# Self-Defense and the Problem of the Innocent Attacker*

*Jeff McMahan*

That there are occasions on which it is permissible intentionally to kill another person in self-defense is an axiom in contemporary ethical theory. While certain Christian moralists once held that killing in individual self-defense is prohibited though killing in war may be permitted, that view is now an antique curiosity and today even many pacifists concede that killing in individual self-defense is sometimes permissible. But, while we are confident that killing in self-defense can be permissible, there is uncertainty about *why* this is an exception to the general proscription of intentional killing. In what follows I will explore the foundations of our views about self-defense by mapping some of the intuitive terrain and examining certain theories of self-defense to determine whether our intuitions can be supported. I will focus on cases in which killing is necessary for self-defense against a lethal threat. These are the hardest cases, since there is no possibility of dividing the costs of conflict among the parties.

## I. THE PRESUMPTION AGAINST SHIFTING HARMS

There is a strong moral presumption against causing harm. This is true even when it is inevitable that someone must suffer harm and shifting the harm from one person to another would reduce the amount of harm that must be suffered. The presumption in these cases is a corollary of the moral asymmetry between doing harm and allowing harm to occur; for to shift a harm is normally to do harm while not to shift it is to allow the initial victim to suffer it.

*This article, which was initially drafted in the summer of 1991, develops certain themes from an earlier and longer manuscript that will form the basis of certain foundational chapters in *The Ethics of War*. I am very grateful to Sissela Bok, Michael Gorr, Shelly Kagan, Gregory Kavka, Steven Lee, and Walter Sinnott-Armstrong for comments on the earlier manuscript and to Ann Davis, Derek Parfit, David Wasserman, and Noam Zohar for comments on earlier drafts of this article. I also gratefully acknowledge support for my work from the National Endowment for the Humanities, the John D. and Catherine T. MacArthur Foundation, the U.S. Institute of Peace, and the Program for the Study of Cultural Values and Ethics at the University of Illinois.

*Ethics* 104 (January 1994): 252–290

We may distinguish two types of case involving the shifting of harm. In the first, harm has already occurred. While this harm cannot be literally redistributed, it nevertheless counts as shifting the harm if a new harm is imposed on another person as a means of canceling the intitial harm by fully compensating the victim. The presumption against shifting a harm in this way is sometimes articulated by the claim that harms should lie where they fall. In the second type of case, it is inevitable that a harm will occur. There are two ways of shifting an inevitable harm. First, it is sometimes possible to shift a future harm by redirecting the threat itself (as, e.g., in the well-known cases involving a runaway trolley that will kill people on one track unless it is diverted to another track where it will kill different people). Call this mode of shifting an inevitable harm *redirection*. Second, it is sometimes possible to avert harm to one person by creating a new and distinct threat that harms another. Call this mode of shifting a harm *creation*. In both redirection and creation, if the person on whom harm is inflicted is himself the agent of the harm that is averted, the act of shifting the harm counts either as *self-defense* (if the person who shifts the harm is the potential victim) or as *other-defense* (if the harm is shifted by a third party). If the person on whom harm is inflicted is not the agent of the harm that is averted, the act is an instance of *self-* or *other-preservation*.[1] Self- or other-defense normally involves the creation of a new threat. Hence most instances of redirection involve self- or other-preservation. But redirection can be self- or other-defensive—as, for example, when a gunman's bullet is deflected back at him.

The presumption against shifting a harm can be overcome. Consider the type of case with which we will be concerned: that involving an inevitable harm (i.e., someone must suffer a harm in the future). It might be considered a way of shifting an inevitable harm, if one is in the path of a threat, simply to duck or dodge the threat so that it then harms someone else.[2] This is self-preservation by a form of redirection. It is generally permissible—primarily, I believe, because it normally involves allowing unintended harm to occur rather than doing or intending harm.[3] If the presumption against shifting harm

1. The latter term suggests that it is a lethal threat that is averted. We need a more general term that covers the shifting of nonlethal harms as well; but, since my concern is with the shfting of lethal harms, I will not offer one here.

