Mistaken Defense and the Unbundling of Rights

Abstract

At the heart of the ethics of war and defense is the project of developing a theory of *liability*: a theory of when and why (apparent) attackers forfeit rights to not be harmed. This essay contributes to this project by developing and defending a heterodox answer to a serious and long-standing challenge to the project — what I call the *Challenge of Merely Apparent Attackers*. I argue that our standard conception of forfeiture is too coarse-grained to adequately answer this challenge, and that we need to distinguish between the forfeiture of one's rights against harm and the forfeiture of the contingent, moral "perks" of those rights. Appreciating this distinction helps us answer the challenge without undermining our ability to make sense of the so-called *necessity* and *proportionality* constraints on defensive harm, and without generating perverse incentives or paradoxical results for defenders operating under conditions of uncertainty. Appreciating this distinction may also bear fruit in other domains of moral philosophy.

1 Introduction

Suppose you need a liver transplant. You'll die without one. But the waiting list is too long, and the only way you can get a liver is if you kill your next-door neighbor and steal theirs.

Or suppose you are attacked by a gunman, and the only way to protect your life is to grab an innocent bystander and use them as a human shield, resulting in their death.

It would be wrong to kill either your neighbor or the bystander in these circumstances. It is normally wrong to intentionally harm others, even if by harming them you can prevent comparable harm to yourself. But there are exceptions to this rule. And one of the most well-recognized exceptions is the case of self-defense. Consider a paradigm example:

> *Paradigm*. A villain attempts to murder you. He will succeed unless you kill him first.

You may not steal your neighbor's liver or use a bystander as a human shield to save your life. But you may kill Villain to save your life.

The difference is that only Villain has made himself *liable* to be harmed — that is, he has done something to lose some of his normal rights against harm.[1] Your neighbor and the bystander have done no such thing. They retain their normal rights against harm. But no such rights stand in the way of harming Villain — hence your permission to harm him.

Liability is *the* central ingredient in most justifications for the use of defensive force.[2] And so a theory of liability must be at the center of our

[1] As I'll use terms, for a person to be liable to some harm just is for them to lack their normal rights to not be so harmed without their consent.

[2] I say 'most' because some harm impositions are *permissible rights-infringements*, as when you redirect the trolley away from killing five people and onto a path where it

theory of the ethics of defense. If we want a theory of defense, we need a theory of when and why someone loses rights against harm.

So what is it that makes someone like Villain liable to be harmed? On the face of it, it may seem to have something to do with the fact that *he would harm someone if he isn't harmed him first*. Indeed, it is widely thought that a person is only liable to harm that is necessary to prevent harm to someone else. There is, however, a serious and long-standing problem with this idea, which is that there are compelling cases where a person seems liable to harm even though they pose no actual threat to anyone.[3] Suppose someone attempts to kill you, but unbeknownst to either of you their gun is jammed.[4] Or suppose someone convincingly *pretends* to attempt to kill you, so as to give you the scare of your life.[5] Neither the "futile attempter" nor the "bluffer" pose any genuine threat of harm. And yet it sure *seems* they are liable to be harmed.

Cases like these pose a serious challenge for the theory of liability. We need a theory that can account for our judgments in such cases — but that does not at the same time *over-generate* cases of liability. I'll call this challenge the *Challenge of Merely Apparent Attackers*. I think it's a challenge of central importance to the theory of liability, not least because the relevant features of these cases are far from exotic. On the contrary: in the real-world, defensive agents very often operate under conditions of false information and belief. Real-world defensive agents are frequently mistaken about all sorts of things: about the threat others pose, about

---

will break one person's foot. Some theorists also claim that so-called "agent-relative-prerogatives" can sometimes justify harming people who are not liable to harm. See, for example, Jonathan Quong, *The Morality of Defensive Force* (Oxford: Oxford University Press 2020): 58–96.

[3]See, for example, Alberico Gentili, *De Jure Belli (On the Law of War)*, trans. John C. Rolfe (Clarendon Press 1933)(1598): Bk. I, ch. 14, p. 62–63.

[4]This case is inspired by Helen Frowe's "Apparent Murderer" case from *Defensive Killing* (Oxford 2014): 85.

[5]This case is inspired by Kimberly Kessler Ferzan's case by the same name from "The Bluff: The Power of Insincere Actions," *Legal Theory* 23 (2017):169.

what defensive options they have available, and about the potential consequences of those options. As such, an answer to the Challenge is an important part of an ethics of defense that speaks to the conduct of defensive agents in the real world — be they soldiers, police, or private individuals.

As I'll show in this paper, it's also an important challenge because of what we learn about the the nature of rights and of rights forfeiture in answering it. To date, the most focused attempts to answer this challenge all agree that merely apparent attackers are liable to harm in the same sense that actual attackers are liable to harm, but on the basis of different grounds.[6] I agree that the challenge shows the need for a more nuanced theory of forfeiture. But in what follows I defend an answer to the challenge where the nuance concerns, not the *grounds* of forfeiture, but the *contents* of forfeiture. I'll argue for the surprising claim that merely apparent attackers are *not* liable to defensive harm, but that we can nonetheless explain why some (and only some) merely apparent attackers bear many of the usual "upshots" of liability. An important lesson to emerge is that we need to rethink the nature of rights and of rights forfeiture. Rights can, at times, float free of some of their typical moral upshots; and forfeiture is a more fine-grained phenomenon than we have previously noticed.

Here's an overview. §2 more carefully sets out the Challenge of Merely Apparent Attackers. §3 summarizes the orthodox answer to the Challenge. §4 unpacks the most serious shortcomings of the orthodox answer. §5 develops my alternative answer to the Challenge (the Partial Forfeiture account), and distinguishes two explanatory paths to such a view. §6 anticipates two worries for Partial Forfeiture. §7 extolls the comparative advantages of Partial Forfeiture over the orthodox approach. §8 offers

[6]Ferzan, "The Bluff"; Frowe, *Defensive Killing*: 85-86; Renée Jorgensen, "The Moral Grounds of Reasonably Mistaken Self-Defense," *Philosophy and Phenomenological Research* 103 (2021):140-156; Jeff McMahan, "Who is Morally Liable to be Killed in War?" *Analysis* 71 (2011): 555-556.

concluding remarks.

2 The Challenge of Merely Apparent Attackers

Consider two cases, one involving a "futile attempter" and the other a "bluffer":

> *Jam*. Jammy intends to murder Defender. He points a gun at her. Defender reasonably believes her life is in imminent danger and knows she can prevent Jammy from pulling the trigger only by breaking his leg. Unbeknownst to anyone, however, Jammy's gun is irreparably jammed.
>
> *Bluff*. Bluffer decides to play a very ill-conceived prank on his workplace manager, Defender. He brings to the office an unloaded gun, points the gun at Defender, and yells, "Time to die!" Defender reasonably believes her life is in imminent danger and knows she can prevent Bluffer from pulling the trigger only by breaking his leg.

Jammy and Bluffer pose no actual threat to Defender; she would not be harmed were she to abstain from harming Jammy or Bluffer. And yet: Jammy and Bluffer sure *seem* liable to have their legs broken by Defender. At least they bear the usual *upshots* of liability. Let me explain. When someone has a right to not be harmed, she is typically permitted to defend against threats to that right by force (so too are third-parties), and if her right is in fact violated, she typically has the standing to complain and is owed compensation. We see this clearly in the Paradigm case. Your rights are threatened and you can fight back against Villain, others can fight back on your behalf, and you would be owed compensation and have the standing to complain were Villain to succeed in injuring you.

But now consider the position of Villain — the position of someone who *has* made himself liable to harm. Suppose you fight back against him.

Is he likewise permitted to fight back — to engage in counter-defense? Certainly not. Furthermore, a bystander is not permitted to fight back against you on Villain's behalf. And if you do injure Villain in self-defense, you don't owe him any compensation and he has (at best) very little standing to complain.

So these are four typical upshots of liability. If you are liable to have some harm imposed on you, then it will typically be true that:

(U1): You would not be permitted to fight back against that harm.[7]

(U2): Others would not be permitted to fight back on your behalf.

(U3): You would not be owed compensation if you were to suffer that harm.

(U4): You have little or no standing to complain about that harm.[8]

Jammy and Bluffer bear all four of these upshots. Neither they nor third-parties would be permitted to fight back against Defender (they wouldn't be permitted, for instance, to break Defender's leg to prevent her from breaking theirs), and if Defender breaks their legs, they have no claim to compensation and little or no standing to complain.

The lesson of Jam and Bluff, then, is that a person can bear these upshots of liability even if they pose no actual threat — even if they are what I'll call a *merely apparent attacker*.[9] A theory of liability should explain why.

[7] At least not with anything like comparable or greater force.

[8] Though it is common for theorists to claim that a person who is liable to harm has *no* standing to complain about that harm (e.g., McMahan, *Killing in War*, p. 8.), I want to leave to leave the door open to the possibility that the standing to complain may come in degrees, and that different forms of (apparent) aggression may compromise this standing to different degrees. This is my reason for making the weaker claim that liable persons have "little or no" standing to complain.

[9] As I'll use terms, A is a merely apparent attacker just in case (i) he poses no actual threat, but (ii) someone, B, believes that A poses an actual threat of harm, or B's evidence gives her sufficient reason to believe that A poses an actual threat.

