximum that could be justified if the victim were not liable to any harm at all, as is true of the innocent bystander in the bystander case. This would be analogous to the view that inflicting a certain amount of punishment on a guilty person beyond what he deserves is less objectionable than inflicting the same amount of punishment on an innocent person (McMahan 2017: 23).

Bazargan’s response is to propose the lesser-evil discounting view, which I have discussed earlier. According to the discounting view, the moral disvalue of the additional harm to the MRT should be discounted in light of her partial responsibility, thereby making it easier for lesser-evil reasoning to justify killing her (Bazargan 2014: 127–129). However, as I argued in Section 4, I reject this view. On my account, the nature of liability itself is best understood as a discounting of the moral disvalue of harming someone, indexed to her degree of responsibility. There is no additional moral reason to add a further “discounting” principle at the lesser-evil justification stage.

If my version of the hybrid justification does not appeal to the lesser-evil discounting view, how can I respond to McMahan’s objection? It is natural for my account to say that McMahan’s objection is largely beside the point. His worry bites only if one assumes that the aim of the theory is to show that the killing of MRTs must always be permissible. In this case, the account would require some additional principle, such as the discounting view, to bridge the gap between liability and permissibility. But my aim here, as I have claimed, is different. I do not need to guarantee the permissibility of killing MRTs in every case. Instead, my aim is explanatory: to offer a framework that accounts for why people’s intuitions *diverge* in MRT cases.

This is why McMahan’s objection, in principle, does not undermine my account. The lesser-evil justification in my framework is not meant to conclusively prove permissibility, but rather to help model the moral structure that produces divergent intuitions. To see this more clearly, consider the MRT case in an abstract way. The MRT poses an objectively unjustified threat. She is *x* percent responsible for that. Thus, she is liable to *n* units of harm. Killing the MRT would be intentionally inflicting 100-*n* units of wrongful harm. Doing nothing would be allowing 100 units of wrongful harm. Let us further assume that the moral disvalue of intentionally inflicting 1 unit of wrongful harm is *a* and the moral disvalue of allowing 1 unit of wrongful harm is *b*. Now there are two difficulties. How to determine values of *x* and *n* in a given MRT case such as Conscientious Driver? How to determine the ratio between *a* and *b* in order to determine the value of *n* so that allowing 100 units of wrongful harm is morally equivalent to intentionally inflicting 100-*n* units of wrongful harm? In fact, I believe there are no definitive answers to these questions. I believe this explains why people have different intuitions about MRT cases. For those who think that it is intuitively permissible to kill MRTs, it might be the case that the value of *n* is larger, and the ratio between *a* and *b* is smaller. For those who think it is impermissible to kill MRTs, the value of *n* might be smaller, and the ratio between *a* and *b* might be larger.

This is exactly how my version of the Hybrid Account is different from Bazargan’s. My version remains largely neutral on whether a given MRT is liable to be killed. Instead, it provides a theoretical framework to account for the disputed intuition regarding this problem.

However, my explanation above is incomplete. It is mistaken to disregard McMahan’s objection only by referring to my aim of the hybrid justification. For the hybrid justification to explain the divergence in intuitions about killing MRTs (why it can be reasonable to think such killings are sometimes permissible and also reasonable to think they are sometimes impermissible), it cannot collapse into one of two extremes.

On the one hand, it must not permit all cases of killing MRTs. On the other, it must not forbid all of them. If it permits all such killings, it could not capture the intuitive hesitation many people feel in these cases. If it forbids all of them, it would fail to accommodate the intuitive judgments of many others who think it is sometimes permissible to kill MRTs.

I believe the moral significance of the doing and allowing distinction is crucial here. If we accept McMahan’s critique of the moral significance of this distinction (that it is a stringent moral constraint), then the hybrid justification risks prohibiting almost all cases of killing MRTs. This is because, in most MRT cases, the defensive killing would count as “doing” harm beyond the aggressor’s liability, and with the doing and allowing distinction being a stringent constraint, the lesser-evil calculation would rarely overcome the constraint against killing someone not fully liable. On the other hand, if we deny that the doing and allowing distinction is morally significant at all, the hybrid account will swing to the opposite extreme. It will permit killing all MRTs even when they bear only a slight degree of liability (such as the falling person) because the lesser-evil reasoning will treat killing as no worse than allowing the same harm to occur.

We therefore need a version of the hybrid account that recognises the moral significance of the doing/allowing distinction while calibrating its strength. I will now try to sketch a comparative picture of the strength of the doing/allowing distinction. I will demonstrate that while doing harm is generally harder to justify than allowing harm, the magnitude of that difference varies across cases because, in some cases, there are other independent moral constraints. Given my purpose and the constraints of space, in what follows, I will not attempt to distinguish between or defend any particular view in detail. My aim is only to show that the line of thought I propose is defensible.

Let us begin with a comparison between two classic cases. In the first case, you are a bystander and you can save five innocent people by diverting a runaway trolley into the side track and killing one innocent person. In the second case, you are still a bystander and you can save five innocent people from a runaway trolley by pushing a fat person off a footbridge and using his body to block the trolley.