2. For a discussion of these cases, see Christopher Boorse and Roy A. Sorensen, "Ducking Harm," *Journal of Philosophy* 85 (1988): 115–34.

3. This understanding of "ducking harm" is supported by my analysis in "Killing, Letting Die, and Withdrawing Aid," *Ethics* 103 (1993): 250–79. The claim is that not to duck is to save the other possible victim (even if one does so inadvertently) by shielding him from the threat. Thus ducking counts, not as an act of harming, but as an active refusal to save the victim from harm. As my later remarks will suggest, the

is a corollary of the asymmetry between doing and allowing, and if ducking is an instance of allowing, then the presumption does not apply in cases of redirection by ducking—unless, for example, the agent has culpably caused the threat or intends for the threat to harm the victim. But even when, as in the majority of cases, redirection involves doing harm rather than allowing harm to occur, the presumption against shifting harm by redirection is generally weaker than that against shifting harm by creation, other things being equal (e.g., the threat is not redirected with the intention that it should strike the victim). Thus the weaker presumption against redirection can sometimes be defeated if redirection would reduce the overall harm or, perhaps, if it would distribute the harm more widely, thereby reducing the per capita harm without increasing the overall harm. Personal partiality may also permit a potential victim to redirect a threat away from herself provided that the harm it causes to another is proportionate—that is, does not exceed by a certain amount the harm that the agent thereby avoids.[4]

These examples suggest that there are cases in which there is either no presumption against shifting an inevitable harm (as is arguably the case in some instances of ducking a harm) or only a relatively weak presumption. But in many cases the presumption is strong. It is strong when the agent who shifts the inevitable harm does so by making a previously unthreatened person the intentional object of his action in a way that the agent believes to be sufficient to harm that person. Shifting an inevitable harm in this way offends against both the presumption against doing harm and the presumption against intentional harming.[5] Thus, for example, it is a paradigm of wrongful

---

morality of ducking is altered if the agent is culpably responsible for the creation of the threat.

4. Determining the role that personal partiality has in fixing the permissibility of action is one of the central problems of ethics. It seems that partiality may legitimately influence the distribution of certain benefits as well as choices as to which unintended harms to prevent and which to allow, given that not all can be prevented. Beyond this, the role of partiality is controversial. I will return to this.

5. As I interpret it, the presumption against intentional harming or killing that underlies the Doctrine of Double Effect (DDE) does not require that harming or killing should itself, *under that description,* be an intended effect in order for the act that causes it to violate the presumption. It is sufficient for the act intentionally to affect the victim in a way that the agent believes to be causally or logically sufficient to harm or kill the victim. I develop this claim in "Revising the Doctrine of Double Effect," *Journal of Applied Philosophy* (in press). The argument of that article is indebted to that in Warren Quinn, "Action, Intention, and Consequences: The Doctrine of Double Effect," *Philosophy and Public Affairs* 18 (1989): 334–51. For convenience, I will continue to use the usual locutions "intentional harming" and "intentional killing" to refer to acts that violate the presumption laid down by the DDE even if the harming or killing is not, in the strictest sense, an intended effect.

action to kill an Innocent Bystander, or otherwise use him in a way that one believes will result in his death, as a means of self-preservation. (An *Innocent Bystander*—henceforth, IB—is a person who is not causally involved in the production of a threat. A person's status as an IB may be relative to a particular threat. A person who poses a certain threat may be an IB relative to a situation involving a different threat.) Normally, neither personal partiality nor a diminution of overall harm is sufficient to defeat the presumption against shifting a harm by intentionally killing an IB. Indeed, this mode of shifting harm is so widely condemned that we reserve the unequivocally pejorative term "terrorism" for its manifestation in political contexts.

Typical instances of killing in self-defense involve this same mode of shifting harm. In order to avert a harm to herself, the agent who engages in self-defense intentionally affects a person in a way that she believes will, if successful, kill that person. Her action offends against both the presumption against doing harm and the presumption against intentional harming. It is, therefore, a case in which the presumption against shifting harm is strong. How, then, can we be so confident about the permissibility of killing in self-defense?