But we also need a theory that can properly discriminate between cases where merely apparent attackers bear these upshots of liability and cases where they do not. Jam and Bluff are two cases where merely apparent attackers bear these upshots. Here are two cases where the merely apparent attacker clearly does not:

> *False Testimony*. Testifier wants to see Innocent harmed by Defender, and so he lies to Defender: he tells her that Innocent is about to kill her. Defender has excellent reason to trust Testifier's testimony, and so she attacks Innocent, breaking his leg.[10]

> *Evil Twin.* While on a road trip, Twin's engine overheats. He walks to the nearest town and enters the first mechanic shop he comes across. Unbeknownst to Twin, however, he has an evil twin brother who has just escaped from prison. Authorities have warned locals that the escapee will shoot anyone he comes across on sight. Reasonably believing Twin to be the murderer and believing himself to be in imminent danger, the mechanic, Defender, lunges at Twin with a crowbar, breaking his leg.[11]

As in Jam and Bluff, the defender in False Testimony and Evil Twin attacks someone who merely appears to pose a wrongful threat. But where Jammy and Bluffer bear the upshots of liability, Innocent and Twin do not. They (or a third-party) would be permitted to fight back against Defender, they are owed compensation for their broken legs, and they have the standing to complain about their injuries.[12]

[10]This case is inspired by Michael Otsuka's "Dignitary" case from "Killing the Innocent in Self-Defense," *Philosophy and Public Affairs* 23 (1994): 91.

[11]This case is inspired by McMahan's "Mistaken Resident" case from "Basis of Moral Liability": 387, and Quong's "Mistaken Attacker" case from *Defensive Force*: 23.

[12]This judgment appears to be widely, though not universally, shared in the literature.

So here's the *Challenge of Merely Apparent Attackers*: to provide a theory of liability that can capture the intuitive differences between characters like Jammy and Bluffer (on the one hand) and characters like Innocent and Twin (on the other) — a theory that explains why the former bear the upshots of liability, but not the latter.

3 The Orthodox Answer

The most focused attempts to answer the Challenge of Merely Apparent Attackers are all versions of the following, very natural idea:

> Orthodoxy. Merely apparent attackers bear the upshots of liability when and because they are, in fact, liable to be harmed — i.e., because they have forfeited rights to not be harmed.

Every version of Orthodoxy on offer attempts to make sense of liability for merely apparent attackers by proposing a *disjunctive* account of the conditions for liability. Thus, for example, Kimberly Kesslar Ferzan argues that what makes an *actual* attacker liable to be harmed is that they are *culpable* for the fact that they would otherwise violate someone's right not to be harmed.[13] What makes characters like Jammy and Bluffer liable, of course, is not that they are culpable for posing an *actual* threat of wrongful harm. Rather, they are liable because they are culpable for *appearing* to pose such a threat in the sense that they are culpable for the defender's (reasonable) *belief* that they pose a such a threat.[14] So on Ferzan's theory, what makes a person liable to (apparent) defensive harm is that either (i) they are culpable for the fact that someone's rights would otherwise be violated, or (ii) they are culpable for the fact that the defender (reasonably)

One partial exception is Larry Alexander, who appears — in the Evil Twin case — to take the view that Twin would be permitted to fight back, but that a bystander would *not* be permitted to fight back on Twin's behalf. See his "Recipe for a Theory of Self-Defense," in Christian Coons & Michael Weber (eds.), *The Ethics of Self- Defense* (Oxford University Press 2016): 29.

[13] "The Bluff": 173.

[14] Ibid: 172.

believes someone's rights would otherwise be violated. On this view, the difference between Jammy/Bluffer and Innocent/Twin is that only the former are culpable, or at fault, for Defender's belief that harming them is necessary to defend herself.

Jeff McMahan and (separately) Helen Frowe propose an account that is roughly analogous to Ferzan's in structure, but that replaces the central notion of *culpability* with the weaker notion of *responsibility*. On this view, what makes a person liable to (apparent) defensive harm is that either (i) they are responsible for the fact that someone's rights would otherwise be violated, or (ii) they are responsible for the fact that the defender believes (or has sufficient reason to believe) that someone's rights would otherwise be violated.[15] On this view, the difference between Jammy/Bluffer and Innocent/Twin is that only the former bear sufficient responsibility for Defender's evidence or belief that harming them is necessary to defend herself.[16]

Finally, a third version of ORTHODOXY comes courtesy of Renée Jorgensen.[17] Jorgensen does not purport to offer a complete theory of liability, but rather an addendum that is meant to be compatible with many different theories of liability in actual-threat cases. So for Jorgensen, there's the usual conditions for liability that we find satisfied in actual-threat cases — *whatever exactly those might be*. But then there's a second path to liability. Instead of meeting the usual conditions for liability, you might make yourself liable by engaging in behavior that conventionally *signals* that you meet the usual conditions for liability. More precisely: on Jorgensen's account, what makes a person liable to (apparent) defensive

[15]Frowe, *Defensive Killing*: 85-86; McMahan, "Who is Morally Liable?": 555-556. Ferzan takes "the relevant moral responsibility for forfeiting rights" to be the merely apparent attacker's responsibility for the defender's *belief* that the former poses an actual threat. McMahan, by contrast, focuses on the merely apparent attacker's responsibility for the defender's *evidence*.

[16]The above formulation glosses over the fact that, on this account, it is *comparative* responsibility that matters for liability. More on this in §5.1.

[17]Jorgensen, "Reasonably Mistaken Self-Defense," p. 140-156.

harm is that either (i) they meet the usual conditions, C, for liability (i.e., the conditions specified by the correct theory of liability in paradigmatic, actual-threat cases), or (ii) they have performed an action that conventionally *signals* to others that p, where conditions C would be satisfied if p were true.[18] On this view, the difference between Jammy/Bluffer and Innocent/Twin is, roughly, that only the former engage in behaviors that conventionally signal aggression.

The above theorists give different answers to the Challenge of Merely Apparent Attackers. But what they all agree on is that, although characters like Jammy and Bluffer are not liable for the same reasons that actual attackers are liable, they are liable nevertheless. They forfeit rights against harm.

4 Problems for Orthodoxy

I have serious concerns about this basic idea. For the sake of space, I won't discuss any concerns that are specific to particular versions of Orthodoxy. Instead I'll present four concerns that generalize to all of the orthodox theories surveyed above.

4.1 The Problem of Necessity

Consider a paradigm case of unnecessary harm:

> *Easy Defense*. A conscientious driver loses control of his vehicle. Defender knows that she has two equally good and easy ways to prevent herself from being killed by the car: she can redirect the car to the right, in which case Driver will be killed,

[18] To illustrate, let's plug a simple Thomsonian account of liability in to the Jorgensen schema. Judith Jarvis Thomson claims that an actual attacker is liable when and because he would violate someone else's rights if he were not harmed himself ("Self-Defense," *Philosophy and Public Affairs* 20 (1991): 283-310). Plugged into the Jorgensen schema, the resulting theory would claim that a merely apparent attacker makes himself liable when and because he conventionally signals that p, such that if p were true, then he would violate someone else's rights if he were not harmed himself.

> or she can redirect the car to the left, in which case Driver will suffer a sprained wrist.[19]

It is impermissible for Defender to kill Driver under these circumstances, and she would wrong him by doing him. It's widely held that this is because it isn't "necessary" for Defender to kill Driver to defend herself, and because

> Necessity Limits Liability. A person who makes themselves liable to harm only makes themselves liable to as much harm as is *necessary* to prevent harm elsewhere.[20]

This principle, however, is straightforwardly inconsistent with Orthodoxy. Characters like Jammy and Bluffer pose no actual threat. And so of course it isn't necessary for Defender to harm them in order to prevent harm to herself (or to anyone else). Orthodoxy claims that such characters are liable to be harmed by Defender, but Necessity Limits Liability implies they are not.

What Orthodoxy needs, then, is an alternative way of accounting for the wrongfulness of killing someone like Driver in a case like Easy Defense — an explanation that does not depend on the claim that liability is restricted to necessary harm. Now it might seem that this isn't so hard to find. One might think that the problem with killing Driver isn't that

[19]This case is a twist on McMahan's case of the conscientious driver in "Basis of Moral Liability": 393.

[20]This principle is widely, but not universally, endorsed in the defensive harm literature. For some defenses of this claim, see David Clark, "The Demands of Necessity," *Ethics* 133 (2023): 473-496; Kaila Draper, "Necessity and Proportionality in Defense," in *The Ethics of Self-Defense*, eds. Christian Coons and Michael Weber (Oxford: Oxford University Press, 2016): 174-175; McMahan, "The Limits of Self-Defense," in *The Ethics of Self-Defense*, eds. Christian Coons and Michael Weber (Oxford: Oxford University Press, 2016): 185–210; and Quong, *Defensive Force*: 124-150. A notable dissenter is Frowe, *Defensive Killing*: 88-119. Frowe argues that unnecessary force is *pro tanto* wrongful, but that it does not violate the rights of its target.

it's not necessary, but rather that it doesn't *appear* necessary. That is, we might explain the case of Easy Defense by appeal, not to the principle that liability is restricted to necessary harm, but to the principle that *liability is restricted to harm that (reasonably) appears necessary*.[21] This reformulation of the necessity constraint is at least consistent with the thought that Jammy and Bluffer are liable to harm, since harming them does at least appear to Defender to be necessary.