I believe the prohibition of harming in the second case can be more plausibly explained by a separate moral principle: the moral constraint on using a person as a means to achieve one's ends. Similar principles and ideas are widely accepted and defended (Quinn 1989; Frowe 2014; Tadros 2020; Quong 2020). In these cases, the doing/allowing distinction is not the primary driver of the moral judgment, and the constraint it generates is correspondingly weak once the use-as-means factor is absent. This is why, although in the first case, you are doing the harm, your action is still intuitively permissible.

This comparison shows that some of the moral work we attribute to the doing/allowing distinction might be better understood as coming from an independent means-based constraint. In such cases, once that factor is absent, the remaining doing/allowing difference is relatively small.

Another comparison can be made between cases in which no person is used as a means. Still, there might be morally significant differences. To see this point, consider another case:

> Power Station: A trolley is heading towards five people. You can stop it by detonating the power station's control unit, but an innocent bystander is standing next to the unit and would be killed in the explosion.

I believe, in this case, the moral constraint against harming the bystander is intuitively stronger than the one in the first case above. I believe the difference can be explained as this. In the first case, the innocent person on the side track is already in a dangerous situation. We can imagine that she is trapped or stuck to the track, which is dangerous. There could be trolleys coming even if this particular trolley is not heading towards her. In Power Station, however, the bystander is not in any dangerous situation until you decide if you will detonate the power station's control unit. You actively involve her into a dangerous situation. Or put it the agent neutral way, the bystander is forced into a dangerous situation. Similar principles and ideas are also discussed and defended (e.g. Bennett 1995; Kamm 2007).

Having examined the two comparisons, I believe that the hybrid justification can preserve the moral significance of the doing/allowing distinction without overestimating its scope and strength. In cases of MRTs, first, harming the MRT is not using the MRT as a means. It is unlike the case in which the fat person's body is used as a means. Also, in these cases, harming the MRT does not involve forcing the MRT into a new dangerous situation. Both the MRT and the potential victim are already in a dangerous situation, and this is why we are discussing these cases as there is unavoidable harm which needs to be distributed. Thus, we can say that the doing/allowing distinction provides a moderate moral constraint on harming the MRTs.

This view thus helps the hybrid justification steer between the extremes. It avoids the total prohibition that would result from treating all acts of doing harm as strongly constrained, and it avoids the complete permission that would result from denying any difference between doing and allowing. In doing so, it better supports the account's overarching aim: to capture the moral ambivalence in MRT cases by showing why both permissibility and impermissibility can be reasonable judgments, depending on the specific moral structure of the case.

7. Conclusion

In conclusion, I have rejected the simple account of liability which holds that the more responsible party is liable to bear all the cost in self-defence. Based on Bazargan's view, I developed my own version of the hybrid justification, which avoids several problems faced by the simple account. I also demonstrated that the hybrid justification has unique advantages as it captures the idea of the ideal distribution of harm. Finally, I defended my version of the hybrid justification against objections and showed it can adequately explain why people’s intuition diverges in various cases.

Concluding Remarks

This thesis has examined the ethics of defensive harm by treating self-defence as a problem of allocating and reallocating unavoidable harms under non-ideal conditions. The central claim is that defensive harming is justified when, and because, it realises—or most closely approximates—the just distribution of harm. On this basis, the thesis developed (i) an account of the nature of liability as discounting the moral disvalue of harms to responsible threateners, (ii) a graded account of the grounds of liability keyed to moral responsibility, and (iii) a hybrid framework to explain persistent divergence of intuitions in minimally responsible threat cases.

Chapter 1 analysed the nature of liability. It argued that rights-forfeiture, rights-enforcement, and enforceable-duty theories are inadequate: they suffer from incompleteness/redundancy and cannot accommodate proportionality, necessity, or scalar liability. The chapter developed the discounting account, on which liability consists in a reduction of the moral disvalue of harms borne by threateners in virtue of responsibility-relevant features. This structure directly supports proportionality (permissions expand as responsibility and threatened severity increase) and necessity (among equally effective means, choose the option with least discounted moral cost).

Chapter 2 addressed the grounds of liability. It argued that culpability, pure causation, moral-status, and duty-based grounds each misfire in characteristic ways. It then proposed a revised moral-responsibility model according to which degrees of liability track three variables: (i) voluntary initiation or maintenance of a risk-imposing activity, (ii) evidence-relative awareness of non-trivial risk to others, and (iii) causal proximity to the impending harm. This graded structure yields stable rankings across contested cases (e.g., culpable attackers, justified-mistaken defenders, careful drivers, coerced agents, bystanders) and integrates with the discounting account to generate determinate proportionality and necessity verdicts without sharp epistemic thresholds.