To try to answer this question, let us consider the paradigm case of justified killing in self-defense—that is, the case in which self-defensive killing seems most clearly justified. Identifying the features of this case should help us to determine which features defeat the presumption against shifting harm. In this and other cases of self-defense, let us refer to the person who initiates the conflict as the *Attacker* and the person who is the target of the initial attack as the *Victim*.[6] (I use "Attacker" as a technical term that refers to anyone who is engaged in action that is causing or threatens to cause harm. According to this use, the verb "attack" is like "kill" in that one can attack inadvertently or accidentally.) In the paradigm case, then, the Attacker (1) poses a lethal threat to the Victim, (2) poses this threat through his present action, not through his mere location or physical movement nor through his past action or anticipated future action, and (3) intends the threat he poses. Because (4) his action is morally unjustified and because (5) it is also unexcused (e.g., there is no diminished responsibility, nonculpable ignorance, or duress), it follows that (6) the Attacker is morally culpable for posing the threat. The Victim (7) has done nothing to provoke the attack (e.g., by attacking first), (8) has no option other than killing the Attacker that would be comparably effective in averting the threat, (9) causes expected harm to the Attacker that is not significantly greater than the expected harm to herself that she

6. For the purposes of this article, Attackers will be male and Victims female. This reads quite naturally given statistics on the actual occasions for individual self-defense.

thereby averts, and (10) does not, in taking self-defensive action, cause harm to IBs.

Many of these features seem morally significant and may contribute to our sense that the killing is permissible. In cases in which one or more of these features is absent, or appears in a weaker form, self-defensive killing may seem less obviously justified. If enough of the features are missing, self-defensive killing will not be justified. The problem is to determine which feature, or combination of features, is decisive in defeating the presumption. One possibility, of course, is that more than one feature or combination of features is decisive. Rather than there being a single, unitary justification for killing in self-defense, there may be several distinct justifications, so that self-defense may be justified in one way in one case and another way in another. This possibility may be obscured by focusing on the paradigm case, in which the different justifications would converge. In the paradigm case, when all of the features that make self-defensive action justifiable are present together, the justification may be *overdetermined*—that is, self-defense may be justified for more than one reason. Since I will argue that the correct justification for self-defense in certain cases is incapable of supporting it in other cases in which we believe it to be permissible, I think it likely that the right of self-defense does indeed have multiple, independent foundations.

## II. THE JUSTICE-BASED ACCOUNT

According to one view, which I call the ORTHODOX VIEW, self-defense is, in effect, self-justifying and absolute. Violence that is necessary to defend oneself from a threat and is directed only at the agent of the threat is by its nature justified. This view has a long history in the traditional theory of the just war, which holds that, while it is wrong intentionally to attack the innocent, those who are noninnocent forfeit their immunity to attack. To be innocent is to be not *nocentes*—that is, not one who is harmful or who injures. To be noninnocent in the relevant sense, therefore, is simply to be engaged in causing harm.[7] Some writers refer to this as "material noninnocence" to distinguish it from the more familiar notion of moral noninnocence. Since the materially noninnocent forfeit their immunity to attack, it is permissible to use violence in self-defense against them. In short, the mere

7. A typical statement of this view holds that "the wrongfulness of killing the innocent is primarily the wrongfulness of killing the causally innocent, those who are doing no harm, not the wrongfulness of killing the juridically innocent, those who have committed no crime, or the morally innocent, those who are not in a relevant respect morally flawed." See Philip Devine, *The Ethics of Homicide* (London: Cornell University Press, 1978), p. 152. I criticize this view in some detail in "Innocence, Self-Defense, and Killing in War," *Journal of Political Philosophy* (in press), sec. 2.

fact that someone is engaged in causing harm defeats the presumption against shifting harm; for it is permissible to harm him to prevent the harm he would otherwise cause. (This view is sometimes articulated by reference to the contrast between self-defense and self-preservation. While the presumption against shifting harm is defeated in the case of self-defense, it remains strong in the case of self-preservative violence since the latter is directed against a Bystander who is, on this view, innocent by definition.[8] While the ORTHODOX VIEW's appeal to such notions as innocence and the forfeiture of immunity make it appear that it offers an explanation of the permissibility of self-defense, these notions are simply components of its analysis of the concept of self-defense. The ORTHODOX VIEW does not explain but merely asserts the permissibility of self-defense.)