But this alternative principle is untenable. Consider:

> *Two Snipers*. There are two snipers hidden in a belltower, each attempting to kill Defender. Defender only spots Sniper_1, and reasonably believes he is the only person in the belltower. She knows that she can save her life only if she throws a grenade into the belltower (she doesn't know that this is also necessary to defend against Sniper_2). She throws the grenade, wounding both snipers.

Sniper_2 is plainly liable to this harm, even though it doesn't appear, to Defender, to be necessary to harm him. So it looks like the principle that *liability is restricted to harm that appears necessary* won't do.

We might instead try a weaker, disjunctive principle: *liability is restricted to harm that either is necessary or that (reasonably) appears to the defender to be necessary*. But this principle also runs into trouble. Consider:

> *Two Snipers Redux*. As before, there are two snipers hidden in a belltower, each attempting to kill Defender. Defender only spots Sniper_1, and reasonably believes he is the only person in the belltower. She knows that she can save her life only if

[21] A formulation of the necessity constraint along these lines is suggested by Seth Lazar, "Necessity in Self-Defense and War," *Philosophy and Public Affairs* 40 (2012): 13.

she disarms him. She can do so by throwing a grenade into one of two windows. If she throws the grenade into the first window it will break the legs of Sniper_1 and Sniper_2. If she throws the grenade into the second window it will break the legs only of Sniper_1. As Defender is unaware of the presence of Sniper_2, she sees no reason to prefer one window to the other. She chooses to throw the grenade into the first window. Both snipers are wounded. As it turns out, Sniper_2's gun was jammed and he posed no actual threat.

Sniper_2 at least bears the upshots of liability. He would not be permitted to kick the grenade back towards Defender, bystanders would not be permitted to kick the grenade back on his behalf, and he is not owed compensation for his injuries. Orthodoxy is committed to explaining these upshots of liability by attributing genuine liability. But notice: it was neither necessary for Defender to harm Sniper_2 (the option to throw the grenade into the other window was available to her), nor did it *appear* to Defender to be necessary to harm him (she doesn't have any reason to believe that he's a threat; she doesn't even know he's there). So Orthodoxy can't help itself to the disjunctive principle either.

Now perhaps there is yet some way for Orthodoxy to have its cake and eat it too. There may yet be some alternative precisification of the necessity constraint I've not considered that can do the job. But there is at least a real burden here for Orthodoxy. It requires seriously rethinking the nature of the necessity constraint.

4.2 The Problem of Proportionality

The so-called proportionality constraint likewise generates trouble for Orthodoxy. Consider:

> *Overkill*. Once again, a conscientious driver loses control of his car, and it careens towards Defender. Defender knows it

> will merely bruise her big toe unless she redirects the car. She can only redirect the car by shooting Driver dead.

It would be impermissible for Defender to shoot and kill Driver. As shooting Driver is Defender's *only* defensive option, the problem isn't that killing Driver is unnecessary. The problem is that it is disproportionate. Killing Driver when he poses only a threat of bruising Defender's toe would wrong Driver because lethal force would be disproportionate and

> Proportionality Limits Liability. A person who makes themselves liable to harm only makes themselves liable to harm that is proportionate to the threat averted thereby.

This principle, too, is incompatible with Orthodoxy. Orthodoxy claims that characters like Jammy and Bluffer are liable to be harmed. But they pose no threat. Harming them thus averts no threat, and so harming them can't possibly be proportionate to their threat.[22]

Again, we might look for alternative precisifications of the proportionality constraint that deliver alternative explanations of the wrongfulness of killing Driver in a case like Overkill. Perhaps, we might think, liability is not restricted to proportionate harm, but rather to harm that (reasonably) *appears* to be proportionate.

This principle, however, runs into the very same trouble as the appearance-centric version of the necessity constraint, as illustrated by the case of Two Snipers. Defender sees both snipers, but falsely believes that only Sniper_1 poses a threat. She throws a grenade that wounds both snipers, and it

[22]To quote Lisa Hecht: "There is no wrongful threat that can be used as yardstick for determining proportionality" ("Provocateurs and Their Rights to Self-Defence," *Criminal Law and Philosophy* 13 (2019): 169.) In that quote she is discussing provocateurs who are not apparent threats, but the point applies equally well to the case of merely apparent attackers.

turns out that throwing the grenade was (in fact, and contrary to appearances) necessary and proportionate to avert Sniper_2's actual threat. Surely, then, Sniper_2 is *not* wronged, despite the fact that his injuries did not *appear* to Defender to be proportionate to his threat.

Again we might try a weaker, disjunctive principle: *liability is restricted to harm that either is proportionate or that (reasonably) appears to be proportionate*. But again we have parallel troubles as with the disjunctive version of the necessity constraint, as illustrated by Two Snipers Redux. In that case, recall, harming Sniper_2 is neither proportionate in fact nor in appearance. But given that he bears the upshots of liability, Orthodoxy is committed to declaring him liable to the harm he suffers.

Again, there may yet be some alternative precisification of the proportionality constraint I've not considered that can do the job. But again the point is that there is a real burden here for Orthodoxy. It requires seriously rethinking the nature of the proportionality constraint.[23]

4.3 The Problem of the Lack of Information as a Ground of Liability

A very curious feature of Orthodoxy is that it makes one person's lack of information a ground of another person's liability to be harmed. To illustrate, compare the original case of Bluff with this variant of the case:

Transparent Bluff. As in the original Bluff case, except that

[23]It's worth noting that the Problem of Necessity and the Problem of Proportionality are problems that generalize to *any* theory that attributes liability to some people who pose no actual threat. Consider, for example, alternatives to Orthodoxy according to which Jammy, but not Bluffer, is liable to harm —perhaps because only Jammy intends or attempts harm, or perhaps because only Jammy performs an action of a risk-imposing type. (I am grateful to an anonymous reviewer at *Ethics* for suggesting such a theory.) This kind of theory — assuming it grounds Jammy's liability in facts about Jammy rather than in facts about Defender's evidence or beliefs — would be immune to the problems I present in §4.3 and §4.4. But it would still be inconsistent with the principle that liability is constrained to necessary and proportionate harm, and would thus stand in need of some alternative principle by which to explain why Driver would be wronged if he is killed in Easy Defense or in Overkill.

Defender learns that Bluffer's gun is unloaded.

Bluffer is certainly not liable to be have his legs broken in this case. A person is not liable to be harmed by a defender who *knows* that harming them would be purely gratuitous. (I expect virtually all defenders of Orthodoxy would agree on this point.)

Orthodoxy, however, tells us that Bluffer *is* liable to be harmed in the original Bluff case. The only difference between Bluff and Transparent Bluff is that Defender lacks information in the latter that she possesses in the former. Thus Orthodoxy is committed to the claim that one person's lack of information can ground another person's liability to be harmed — indeed, that it can be the decisive difference between liability and non-liability.

This seems hard to believe. While it's very plausible that a person's lack of information can ground their *excuse* for harming someone, it seems wild to think that it can ground the *loss of another's rights*.[24] It's an implausible idea in the abstract, but it also has implausible implications.

Consider, for instance, the weird kinds of incentives this would generate for information sharing, as in:

> *Mixed Bluff*. As in the original Bluff, except that, although Defender reasonably believes Bluffer's gun is loaded, Observer knows the gun is unloaded.

Observer knows that Bluffer is bluffing. Suppose that she is able to pass this information along to Defender in time for Defender to make use of it. Suppose, however, that Observer is unsure how Defender will act in light of that information. On Observer's evidence, there is a 90% chance

[24] C.f., McMahan, "Basis of Moral Liability": 391.

Defender harms Bluffer if Observer does *not* pass the information along; and there is a 50% chance that Defender still harms Bluffer if Observer *does* pass the information along. Given Orthodoxy, then, Observer is faced with the following choice. Option 1: she can withhold from Defender the crucial information that Bluffer is merely bluffing, in which case she can be certain that Defender will *not* wrong Bluffer. Option 2: she can pass along the information, in which case there's a 50% chance Bluffer will be wronged (since there would be a 50% chance that Defender harms someone she knows is merely bluffing). Given the greater importance of preventing rights infringements over minimizing harm, it seems that what Observer should do in light of her evidence is to go for the first option and withhold the crucial information from Defender. But this seems wildly implausible. Surely Observer should pass along the information that the gun is unloaded. And surely her passing along this information would be better for both Defender and Bluffer and would minimize the chance that a wrong is done.

4.4 A Paradox of Aggregation

A final worry is that Orthodoxy generates paradoxical results across cases involving uncertainty. To see the trouble, compare three cases. The first:

> *Transparent Tracks.* You must choose whether to redirect a trolley down track A or track B. You can't see down either track, but (i) you know there is someone on track A who poses an actual lethal threat to an innocent victim, and (ii) you know there is someone on track B who is merely bluffing about posing a lethal threat. You don't know the identities of the persons on the tracks, but you know that sending the trolley down a track will kill the person on that track and scare off the person on the other track (thus, either way the innocent victim will be saved).