Chapter 3 examined minimally responsible threats and reconstructed a hybrid strategy. Retaining a proportional, responsibility-sensitive account of what an agent is liable to, it rejected both a rights-forfeiture reading and any second responsibility-grounded “discount” at the lesser-evil stage. On the proposed hybrid explanatory account, liability itself is the responsibility-indexed discount; any further harm is assessed by ordinary lesser-evil reasoning subject to familiar constraints (including the doing/allowing distinction and means-based restrictions). The chapter showed how this framework (a) explains divergent intuitions about killing minimally responsible threats, (b) handles multi-threat scenarios, and (c) clarifies the role of an ideal distribution of harm and the appropriateness of compensation where justified outcomes remain non-ideal.

Across the chapters, the thesis offers a unified picture: self-defence permissions emerge from minimising and fairly allocating discounted moral costs. This recasts proportionality and necessity as consequences of a single distributive aim, explains why intuitions diverge in hard cases, and supplies a responsibility-sensitive structure that improves on binary rights-loss and duty-enforcement models.

Bibliography

Bazargan, Saba. 2014. “Killing Minimally Responsible Threats.” *Ethics* 125(1): 114–136.

Bennett, Jonathan. 1995. *The Act Itself*. Oxford: Clarendon Press.

Draper, Kai. 2009. “Defense.” *Philosophical Studies* 145(1): 69–88.

Draper, Kai. 2016. *War and Individual Rights: The Foundations of Just War Theory*. Oxford: Oxford University Press.

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. “Forfeiture and Self-Defense.” In *The Ethics of Self-Defense*, edited by Christian Coons and Michael Weber, 199–220. Oxford: Oxford University Press.

Frowe, Helen. 2014. *Defensive Killing*. Oxford: Oxford University Press.

Frowe, Helen. 2022. “Risk Imposition and Liability to Defensive Harm.” *Criminal Law and Philosophy* 16: 511-524.

Gordon-Solmon, Kerah. 2018. “What Makes a Person Liable to Defensive Harm?” *Philosophy and Phenomenological Research* 97(3): 543–567.

Kamm, Frances. 2007. *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*. Oxford: Oxford University Press.

Lang, Gerald. 2014. “Why Not Forfeiture?” In *How We Fight: Ethics in War*, edited by Helen Frowe and Gerald Lang, 127–144. Oxford: Oxford University Press.

Lazar, Seth. 2009. “Responsibility, Risk, and Killing in Self-Defense.” *Ethics* 119(4): 699–728.

McMahan, Jeff. 1994. “Self-Defense and the Problem of the Innocent Attacker.” *Ethics* 104(2): 252–290.

McMahan, Jeff. 2005. “The Basis of Moral Liability to Defensive Killing.” *Philosophical Issues* 15: 386–405.

McMahan, Jeff. 2009. *Killing in War*. Oxford: Oxford University Press.

McMahan, Jeff. 2011. “Who Is Morally Liable to Be Killed in War.” *Analysis*, 71(3): 544–559.

McMahan, Jeff. 2017. “Liability, Proportionality, and the Number of Aggressors.” In *The Ethics of War: Essays*, edited by Saba Bazargan and Samuel C. Rickless, 105–133. Oxford: Oxford University Press.

Nozick, Robert. 1974. *Anarchy, State, and Utopia*. New York: Basic Books.

Otsuka, Michael. 1994. “Killing the Innocent in Self-Defense.” *Philosophy & Public Affairs* 23(1): 74–94.

Otsuka, Michael. 2016. “The Moral-Responsibility Account of Liability to Defensive

Killing." In *The Ethics of Self-Defense*, edited by Christian Coons and Michael Weber, 121–140. Oxford: Oxford University Press.

Quinn, Warren. 1989. "Actions, Intentions, and Consequences: The Doctrine of Double Effect." *Philosophy and Public Affairs* 18: 334-51

Quong, Jonathan. 2012. "Liability to Defensive Harm." *Philosophy & Public Affairs* 40(1): 45–77.

Quong, Jonathan. 2020. *The Morality of Defensive Force*. Oxford: Oxford University Press.

Quong, Jonathan. 2022. "The Morality of Defensive Force: Replies to Otsuka, Frowe, Fabre, and Burri." *Criminal Law and Philosophy* 16: 555–574.

Renzo, Massimo. 2017. "Rights Forfeiture and Liability to Harm." *Journal of Political Philosophy* 25(3): 324–342.

Tadros, Victor. 2012. "Duty and Liability." *Utilitas* 24(2): 259–277.

Tadros, Victor. 2016. "Causation, Culpability, and Liability." In *The Ethics of Self-Defense*, edited by Christian Coons and Michael Weber, 141–168. Oxford: Oxford University Press.

Tadros, Victor. 2020. *To Do, To Die, To Reason Why: Individual Ethics in War*. Oxford: Oxford University Press.

Thomson, Judith Jarvis. 1985. *The Trolley Problem*. *Yale Law Journal* 94: 1395–1415.

Thomson, Judith Jarvis. 1991. “Self-Defense.” *Philosophy & Public Affairs* 20(4): 283–310.