The ORTHODOX VIEW treats the question whether the Attacker is morally culpable, morally innocent, or indeed morally justified in what he does as irrelevant. This, I believe, accounts for its appeal to theorists of the just war, since this enables it to support the view embodied in the laws of war that soldiers on each side in a war are permitted to kill soldiers on the other side, irrespective of which side's cause is just. Treating these considerations as irrelevant is, however, wholly implausible outside the context of war.[9] Suppose, for example, that a police officer begins shooting at a murderer to prevent his committing a further murder. The murderer is engaged in causing harm in a way that is both unjustified and unexcused. He is a *Culpable Attacker* (henceforth, CA). The police officer's action is, by contrast, clearly justified, assuming that there is no less harmful means of preventing the murder. We should, however, distinguish two types of Attacker whose action is morally justified. A *Justified Attacker* is justified in acting even though there is a victim who is wrongfully harmed, or wronged, by his action. A *Just Attacker* is also justified in his action, but those who are harmed by it are liable to the harm—that is, they are not wronged, nor are their rights violated by it. While most of us believe that those who would be wronged by the action of a Justified Attacker are normally permitted to engage in self-defense against him, we also believe that there is no right of self-defense against a Just Attacker. For example, since the police officer is a Just Attacker, the murderer must not fight back, even in self-defense. Yet, according to the ORTHO-

8. See, e.g., Jenny Teichman, *Pacifism and the Just War* (Oxford: Basil Blackwell, 1984), pp. 84–85. Teichman's definition of self-defense is unusually narrow, since it insists that the Attacker must intend to harm his Victim.

9. I argue in "Innocence, Self-Defense, and Killing in War" that these considerations are in fact relevant to the morality of war and that proponents of the ORTHODOX VIEW are led to believe otherwise by the mistaken assumption that principles that are plausible as laws of war also provide a plausible account of the morality of war.

DOX VIEW, the murderer is justified in killing the police officer, since the latter is materially noninnocent and has forfeited his immunity to attack. This is clearly wrong.

The case of the officer shows that a person's being materially noninnocent is not sufficient to justify self-defense against him. Mere material noninnocence may not compromise a person's immunity at all. Material noninnocence is not, moreover, necessary for the forfeiture of immunity. Suppose, for example, that a mine shaft has collapsed, leaving two miners trapped in a small open space. A radio communication has informed them that rescuers will reach them in five hours, but their instruments indicate that, while there is enough oxygen to allow one of them to survive for more than five hours, their oxygen supply will be depeleted within three hours if both continue to breathe. Suppose further that one of the miners (call him the "innocent miner") then learns two facts: first, that the other miner (the "culpable miner") deliberately engineered the collapse of the shaft in an effort to kill him and, second, that the culpable miner has a small oxygen tank that will allow him to survive for two hours after the oxygen in the shaft runs out. The innocent miner therefore has two options: he can do nothing and die of asphyxiation while the culpable miner survives or he can take the culpable miner's tank by force, thereby killing the culpable miner in self-preservation. (He might kill the culpable miner *in order* to take the tank from him or he might kill him *by* taking it from him. Either would count as killing in self-preservation.)

Let us say that one whose past action is causally responsible for a present threat is a *Cause*. The culpable miner is a *Culpable Cause*, since the past action by which he has caused the present threat was culpable. But he is not presently engaged in causing harm, nor is he part of the threat to the innocent miner. He is now, in fact, a Bystander—indeed an Innocent Bystander if "innocent" means "materially innocent." Hence killing him to secure the oxygen tank would not be an act of self-defense but an act of self-preservation. According to the ORTHODOX VIEW, he retains his immunity (just as civilian noncombatants who have culpably caused their country to launch an unjust war retain theirs). This, however, is a case in which killing in self-preservation seems clearly permissible.[10] The presumption against shifting the harm is defeated, not because the person onto whom it

10. The view that there is a fundamental moral difference between the self-defensive killing of a CA and the self-preservative killing of a Culpable Cause is defended by David Wasserman in "Justifying Self-Defense," *Philosophy and Public Affairs* 16 (1987): 356–78, esp. pp. 365–78. I criticize Wasserman's argument in *The Ethics of War* (New York: Oxford University Press, in press).

is shifted is materially noninnocent (for he is not), but because he is *morally* noninnocent.