I assume defenders of Orthodoxy would agree with me that you are not permitted to kill the person you *know* to be bluffing over the person who you know poses an actual threat.[25] I take this to be a fixed point.

Before we consider the second case, we need to make a toy stipulation about Orthodoxy's criteria of liability for merely apparent attackers. Presumably, defenders of Orthodoxy are not thinking that defenders must be *certain* (or have evidence that justifies certainty) that an apparent attacker poses a wrongful threat in order for the latter to be liable to be harmed. Presumably they think something less than certainty will do. Let's suppose, just for the sake of illustration, that the view is that merely apparent attackers are liable to be harmed by a defender only if the defender reasonably believes (or has evidence to justify believing), *with credence 0.7*, that they pose a wrongful threat.

This stipulation to hand, consider:

> *Uncertain Tracks.* As in Transparent Tracks, except that you reasonably believe, with credence 0.7, that the person on track B is an actual attacker and not a mere bluffer.

Again, it seems to me that in this case you should prefer to kill the person you know is an actual threat over the person who only seems likely to you to pose a threat. There is more reason to send the trolley down track A than track B. Orthodoxy, however, doesn't seem to get us this result. According to Orthodoxy, both apparent attackers are liable to be killed. That is, neither has a right to not be killed. So rights-based considerations do not give us more reason to send the trolley down track A. But neither do interest-based considerations give us any more reason to send the trolley down track A, since both options impose the same amount of defensive harm (one liable person killed) and prevent the same amount of wrongful

[25] C.f., Ferzan, "The Bluff," p. 171.

harm (one innocent person saved). It's hard to see, then, why we should prefer to kill the known attacker over the merely-probable attacker given Orthodoxy.

I find this a troubling result in its own right. But things get much more troubling when we add a third case to the table:

> *Busy Uncertain Tracks*. Again, there are two tracks. There are 1,000 people on track A, all of whom you know to be actual attackers. There are 1,000 people on track B: you know that 700 are actual attackers and that 300 are mere bluffers (though you don't know which members of the group belong to which category). Given this fact, for each person on track A, you know that they pose an actual threat; and for each person on track B, you reasonably believe with credence 0.7 that they pose an actual threat. Sending the trolley down a track will kill all the persons on that track and scare off all the persons on the other track. (Either way, all the actual attackers will be prevented from harming anyone.)

On the one hand, Orthodoxy would seem to enjoin us to evaluate Busy Uncertain Tracks as simply a version of Uncertain Tracks writ large. Given Orthodoxy, every person on each track meets the conditions for liability. There are thus no rights-based reasons to prefer sending the trolley down one track or the other, since every person on each track meets the conditions for liability. And there are no interest-based reasons to prefer sending the trolley down one track or the other, since both options kill the same number of liable persons and save the same number of would-be victims. Given Orthodoxy, then, it seems permissible to send the trolley down either track.

But the problem is that this verdict is at odds with our verdict in Transparent Tracks. There we said that you are not permitted to kill someone

you know to be merely bluffing over killing someone you know to pose an actual threat (all else equal). But this is precisely what Orthodoxy permits you to do in Busy Uncertain Tracks. By sending the trolley down track B you *know* you would be killing 300 bluffers in place of 300 actual attackers.

The fact that Orthodoxy leads to paradox in a case like Busy Uncertain Tracks is due to the fact that it makes liability sensitive to a defender's evidence such that there is some probabilistic threshold (< 1.0) above which merely apparent attackers can become liable. This has paradoxical implications when we consider cases where a defender clears that probabilistic threshold with respect to her information about any given apparent attacker, but falls below that threshold with respect to her information about the apparent attackers in the aggregate.

5 Rethinking Forfeiture

In the remainder of this essay I develop and defend an alternative answer to the Challenge. I argue that we needn't commit ourselves to the claim that Jammy and Bluffer forfeit rights against harm in order to explain why they bear the upshots of liability. We can explain why they bear those upshots even if they maintain their rights against harm.

Indeed, I think it's a mistake to conflate the kind of rights forfeiture that actual attackers undergo with the kind of forfeiture that merely apparent attackers undergo; I think that merely apparent attackers forfeit *only some* of the moral perks that actual attackers forfeit. Their moral status falls somewhere between that of an actual attacker and an innocent bystander. Specifically, I will defend a thesis I call

> **Partial Forfeiture**. When merely apparent attackers bear the upshots of liability, this is not because they forfeit their rights to not be harmed. They maintain their rights against

> harm despite forfeiting many of the contingent moral upshots of those rights — upshots that include their permission to defend those rights and have those rights defended by others, their right to compensation, and their standing to demand that they not be harmed.

I want to be clear about what is being claimed. There are a number of ways in which theorists have previously noted that forfeiture can be "partial". Virtually all theorists agree that forfeiture can be partial in the sense that one can make themselves liable only to certain amounts of harm (e.g., only to proportionate and necessary harm). And some theorists think forfeiture can be partial in the sense that you can be liable to be harmed by some people and not others,[26] or liable to be harmed in service of some goals but not others.[27] What I'm advancing in this paper is the idea that forfeiture can be partial in a further, very different sense: one can forfeit some, but only some, of the usual perks of a right.

In this section, I will present two different paths to Partial Forfeiture that issue from two different general approaches to the theory of liability.

5.1 Liability by Distributive Fairness

Numerous advocates of the very popular "responsibility" theory of liability (canvassed above in §3) motivate the theory by appeal to a deeper explanatory story about *why* responsibility for an (apparent) threat grounds liability — one that appeals to a responsibility-sensitive theory of distributive justice. So, for example:[28]

[26]See, for example, David Clark, "Refusing Protection," *Philosophy and Public Affairs* 51 (2022): 33-59; and McMahan, "Limits of Self-Defense": 201-203.

[27]See §5.2 of the present essay, and McMahan, *Killing in War*, 8-9.

[28]To the best of my knowledge, Phillip Montague is the first to defend the idea that liability to defensive harm is a function of distributive fairness in his "Self-Defense and Choosing between Lives," *Philosophical Studies* 40 (1981): 207–19 (though he does not use the term 'liability').

> The determination of liability to defensive harm is a matter of justice in the *ex ante* distribution of unavoidable harm. — *Jeff McMahan*[29]
>
> In its appeal to distributive fairness, the responsibility account of liability is the parallel, in the domain of preventive justice, to luck egalitarianism in the domain of distributive justice. — *Kerah Gordon-Solmon*[30]

I will call the common core of this story:

> **Liability by Fairness**: What makes someone liable to some harm is the fact that he would suffer that harm under the most fair distribution of unavoidable harm that is available to the defender.

As comes out in the Gordon-Solmon quote above, many proponents of Liability by Fairness believe that it lends support to the view that persons are liable to harm when and because they are *responsible* for posing (or appearing to pose) a wrongful threat of harm. They arrive at this criterion by supplementing Liability by Fairness with certain assumptions — familiar from the luck-egalitarian literature — about the way in which facts about distributive fairness are sensitive to facts about responsibility for the benefits and burdens to be distributed. To illustrate, consider the Paradigm case of defensive harm from §1: Villain attempts to murder you; he will succeed unless you kill him first. Villain's attempt to murder you makes him responsible for creating what Phillip Montague calls a "forced choice" between harms — that is, he is responsible for putting you in a situation where harm is inevitable and you must choose between

[29] "Debate: Justification and Liability in War," *Journal of Political Philosophy* 16 (2008): 234.

[30] "What Makes a Person Liable to Defensive Harm?" *Philosophy and Phenomenological Research* 97 (2017):546.

two distributions of harm: one where you are killed and Villain is unharmed, and one where Villain is killed and you are unharmed.[31] Given the assumption that fairness is sensitive to responsibility in such a way that Villain's responsibility for creating the forced choice makes the latter distribution more fair than the former, it follows from Liability by Fairness that Villain is liable to be killed. We can represent the explanatory structure as follows:

$$\text{Responsibility facts} \xrightarrow{\text{explain}} \text{Fair distribution facts} \xrightarrow{\text{explain}} \text{Liability facts}$$

Recall from §3 that McMahan embraces a version of Orthodoxy according to which merely apparent attackers are liable to harm when and because they are responsible for appearing to pose a wrongful threat. This idea, however, seems at odds with McMahan's additional commitment to Liability by Fariness. Think about Jammy and Bluffer's situation from the perspective of Liability by Fariness. Jammy and Bluffer pose no actual threat. Since they pose no actual threat, harm is *not* unavoidable; there is no "forced choice"; a distribution is available to Defender where no one suffers harm (namely, the option where Defender abstains from breaking Jammy or Bluffer's legs). And so there are no distributive-justice grounds on which to attribute liability to Jammy and Bluffer — there only *seem* to be such grounds from Defender's perspective.[32] Liability by Fairness, then, more naturally leads, not to Orthodoxy, but to the view that Jammy and Bluffer have *not* forfeited rights against harm.

Liability by Fairness nevertheless gives us reason to insist that Jammy and Bluffer forfeit many of the normal moral advantages of those rights they maintain. That is, given Liability by Fairness, even if Jammy and Bluffer are not liable to harm, they nonetheless bear the four upshots of liability (discussed in §2). Consider these upshots in turn.