If moral noninnocence alone can defeat the presumption in a case involving self-preservation, then it can clearly do so in cases of self-defense in which the Attacker is both morally and materially noninnocent. In cases, such as the paradigm case, involving necessary defense against a CA, the CA's culpability provides a sufficient justification for shifting the harm (even if other factors also contribute, so that justification is overdetermined). The reason is that, in cases in which a person's culpable action (whether past or present does not matter) has made it inevitable that someone must suffer harm, it is normally permissible, as a matter of justice, to ensure that it is the culpable person who is harmed rather than allowing the costs of his wrongful action to be imposed on the morally innocent.[11] In particular, if one person's culpable action threatens the life of another, it is permissible as a matter of justice to kill that person rather than to allow his culpable action to kill a morally innocent person, for considerations of justice give the innocent priority over the morally noninnocent. Call this the Justice-based Account of the right of self-defense (henceforth, JUSTICE).[12]

According to JUSTICE, it is a person's culpability that defeats the presumption against harming him. The harm imposed must, however, be appropriately related to the culpable action. Normally, for example, one may not harm a culpable person, even in self-preservation, if his culpability is unrelated to the threat to oneself. For, relative to that threat, the culpable person is a Bystander who is innocent both materially and morally. In short, culpability engenders liability only to those harms that are necessary to shift harms caused or made inevitable by the culpable action itself.[13]

It is important to stress that, as an account of the permissibility of shifting harms, JUSTICE is concerned with liability rather than with desert. As I will use the term, a person's being liable to a harm implies only that it is just that he should suffer it *if* it is necessary that someone must suffer harm. If it turns out not to be inevitable or necessary that someone must be harmed, then there is no reason why the liable

11. This requires qualification in various ways (e.g., it may not be permissible to inflict a great harm on a person in order to prevent his culpable action from causing a trivial harm; or, in some cases, it may be permissible to divide inevitable costs between the innocent and the culpable). These qualifications will be explored in *The Ethics of War*.

12. One understanding of this view is developed by Phillip Montague in "Self-Defense and Choosing among Lives," *Philosophical Studies* 40 (1981): 207–19, and in "Punishment and Societal Defense," *Criminal Justice Ethics* 2 (1983): 30–36.

13. It is possible to attribute a wider significance to culpability by developing the intuition that it is unjust if the wicked should flourish to the same extent as the virtuous. I will not pursue this.

person should be harmed.[14] If, by contrast, a person *deserves* a certain harm, then it is bad, other things being equal, if he fails to suffer it. Retributivists believe that culpable action can engender desert as well as liability and that punishment inflicts harms that are deserved. On this view, punishment is not intended to shift harms but to add new harms to those that have already occurred. JUSTICE, by contrast, implies only that culpable action gives rise to liability to harms inflicted in self- or other-defense or self- or other-preservation. (It is, however, possible to interpret JUSTICE not just as an account of the rights of self-defense and self-preservation but as a theory of punishment as well. So interpreted, it would reject the retributive function of punishment and would treat punishment as a means of shifting possible harms. Punishment would involve inflicting harm on those guilty of culpable action as a means of preventing or deterring possible harms that might otherwise be caused to the innocent. I will not discuss this broader version of the theory.)

As my references to other-defense and other-preservation suggest, the same considerations that justify the self-defensive killing of a CA and the self-preservative killing of a Culpable Cause also justify third-party intervention on behalf of the innocent in these cases, other things being equal (e.g., assuming that the innocent would not regard intervention as objectionably paternalistic). For the claims of justice that make it permissible for the Victim to kill the CA or Culpable Cause are impartial and ground a reason for killing that is agent-neutral in character. That the Victim has a stronger reason than most third parties is explained by the fact that the Victim's agent-neutral reason is augmented by agent-relative reasons that are too obvious to require comment.

Although it implies an agent-neutral reason for shifting harms from the innocent to the culpable, JUSTICE is best interpreted as a deontological theory. It holds, in other words, that self-defensive *action* against a CA is itself just, not that it is permissible because it produces a better or more just *outcome*. The permissibility of self-defense is to some extent independent of considerations of consequences.

It is, however, possible to interpret JUSTICE in such a way that it treats justice as a feature of outcomes rather than acts.[15] On this interpretation, self-defense against a CA is permissible because the outcome in

14. Compare Phillip Montague, "The Morality of Self-Defense: A Reply to Wasserman," *Philosophy and Public Affairs* 18 (1989): 81–89, p. 88.

15. The view that what morality requires is the maximization of justice in outcomes is developed by Fred Feldman in *Confrontations with the Reaper* (New York: Oxford University Press, 1992), pp. 182–90. A less monistic view is suggested by Laurence Alexander's claim that "a consequentialist approach can, and should, consider such factors as moral innocence." See his "Justification and Innocent Aggressors," *Wayne Law Review* 33 (1987): 1177–89, p. 1188.

which the CA is killed is more just than that in which the Innocent Victim is killed. If, however, the theory is to capture our intuitions when interpreted in this way, it must, at least in cases in which the CA poses a lethal threat, assert that considerations of justice outweigh all other features of the possible outcomes. For, according to commonsense morality, an Innocent Victim is permitted to kill a CA irrespective of differences in age, quality of life, or usefulness to society; she may do so even if the probability that the threat the CA poses would otherwise prove lethal is relatively low (e.g., because he is inept); and she may kill any number of CAs if this is necessary for self-defense.

In addition to giving considerations of justice absolute priority in the evaluation of outcomes in certain cases, a consequentialist interpretation of JUSTICE would also have to allow that one is not always required to do the act that would have the best expected consequences.[16] Otherwise the theory would make self-defense against a CA mandatory rather than optional. (The deontological version is compatible with the idea that self-defense is optional; for deontological ethics holds that self-sacrifice, even for the sake of those who are culpable, can be meritorious and cannot be opposed by impersonal considerations of justice, provided, of course, that no one other than the agent and the beneficiary is involved. According to this version, when a person voluntarily forgoes the option of self-defense, the reason she has to sacrifice herself does not eliminate the injustice of the CA's action, but it does, if she chooses freely, cancel the reason she would otherwise have to prevent the injustice.)

There are two possible consequentialist versions of JUSTICE, each based on a different conception of how considerations of justice affect the value of outcomes. According to one version, considerations of justice function to discount the moral significance of the harm that a CA might suffer from the Vicitim's self-defensive action. Self-defense is then justified on the ground that, because the CA's interests count for less, killing him leads to a better outcome than if he kills the Innocent Victim. According to George Fletcher, much of the legislation concerning the right of self-defense in the Anglo-American common-law tradition is founded on a view of this sort, which holds that, "as the party morally at fault for threatening the defender's interests, the aggressor is entitled to lesser consideration in the balancing process. His interests are discounted, as it were, by the degree of his culpability"[17] But, however

16. Some believe that the notion of a nonoptimizing consequentialism is incoherent. I will assume, however, that any theory that determines the rightness or wrongness of acts solely in terms of their consequences counts as consequentialist.