[31] "Choosing Between Lives": 209-210.
[32] C.f., Ferzan, "The Bluff": 170.

The **first upshot** of liability is the loss of one's own defensive permissions. Jammy and Bluffer clearly lose this permission: they would not be permitted to fight back against Defender with comparable or greater force. Liability by Fairness explains why this is, even on the assumption that Jammy and Bluffer maintain their rights against harm. If Defender does what she (fact-relatively) ought not to do and attacks Jammy and Bluffer, she transforms the situation from one where harm is avoidable into one where harm is now unavoidable. Now that there's a forced choice, considerations of distributive fairness become relevant. And here I think it is independently plausible that — even though Defender is the most direct cause of the fact that there is a forced choice — it is *more fair* for the harm of broken legs to fall on Jammy and Bluffer than on Defender. This is because, even though Defender is the most direct cause of the forced choice, Jammy and Bluffer nonetheless bear greater culpability and responsibility for that forced choice in virtue of their unjustly and foreseeably provoking Defender to reasonably respond with force.[33] Given Liability by Fairness, then, if Defender were to attack Jammy and Bluffer, she would not make herself liable to be harmed. This is why Jammy and Bluffer are not permitted to fight back. This is also why a *third-party* may not fight back on Jammy and Bluffer's behalf (the **second upshot** of liability).

The **third upshot** is that Jammy and Bluffer are not owed compensation if Defender breaks their legs. Consider McMahan's own claims about the correlation between the morality of defensive harm and the morality of corrective justice: "Just as we may think of liability in torts as a matter of corrective justice, or justice in the distribution of harm *ex post*, so we may think of liability to defensive action as a matter of preventive justice, or justice in the distribution of harm *ex ante*".[34] I take McMahan to be claiming that the conditions for owing compensation parallel the

[33] C.f., Hecht, "Provocateurs": 175.
[34] "Basis of Moral Liability": 395.

conditions for liability to defensive harm: X owes compensation to Y only if the most fair distribution of burdens is one where X assumes some of Y's burdens by compensating Y. Given this conception of corrective justice — which is a very natural bedfellow to Liability by Fairness — it's straightforward why Jammy and Bluffer are not owed compensation from Defender. If it was more fair that Jammy and Bluffer bear the costs of a broken leg than Defender *ex ante*, then presumably the same is true *ex post*.

Finally, the **fourth upshot**: Jammy and Bluffer have little or no standing to complain about Defender's use of force. There is some temptation to think that part of what it is to have a right against some treatment is to have the standing to complain about such treatment, and thus that one can lose the standing to complain only by losing the right.

This seems to me false, however. It seems a person can lose grounds for complaint without the loss of rights. Considerations of reciprocity, for instance, may independently undercut such grounds. Suppose I'm a pathological promise-breaker; I break almost every promise I make. For this reason, I lack much if any standing to complain if you break a promise that you made to me. But this isn't because I lack a right — and you the correlative duty — that you keep your promise.[35] (The fact that I don't keep my promises doesn't strip you of your ability to choose to put yourself under a promissory duty towards me.) Rather, my standing to complain is weakened *in spite of* my having that right. My standing is weakened because to complain would be to wrongfully hold you to standards to which I do not hold myself.[36]

[35][Redacted].

[36]It seems to me the basic problem with a hypocritical complaint like this is roughly the same problem that R. Jay Wallace identifies with hypocritical *blame* ("Hypocrisy, Moral Address, and the Equal Standing of Persons," *Philosophy & Public Affairs* 38 (2010): 326-329). In making a hypocritical complaint, I express a demand that others abide by some standard that I do not hold myself to, despite my also being subject to that standard. There is a sort of internal inconsistency in this. But the more serious problem — at least as the standing to complain is concerned — is interpersonal, not intrapersonal.

Another way to weaken one's standing to complain without losing the relevant right is by foreseeably (and without justification) inciting someone to violate that right. Suppose I'm the following sort of provocateur. I know how to *really* get under my colleague's skin. I know that Colleague will be deeply offended and angry if I mock his physical appearance, and that he'll almost certainly respond violently. So I go ahead and slip him a letter that mocks his appearance. As predicted, he responds by attacking me. This is something he shouldn't do; it's gratuitous harm that serves no defensive purpose. I may well deserve it, but it's clearly not harm to which I've made myself *liable*. And yet: I plainly have very little (if any) standing to complain about his attack, in virtue of my role in inciting him to violate my rights in this way.

But Jammy and Bluffer surely have at least as little as this to stand on to complain about Defender's use of force. They are at least as guilty as I am of inciting a violent response, and Defender has an even stronger excuse than Colleague for her use of violence. So even on the assumption that Jammy and Bluffer have a right that Defender not harm them, we should think they have little standing to complain.

To summarize: if we endorse the popular Liability by Fairness as the "deep story" of liability, it should lead us to reject Orthodoxy and insist that characters like Jammy and Bluffer are not, in fact, liable to be attacked. But Liability by Fairness also leads to the conclusion that Jammy and Bluffer nonetheless forfeit many of the non-constitutive, contingent perks that normally come with those rights. They greatly weaken the moral significance of certain rights without losing those rights entirely. In short, Liability by Fairness leads to Partial Forfeiture.

As for the sorts of merely apparent attackers that do *not* bear the upshots of

By subjecting others to demands to which I do not subject myself, there is a failure of equal respect. I am, in effect, saying that my being wronged is more serious than my wronging someone else in the same way.

liability — as in False Testimony and Evil Twin —, here is what Liability by Fairness has to say. While Jammy and Bluffer plausibly bear *more* culpability or responsibility for Defender's use of force, this surely isn't true of Innocent and Twin. Thus, the grounds of the fact that it is *more fair* for harm to be distributed onto Jammy and Bluffer than on Defender are not present in the case of Innocent and Twin. This is why, given Liability by Fairness, Innocent and Twin fail to bear the upshots of liability.

5.2 Liability by Duty

My main project in the present essay is to defend Partial Forfeiture and show its virtues as an answer to the Challenge of Merely Apparent Attackers. I have discussed the Liability by Fairness route to Partial Forfeiture because the former is a very prominent theory in the defensive harm literature. I hope to convince advocates of this theory that they would do better to embrace Partial Forfeiture over Orthodoxy.

That said, I feel obliged to air some of my reservations about Liability by Fairness. The most straightforward problem is that it's not hard to find cases where harm is unavoidable, where it would be more fair for the harm to fall on X than Y, but where X is plainly not liable to be made to suffer that harm. Consider, for example:

> *Transfusion*. Bob has had terrible luck throughout his life: he has suffered from a very serious disease numerous times, through no fault of his own. And his luck just got worse: he's come down with the disease for the third time. Alice, by contrast, has enjoyed tremendous health throughout her entire life. (Alice and Bob are indistinguishable in all other fairness-relevant respects.) Bob finds himself in a situation where he can rid himself of his disease by performing a complete blood transfusion with Alice while she is sleeping. Swapping his blood with Alice's will result in Bob being cured and Alice taking on the disease.

From the perspective of distributive fairness, it would certainly be more fair for Alice to have this disease for the first time than for Bob to have it for the third time; this would more fairly allocate the burdens of brute luck. But Alice is plainly not liable to have this disease thrust on her by Bob, even if this is the only way Bob can prevent himself from suffering it.[37]

The reason Liability by Fairness has these counterintuitive implications is that, to quote Susanne Burri, "liability to defensive harm is a distinctly *localised* affair, whereas determinations of distributive justice are sensitive to wider societal considerations."[38] This seems the right diagnosis of the problem. Distributive justice is sensitive to wider societal considerations, including facts about events unrelated to the act of aggression, precisely because fairness is sensitive to such considerations. Liability is sensitive to much narrower considerations.

The non-localized nature of distributive justice also means that Liability by Fairness has trouble explaining the extent to which liability is *goal-constrained*. As McMahan himself insists upon, a person is never liable to harm *simpliciter*. When a person is liable to be harmed, they are only liable to be harmed *as a means or side-effect of certain goals*.[39] To illustrate, consider the following pair of cases:

> *Reckless Driver 1*. Driver, in a hurry to make his dinner reservations, is driving with extreme recklessness. He loses control of his car on a patch of ice. Defender will be killed by the car unless she kills Driver by redirecting the car away from her and into a tree. Driver will be unharmed otherwise.

[37] Quong (*Defensive Force*: 7) raises this objection with a case involving the fair distribution of *risk*: "It's not permissible . . . to toss a coin to decide whether to use an innocent bystander's body as a shield against a lethal projectile, even though this distributes the risk of death equally between the bystander and yourself."

[38] "Defensive Liability: A Matter of Rights Enforcement, not Distributive Justice," *Criminal Law and Philosophy* 16 (2022): 545.

[39] McMahan, *Killing in War* (New York: Oxford University Press 2009): 9.

> *Reckless Driver 2*. As before, except that Defender's defensive options are different. Option 1: she can simply step out of the way of the car, in which case Driver will be unharmed. Option 2: Defender can redirect Driver's car, in which case his car will shield a bystander from a nearby, unrelated avalanche. If Driver is redirected into the avalanche, he will be killed and the bystander saved. The bystander will be killed otherwise.

In either case, killing Driver saves one life. And yet it clearly seems Defender may only kill Driver in the first case. The best explanation for this difference is that Driver's liability is goal-constrained. He is not liable to be killed to prevent just any loss of life. He is, in this case, only liable to be killed where killing him is necessary to prevent Defender from being killed. (In a moment I'll articulate my preferred theory for identifying the goals that constrain a person's liability to harm.)

The problem is that Liability by Fairness doesn't get us this intuitive difference between these two cases. By the lights of Liability by Fairness, it seems that Driver is liable to be harmed in both cases, since harming him in either case yields the most fair distribution of harm. It is more fair that Driver be killed than Defender. But, if we stipulate that the threat to bystander's life was due to nothing more than bad brute luck, it is also more fair that Driver (in virtue of his recklessness) be killed than bystander. Again, the non-localized nature of distributive justice renders Liability by Fairness insufficiently sensitive to the purposes for which people may be liable to be harmed.

To be clear: my aim here is not to offer a decisive rebuttal of Liability by Fairness. The main purpose of this essay is to convince the reader of the truth of Partial Forfeiture, not the falsity of Liability by Fairness. I mention my misgivings with the latter primarily to explain my motivation for not settling with just one path to Partial Forfeiture, and for suggesting a second.

The second path I propose combines certain elements of Liability by Fairness with an idea introduced by Victor Tadros. This is the idea that a person can lose certain of their claim rights by taking on certain kinds of duties. Specifically, Tadros claims that X can become liable to be made to bear certain costs as a means or side-effect of some goal in virtue of X's *having a duty to pursue that goal*.[40]

Liability by Fairness takes liability to defensive harm to ultimately come down to the question: Which distribution of the unavoidable costs is *most fair*? I propose that the question to ask, rather, is one about the duties of individuals: Who has a duty to assume (a share of) the costs that must be distributed?

The view I propose:

> **Liability by Duty**. Someone is liable to defensive harm when and because (i) there is harm that must be distributed and (ii) they have a duty to assume (a share of) that harm — that is, a duty to see that the harm does not fall on anyone else.

Here's how the idea plays out in actual-threat cases like Paradigm. When Villain attacks you, he creates a forced choice. Either Villain will be harmed or you'll be. Clearly, in this case, only Villain has a duty to assume these unavoidable costs; you are under no such duty. By harming him,

[40]Drawing from Tadros ("Responses," *Law and Philosophy* 32 (2013): 296), a possible rationale for this idea is the following. The reason it is normally wrong to harmfully use someone to benefit someone else is, fundamentally, because of the importance of our being free to set our own ends, and to choose which ends to bear costs in service of. This is at the heart of the default prohibition against using others as mere means. When we use others as a mere means, they can object that we are imposing ends on them that are not their own. But when that person has a duty to serve goal G, that duty undermines this objection. They may not have chosen to serve G, but it is nonetheless "their goal" on account of the fact that morality binds them to that duty. They lose their right to not be made to bear costs in service of G, then, because they have lost that which would normally ground their right to not be so used in the first place — namely, the objection that we are imposing ends on them that are not their own.

then, you are just making him do what he has a duty to do himself (and which you do not likewise have a duty to do) — namely, to see that the costs of his attack not fall on anyone but himself. His duty to assume the costs of the attack is what makes him liable to be *made* to assume the costs of the attack.

Again, my main purpose in this essay is not to argue for Liability by Duty over Liability by Fairness. But having aired my worries about the latter, let me briefly note how the former avoids these worries.

First, by appealing to the directed duties of a particular agent, Liability by Duty, does make liability a "localized affair." Consider a case like Transfusion. It may make for a more fair distribution of the burdens of brute luck for Alice to suffer the disease for the first time than for Bob to suffer it for a third time. Nonetheless, Alice is plainly under no duty to assume this disease from Bob. (The mere fact that it would be more fair for X to suffer a certain harm than Y does not, in general, suffice to place X under a duty to assume that harm from Y.) That is why Alice is not liable to be made to suffer the disease, per Liability by Duty.

Liability by Duty also offers a potential explanation for the goal-constrained nature of liability, in light of the fact that duties are always duties to *do something* or *to see to some goal*. Consider the two Reckless Driver cases. Driver may be killed in the first case but not the second. According to Liability by Duty, this is because only in the first case is killing Driver necessary to serve a goal that Driver, himself, has a duty to serve. He has a duty to not kill Defender — a duty stringent enough that he must be willing to accept even death if this is only way to discharge this negative duty. But his positive duty to rescue the bystander is not nearly so demanding; he would not be morally required to sacrifice something as great as his life to save the bystander.[41]

[41] [Redacted.]

Those are some comparative advantages of Liability by Duty over Liability by Fairness. But let's return to the central project at hand, which is the Challenge of Merely Apparent Attackers. How should we diagnose cases like Jam and Bluff from the perspective of Liability by Duty? Firstly and most obviously, Liability by Duty joins Liability by Fairness in implying that Jammy and Bluffer are *not* liable to be attacked by Defender, and for exactly the same reason: before Defender attacks them, there is no forced choice. There are no costs that need distributing or that need assuming.

Liability by Duty nonetheless predicts that Jammy and Bluffer bear the upshots of liability. The key to seeing this is to notice that a person can incur a duty to assume the costs of *someone else's* attack. There are various ways to incur such a duty. For example, X could incur a duty to assume the costs of Y's attack by having promised to do so, or in virtue of the fact that Y is acting as X's authorized agent. But another way X can incur a duty to assume the costs of Y's attack is by X's bearing greater culpability or responsibility than Y for that attack. This is, I think, very plausible regardless of what one thinks about Liability by Duty.[42]

[42]Liability by Duty agrees with theorists like Ferzan that culpability for a wrongful threat can ground liability, and it agrees with theorists like McMahan that responsibility for a wrongful threat can ground liability. This is because culpability or responsibility for a wrongful threat are among the things that can ground a duty of cost assumption. But because a person can incur a duty to assume costs for which they are neither culpable nor responsible, Liability by Duty requires *neither* culpability nor responsibility for liability. Consider, for example: an adult lifeguard has, with clear eyes about the risks of lifeguarding, contracted to save drowning swimmers in his vicinity, even if doing so requires moderate injury. He sees a child drowning, and can rescue the child only by suffering minor cuts to his own arms and legs. Not wanting to suffer these minor injuries, he chooses not to save the child. You, however, are in a position to compel the lifeguard to keep his contractual duty by shoving him into the ocean, causing him to suffer the minor cuts to his arms and legs (you know he'll save the child once he's already in the water). The lifeguard is not responsible or culpable for the fact that harm must fall on either the child or the lifeguard, but Liability by Duty predicts (plausibly, to my mind) he is liable to be harmed by you nonetheless, in virtue of his duty to see that harm falls on himself rather than the child.

Though I lack the space to defend the idea, I think *complicity* can also ground duties of cost assumption, even when that complicity is not itself a product of responsibility or culpability for causing an unjust threat. Consider, for example, the case of a getaway driver for an assassination who does not make any contribution to the assassination, but

For reasons already canvassed in §5.1, Jammy and Bluffer bear more culpability and responsibility than Defender for Defender's attack on them.[43] (This is especially clear with respect to comparative *culpability*. Defender is not at all culpable.). It is independently plausible that this fact puts Jammy and Bluffer under a duty to assume the costs of Defender's attack, despite the fact that they have a right that she not attack them and despite the fact that Defender is not fact-relatively justified in attacking them.

Suppose that's right; suppose Jammy and Bluffer have a duty to assume the costs of Defender's attack. The four upshots of liability follow.

The fact that Jammy and Bluffer would not be permitted to fight back (**first upshot**) follows very straightforwardly from the fact that they have a duty to assume the costs of Defender's attack. Fighting back would just

---

who contributes only *after* the assassination by helping the assassin to escape. The driver is later taken into custody, while the assassin remains at large. I find it plausible that the driver would have a duty to pay at least partial compensation for injuries suffered during the assassin's attack, in virtue of the driver's (non-contributory) complicity in the attack. (C.f., Saba Bazargan-Forward, "Complicitous Liability in War," *Philosophical Studies* 165 (2013): 177-195.)

[43]When a merely apparent attacker provokes and is responsible for a defender's use of force, is this because he causes her to *believe* he poses an actual threat, or because he provides her with *evidence* sufficient to justify this belief? Typically, both ingredients matter. Evidence matters because (excepting unusual cases of direct belief manipulation) X will typically be responsible for Y's belief only when and because X is first responsible for giving Y evidence on the basis of which to form that belief. But belief also matters in most cases. Consider a version of Bluff in which, although Bluffer provides Defender with evidence that he poses a threat, Defender irrationally (but correctly) believes that he poses no threat. Defender hates Bluffer, however, and takes his bluff as an opportunity to harm him without legal repercussions. In this case, Bluffer would be responsible for providing Defender with evidence, but would not be responsible for Defender's use of force, in light of the lack of a responsibility-transmuting connection between the evidence and the use of force. Evidence on its own is typically not enough. Responsibility for evidence will rarely translate into responsibility for the defender's use of force unless *the defender's action is explained by a belief she forms on the basis of that evidence*. In all but the most unusual cases, belief is the "responsibility-transmuting connection" between the defender's evidence and the defender's use of force. I am grateful to an anonymous editor at *Ethics* for encouraging me to think about this matter, and about cases where belief and evidence come apart.

be a way of failing this duty, since to fight back would be to attempt to redirect the costs of the attack onto Defender.