17. George Fletcher, *Rethinking Criminal Law* (Boston: Little, Brown, 1978), p. 858. The idea that a person's guilt diminishes the extent to which his interests count morally is defended, though not in the context of self-defense, by Shelly Kagan in "The Additive Fallacy," *Ethics* 99 (October 1988): 5–31, p. 20.

influential this variant of JUSTICE may be, it suffers from a fatal flaw. For, unless self-defense against a potentially lethal attack by a CA is subject to a proportionality restriction—which, as I noted earlier, commonsense morality rejects—the interests of the CA must be discounted all the way to zero.[18] Otherwise there must in principle be some number of CAs whose combined interests would outweigh those of the Victim, making the self-defensive killing of some or all of them impermissible. But the claim that the interests of the CA do not count at all is incompatible with our conviction that even self-defense against a CA is subject to a requirement of minimal force. If, for example, a Victim facing a potentially lethal assault by a CA can thwart the assault equally effectively either by incapacitating him or by killing him, she is morally required not to kill him. For even in this case it is wrong to cause more harm than is necessary for effective self-defense. And, intuitively, this is because the Attacker's interests do still count for something.[19]

There is, however, another consequentialist version of JUSTICE. According to this version, guilt and innocence function as independent factors that affect the evaluation of outcomes directly rather than by modifying the weights that attach to different people's interests. In a case involving self-defense against a CA, the CA's interests retain their normal moral significance but are simply opposed by considerations of justice, which provide a positive reason to harm him. It is not that the harm that the CA suffers is in itself a good feature of the outcome, as it might be if he deserved to be harmed; it is only that his suffering harm makes the outcome more just than it would be if the Victim were to be harmed instead.

This version may seem more plausible. But there is a general objection to the idea that JUSTICE can take a consequentialist form that is decisive, at least if it is a criterion of a theory's acceptability that it not be wildly at variance with our most fundamental and widely shared beliefs about self-defense. This is that the criteria for determining guilt and innocence that underlie our intuitions about liability are deontological in character. Consider, for example, a fundamental intuition that underlies JUSTICE: that, unless his action is excused by ignorance, duress, or diminished responsibility, one who attempts intentionally to kill an IB thereby becomes relevantly noninnocent, rendering himself liable to defensive violence as a matter of justice. This

18. Compare Wasserman, p. 359.

19. The deontological version has a plausible explanation of the requirement of minimal force. Since the interests of the CA retain their full normal weight and since justice is satisfied simply by ensuring that the Victim is not harmed by the CA's culpable action, any harm that is caused to the CA beyond what is necessary to prevent the harm to the Victim does not serve the purpose of justice and is objectionable for the same reason that causing harm is normally objectionable.

presupposes that there is a presumption or constraint against intentionally killing an IB that can be overridden, if at all, only in extreme circumstances. But theories that evaluate the morality of action solely in terms of consequences cannot hold, as a general matter, that there is such a constraint. For whether or not intentionally killing an IB is wrong depends, on such a view, on what the expected consequences would be; and whenever the consequences would be as good as or better than those of any alternative (e.g., when intentionally killing an IB would prevent the intentional killing of several other IBs by another agent), it is not wrong, but may be obligatory, intentionally to kill the IB. When that is true, the agent who intentionally kills the IB cannot, given the logic of consequentialism, thereby become guilty. For the theory judges his action to be morally permitted or even required and one can be guilty or culpable only for wrongful action.

But, while the agent who attacks the IB would retain his innocence, the IB herself, if she were to try to defend herself against the attack, would forfeit hers. For, if the attack were required because of its beneficial consequences, she would have no right of self-defense but would be required to submit to it (unless, perhaps, provoking resistance was for some reason among the desirable consequences of the attack). Or suppose, alternatively, that the agent refused to attack the IB. He would then be culpable with respect to whatever harm the killing would be needed to avert, thereby making himself liable to self-preservative violence by the potential Victims of that harm. And so on. I will assume that a theory that has these implications is unacceptable as an account of the rights of self-defense and self-preservation. JUSTICE must therefore be understood as a deontological theory.

## III. THE PROBLEM OF THE INNOCENT ATTACKER

So interpreted, JUSTICE provides a compelling—indeed, in my view, the best—explanation of the permissibility of self-defense against a CA. Yet it seems to suffer from narrowness of scope. For most people believe that there are cases in which self-defensive killing is justified but in which the Attacker is not morally culpable for the threat he poses to the Victim. In such a case, JUSTICE provides no ground for assigning priority to the Victim. If self-defense is justified, the justification cannot come from JUSTICE, which then fails to provide a complete account of the foundations of the right of self-defense.