The fact that a bystander would not be permitted to fight back against Defender (**second upshot**) is less straightforward, since a bystander, of course, is not herself under a duty to assume the costs of the attack. But — and here's the key — neither is *Defender* in this case required to assume from Jammy and Bluffer the costs of her attack. Even though Defender is the most direct cause of the fact that harm must fall on someone or other, Jammy and Bluffer nonetheless bear considerably greater responsibility and culpability for this fact than Defender. As such, it seems Jammy and Bluffer have a duty to assume unavoidable costs from Defender, but not vice versa. And from this it follows from Liability by Duty that Defender is *not* liable to be made to assume the costs of the attack. This explains why a bystander may not harm her in defense of Jammy and Bluffer.

The explanation for the fact that Jammy and Bluffer are not owed compensation if they are injured by Defender (**third upshot**) likewise turns on the fact that Defender lacks a duty to assume the costs of her attack from Jammy and Bluffer. A duty to pay compensation for harming Jammy and Bluffer would just be an *ex post* instantiation of the duty to assume the costs of her attack; paying compensation is just what it looks like to assume the costs of an attack after the damage has been done. Since Defender has no duty to assume the costs of her attack *ex ante*, she has no duty to pay compensation after the attack has concluded.

As for the fact that Jammy and Bluffer have little or no standing to complain about Defender's attack (**fourth upshot**), this isn't explained by any unique feature of Liability by Duty. To rehearse the point made in §5.1, there are clear cases where a person can compromise their standing to complain about some harm without losing their right against that harm. Among the ways in which this can happen is by foreseeably (and without justification) inciting someone to violate that right, as illustrated by

the case where I provoke Colleague to violence, and as illustrated in Jam and Bluff. For the same reasons Jammy and Bluffer have little standing to complain given Liability by Fairness, they have little standing to complain given Liability by Duty.

The Liability by Duty approach, then, also leads us to Partial Forfeiture. And as for those merely apparent attackers who do *not* bear the upshots of liability (e.g., Innocent and Twin), Liability by Duty gives the following explanation. Characters like Innocent and Twin have done nothing that plausibly puts them under a duty to assume the costs of Defender's mistaken use of force. This is why they do not forfeit the moral upshots of certain rights in the way that Jammy and Bluffer do.[44]

6 Two Concerns about Partial Forfeiture

Before turning to the comparative advantages of Partial Forfeiture over Orthodoxy, I want to head off two concerns I anticipate at this juncture.

[44] A point of nuance. As we've noted, Liability by Fairness and Liability by Duty both depart from Orthodoxy in claiming that Jammy and Bluffer's actions do not make them liable to be attacked by Defender. But they may be *conditionally liable* to certain harms. Imagine that Defender attacks Bluffer by rolling a boulder towards him, and in doing so makes harm inevitable: Bluffer will have his leg broken by the boulder unless he rolls it back towards Defender. According to Liability by Fairness and Liability by Duty, Defender wrongs Bluffer by setting the boulder in motion. But once the boulder is in motion and a forced choice has been created, Bluffer does become liable to something — specifically, he becomes liable to harm that is necessary to prevent him from redirecting the boulder towards others. Thus, for example, neither Defender nor a bystander would wrong Bluffer by preventing him from redirecting the boulder back towards Defender.

Liability by Fairness/Liability by Duty, then, does make room for a kind of liability, but one that is entirely conditional on Defender reacting in certain ways and that arises only *after* the Defender has reacted in those ways. And even where there is such liability, the view differs from Orthodoxy with respect to *what* the apparent attacker is liable to. According to Liability by Fairness/Liability by Duty, the apparent attacker is only liable to harm that is necessary to prevent the costs of the defender's reaction from falling on someone else, and proportionate to that end. Orthodoxy, as we've seen, attributes to merely apparent attackers a liability that is not so constrained by considerations of necessity and proportionality.

6.1 What of Permissibility?

One might think that it's not just that characters like Jammy and Bluffer are *not wronged* when they are harmed by Defender, but that it is *permissible* for Defender to harm them. Jorgensen expresses such a thought:

> [Defender] has of course acted suboptimally; it would be better if her mistake had not occurred ... We can say that in the evaluative sense of 'ought', her mistake ought not have happened, but it does not follow that it was objectively impermissible ... [such] reasonable mistakes are *permissible* as the byproduct of a justified social practice for fairly managing the risk of suffering unjust harm.[45]

This invites a worry for Partial Forfeiture, which is that while the theory may explain why characters like Jammy and Bluffer bear the upshots of liability, it fails to explain the fact that it is permissible for Defender to harm Jammy and Bluffer.

I doubt there is any real problem here, however, or at least not one that isn't equally shared by Orthodoxy. This is because, regardless of whether you endorse Orthodoxy or Partial Forfeiture, you should agree on two things: (1) It is *fact-relatively impermissible* for Defender to harm Jammy or Bluffer; and (2) it is *evidence-relatively permissible* for Defender to harm Jammy or Bluffer. All parties should agree on (1) because liability is a necessary, but not a sufficient, condition on permissible harm. To say that someone is liable is only to say that a certain reason to *not* harm them is *absent*; it does not imply that there is sufficient positive reason *for* harming them. Permissible harming requires a sufficient balance of reasons in favor of harming. This requirement is plainly not met in the case of Jam and Bluff, since there is no harm to be prevented by harming them.

[45] "Reasonably Mistaken Self-Defense," fn. 3, p. 143.

As for (2), even defenders of such heretical views as Partial Forfeiture (like myself) can agree that Defender has an evidence-relative justification (and excellent excuse) for harming Jammy and Bluffer, since Defender's evidence gives her overwhelming reason to believe that Jammy and Bluffer are liable and that there is much to be gained by harming them.

So the first thing to say about the purported intuition that it is permissible for Defender to harm Jammy and Bluffer is that there is no special problem here for Partial Forfeiture. If that intuition is an intuition about evidence-relative permissibility, then Partial Forfeiture, no less than Orthodoxy, easily accommodates the intuition. And if the intuition is one about fact-relative permissibility, then Partial Forfeiture, no less than Orthodoxy, is inconsistent with that intuition. But also like Orthodoxy, Partial Forfeiture provides the resources to tell a plausible debunking story, according to which this intuition that may seem to be about fact-relative impermissibility is really just tracking other factors: perhaps evidence-relative permissibility, perhaps excusability, or perhaps the presence of the various upshots of liability.[46] In short, this purported problem is no more a problem for my view than for the orthodox view — though I don't think it is a genuine problem for either view in any case.

6.2 What's Left Over?

A second natural worry is that Partial Forfeiture makes a distinction without any real difference.[47] That is, one might wonder what is really "left over" if a person maintains a right while being stripped of the defensive permissions, compensatory claims, and the standing to complain normally associated with that right. What is the significance of such a naked right?[48]

[46]My thanks to an anonymous reviewer at *Ethics* for encouraging me to think of Partial Forfeiture as offering a debunking story of this sort.

[47]I'm grateful to numerous colleagues and to reviewers at *Ethics* for raising this concern.

[48][Redacted.]

The answer is that a right without those upshots still performs the most fundamental function of a claim right, which is to give others reasons for action: my having a right that you ϕ *necessarily* gives you moral reason to ϕ, and thus counts against the (fact-relative) permissibility of your not doing ϕ. Merely apparent attackers who lose some of the contingent, non-constitutive upshots of their rights do not lose the reason-giving force of those rights. Those rights still make an important difference as to how others may treat them. This is hard to notice in Jam and Bluff, where the (fact-relative) impermissibility of harming Jammy and Bluffer is overdetermined in light of the absence of any benefits to harming them. But we can see the difference that even "naked" rights make by considering cases where there are positive reasons for harming a merely apparent attacker. Consider:

> *Vehicular Bluff*. Bluffer wants Defender to have a good scare, and so pretends that he is about to hit her with his car. Defender reasonably believes that she can defend herself only by shooting out Bluffer's tires to redirect his car. Redirecting his car will have the result that the car shields a bystander from a nearby, unrelated avalanche. If Driver is redirected into the avalanche, he will be killed and the bystander saved. The bystander will be killed otherwise.

There is much to be gained by killing Bluffer in this case. So if Bluffer lacks his right to not be killed, then it looks like the two ingredients for permissible defense are present: liability and sufficient reason for harming. If Bluffer does not lose his right, however, then it would not be (fact-relatively) permissible to kill him. This is because rights impose very weighty reasons for action such that a person cannot permissibly infringe one person's right to life in order to save someone else's life. What this case illustrates, then, is that a person's right against harm, even stripped

of its usual contingent upshots, continues to have reason-giving force and make a considerable deontic difference.

Indeed, as we'll see in a moment, this is precisely why Partial Forfeiture departs from Orthodoxy in recommending that Observer share information with Defender in a case like Mixed Bluff (from §4.3), and also why it avoids paradoxical results in the triad of Tracks cases (from §4.4).

7 Advantages of Heresy

The strongest selling point of Partial Forfeiture is the fact that it is not plagued by the problems of Orthodoxy — problems canvassed in §4. (Of course, for those independently attracted to Liability by Fairness or to Liability by Duty, another selling point of Partial Forfeiture is that it is a consequence of those theories.)

The **first** challenge for Orthodoxy, recall, is to explain why it is wrongful to impose unnecessary harm on a character like Driver in a case like Easy Defense. The standard explanation is that this is wrong because *liability is restricted to necessary defensive harm*. As we saw, Orthodoxy is incompatible with this claim, since it implies that some persons (like Jammy and Bluffer) can be liable to unnecessary harm.

Partial Forfeiture, by contrast, is compatible with the standard explanation. This for the simple reason that it does not imply that merely apparent attackers are liable to harm. It only implies that persons sometimes lose their permission to fight back against unnecessary harm, or their standing to complain about such harm, or their claim to compensation.

Likewise for the **second** challenge for Orthodoxy, which concerns the proportionality constraint. The standard explanation for why — as in Overkill — it is wrong to kill someone who merely poses a threat of bruising your toe, is that *liability is restricted to proportionate defensive harm*. Again, we saw that Orthodoxy was incompatible with this claim, since it

implies that people (like Jammy and Bluffer) can be liable to disproportionate harm.

No trouble here for Partial Forfeiture, however. It does not imply that a person can be liable to disproportionate harm — again, it only implies that people sometimes lose their permission to fight back against unnecessary harm, or their standing to complain about such harm, or their claim to compensation.

The **third** worry for Orthodoxy is that it makes one person's lack of information a ground of another person's loss of rights. This seems implausible in its own right, but also has implausible implications for how agents should share information in defensive situations. In the case we considered (Mixed Bluff), Defender reasonably believes that Bluffer poses an actual threat, but Observer knows that Bluffer is only bluffing. Observer is considering whether to communicate this fact to Defender. On Observer's evidence, there is a 90% chance Defender harms Bluffer if Observer does *not* pass the information along; and there is a 50% chance that Defender still harms Bluffer if Observer does pass the information along. Given Orthodoxy, this means that Observer is faced with the following choice. Option 1: she can withhold from Defender the crucial information that Bluffer is merely bluffing, in which case she can be certain that Defender will *not* wrong Bluffer. Option 2: she can pass along the information, in which case there's a 50% chance Bluffer will be wronged.

The problem here for Orthodoxy is that it seems to imply — very counterintuitively — that Observer should choose to withhold the crucial information from Defender, and that, in fact, this would be best, not just for Defender, but also for the purposes of protecting *Bluffer's* rights.

Partial Forfeiture delivers more plausible recommendations for Observer. According to Partial Forfeiture, Bluffer has a right that Defender not attack him, regardless of Defender's epistemic position. Thus, if Ob-

server withholds the information from Defender, there is a 90% chance that Defender violates Bluffer's rights; and if Observer passes along the information, there is only a 50% chance that Defender violates Bluffer's rights. Passing the information along would be better for all concerned. That seems the right result!

Finally, the **fourth** worry for Orthodoxy is that it leads to paradoxical forms of moral decision-making under uncertainty. Recall:

> *Busy Uncertain Tracks*. On track A you know there are 1,000 people who each pose an actual threat. On track B, you know there are 700 actual threats and 300 mere bluffers (so for each person on track B you are 70% sure he poses an actual threat).

The trouble for Orthodoxy, recall, is that, it seems to permit you to kill 300 people you know to be mere bluffers over killing 300 people you know to be actual attackers. But Partial Forfeiture does not have this implication, since it does not permit you to infer that someone is liable to be killed from the fact that it is merely probable that they are an actual attacker. All that Partial Forfeiture permits you to infer from the fact that it is probable that someone is an actual attacker is that *it is probable that they are liable*. For this reason, Partial Forfeiture delivers consistency across all three of the "Tracks" cases. In Transparent Tracks, you know that sending the trolley down track A violates no one's rights, and that sending the trolley down track B violates someone's right. Moral decision-making requires you to minimize your expectation of violating rights, all else equal. Thus, you are not permitted to send the trolley down track B. In Uncertain Tracks, you again know that sending the trolley down track A violates no one's rights. But given Partial Forfeiture your evidence only makes rational a credence of 0.7 that sending the trolley down track B will violate no one's rights. Minimizing the expectation of violating rights once again forbids sending the trolley down track B. The same

goes for Busy Uncertain Tracks. You know that sending the trolley down track A violates no rights, and you know that sending the trolley down track B violates 300 rights. Minimizing the expectation of violating rights forbids sending the trolley down track B.

8 Concluding Remarks

Rights and their moral upshots are standardly assumed to be a package deal. The Challenge of Merely Apparent Attackers gives us reason to rethink this idea, since there are, we've seen, significant advantages to answering this challenge by appeal to the idea that a person's right can come "unbundled" from some of the typical moral upshots of that right. Let me conclude with a comment about this revisionary way of understanding rights and rights forfeiture.

Though I've focused in this paper on a particular challenge within the ethics of harm, I want to suggest that taking the "unbundling" idea seriously may have fruit to bear in other domains of moral philosophy. I am inclined to think that the assumption that rights and their upshots are a package deal has led us to overlook theoretical possibilities that present themselves once we entertain the unbundling possibility.

One illustrative example comes from promissory morality. Familiarly, not all (attempted) promises are *valid* promises. That is, some attempts to promise may fail to put the promisor under a moral duty to perform the action promised. Sometimes this is for *procedural* reasons — for example, a lack of competence or knowledge on the part of the promisor. Other times this is for reasons of *content*. For example, a promise may fail to be valid on account of the wrongfulness of the action promised: I cannot validly make a promise to you to murder your philosophical rival. Likewise, a promise may fail to be valid on account of ill-formed content: I cannot validly make a promise to both show up and not show up to your party.[49]

[49] For discussion of the relationship between promissory content and validity, see: J.E.J

There is, however, an interesting subclass of promises that are often thought to be invalid on grounds of content, despite the fact that the actions promised are well-formed and morally permissible. Suppose, for instance, that Dara promises to donate her kidney to a coworker if they are ever in need of a kidney. This promise may seem not to be valid; it may seem not to confer on the promisee a right to performance. It may seem not to confer a right to performance, I think, because the promisee is not permitted to demand or enforce the performance of this kind of promise (even with forms of enforcement appropriate to promises).

Such a promise, however, seems to have *some* moral import, in a way that invalid promises to do what is grossly immoral do not. Dara's coworker cannot demand or enforce performance, but some kind of apology seems appropriate if Dara decides not to keep her promise, and she seems to have some kind of duty to "make up for" the broken promise. Invalid promises to do what is immoral are not like this. Apology or compensation is in no way called for if I decide not to commit murder on your behalf. Moreover, consider the fact that it seems Dara would wrong her promisee by breaking her promise for the wrong kinds of reasons.[50] Suppose, for example, that Dara decides not to give up her kidney only because she learns that a different coworker will give Dara his parking spot at work if she gives the kidney to him instead. It sure seems that she would thereby wrong her original promisee. (And, again, my promise to murder for you is not like this.)

Altham, "Wicked Promises," in Ian Hacking (ed), *Exercises in Analysis* (Cambridge University Press 1985): 1-22; David Owens, "Promises and Conflicting Obligations," *Journal of Ethics and Social Philosophy* 11 (2016): 1-19; Seana Shiffrin, "Immoral, Conflicting, and Redundant Promises," in R. Jay Wallace, Rahul Kumar, and Samuel Freeman (eds), *Reasons and Recognition: Essays on the Philosophy of T.M. Scanlon* (Oxford: Oxford University Press 2011): 155-178; Judith Jarvis Thomson, *The Realm of Rights* (Cambridge: Harvard University Press 1990): 313-316; Gary Watson, "Promises, Reasons, and Normative Powers," in David Sobel, Steve Wall (eds), *Reasons for Action* (Cambridge University Press 2009): 155-178.

[50] [Redacted.]

This all seems rather puzzling if we make the usual assumption that rights and their upshots are a package deal. But if take the possibility of "unbundling" seriously, this opens up space to contemplate the possibility of a class of promises that occupy a middle ground between standard valid and invalid promises. It opens up the following diagnosis of Dara's promise. When she promises to donate her kidney, this promise is valid in the sense that it confers on the promisee a right to performance. But promises that implicate one's bodily autonomy in important ways are restricted in that they confer on the promisee rights that are stripped of some of their usual upshots. Specifically, they are stripped of those upshots that would, if present, compromise the promisor's bodily autonomy in unacceptable ways. A partial unbundling of this sort would explain why the promise has some moral import (e.g., why apology and compensation is appropriate, and why breach may only be done for certain reasons) while also explaining why the promisee may not demand or enforce performance.

To be clear, my aim at present isn't to make the case for such a class of promises, nor to argue that this is how we should understand a promise like Dara's. That would take us much too far afield. My aim is just to give an example of a theoretical possibility that has some *prima facie* plausibility and that will go overlooked if we fail to consider the possibility that the typical upshots of rights may not always be a package deal. This paper has made a case for unbundling in the service of answering a particular challenge in the theory of the ethics of harm. But I am optimistic that the idea has fruit to bear elsewhere in moral philosophy.[51]

[51] [Redacted.]