There are two types of case in which self-defense seems justified but in which the Attacker is not culpable. One involves self-defense against a Justified Attacker. I will return to this type of case in Section VI. The other, on which I will focus because of the challenge it poses to JUSTICE, involves self-defense against an *Innocent Attacker* (henceforth, IA)— that is, an Attacker whose threatening action is morally unjustified but nevertheless excused or nonculpable. There

are various types of IA corresponding to three basic excusing conditions: nonnegligent ignorance, diminished responsibility, and duress. Among those IAs excused on grounds of ignorance are the *Inadvertent Attacker* who, without being either reckless or negligent, creates a threat to another that is unforeseen and therefore accidental, and the *Mistaken Attacker* who attacks in the reasonable though mistaken belief that his act is justified (e.g., Elliot Ness when he mistakes Al Capone's identical twin brother for Capone himself). Second, IAs excused for absence of moral responsibility are *Nonresponsible Attackers*. These include the *Insane Attacker* and the *Hypnotized Attacker* whose will is utterly subjected to that of another.[20] Finally, there may be instances of *Compelled Attackers* whose unjustified action is the result of irresistible coercion. (To count as an IA, however, one's action must be fully excused; and it may be doubted whether duress can ever be fully exculpatory. One who kills an IB to survive is surely under extreme duress, but it is difficult to believe that the duress fully excuses the killing rather than simply mitigating the agent's guilt.)

The IA constitutes a case in which JUSTICE fails to account for common intuitions about the justifiability of self-defense. Although the theory deals well with the problem of the Culpable Cause, there are other cases in which it also fails to support common intuitions about self-preservation. Let us define a *Threat* as a person who is causally involved in a threat of harm to another though not through his agency. An example of an Innocent Threat is the *Innocent Projectile* whose body is, through no fault of his own, hurled against another person.[21] To kill the Innocent Projectile to prevent oneself from being crushed by his body counts as an act of self-preservation since self-defense, as defined earlier, must be directed at an Attacker (i.e., one who poses a threat through his agency). Because the Innocent Projectile is not morally culpable for the threat he poses, JUSTICE offers the potential Victim no justification for killing him; yet most people believe that it would be permissible to kill him.

There is at least a strong presumption that the justification, if there is one, for the self-defensive killing of an IA should be the same as that for the self-preservative killing of an Innocent Projectile. Consider, for example, an Innocent Projectile who, while working on his roof, is swept off by a sudden gust of wind that could not have been anticipated. Unless an innocent passerby kills the Projectile by

20. These categories may overlap. A small child firing a real gun that he believed to be a toy might be both a Nonresponsible and an Inadvertent Attacker.

21. This example first appears in Robert Nozick, *Anarchy, State, and Utopia* (Oxford: Basil Blackwell, 1974), pp. 34–35. Also see Nancy (Ann) Davis, "Abortion and Self-Defense," *Philosophy and Public Affairs* 13 (1984): 175–207, esp. pp. 190–92.

deflecting his body with a pitchfork she is carrying, she will be crushed and killed. Next consider a Hypnotized Attacker who, through no fault of his own, has been subjected to an irresistible form of mind control and commanded to fling himself from a roof onto an innocent passerby (an attack that the Hypnotized Attacker will survive unless the passerby deflects him with a pitchfork). Most people believe that the passerby would be justified in deflecting the falling person in each case. And it is hard to find a relevant difference between the two, for each is simply being hurled (or made to hurl himself) toward his Victim by forces entirely beyond his control. Hence the presumption that the same form of justification should apply to both.[22]

Since most of us believe that it is permissible to kill an IA in self-defense and an Innocent Projectile in self-preservation, and since JUSTICE cannot support these intuitions, we should search for an alternative account of the right of self-defense. Suppose that we find one that supports these intuitions and also justifies the self-defensive killing of a CA. Would that make JUSTICE superfluous? I think not. For an account (such as the ORTHODOX VIEW) that applied the same justification to self-defense against a CA that it applied to self-defense against an IA would miss the fact that the CA's culpability contributes importantly to the justification for killing him. That the two cases are relevantly different is revealed by the fact that they are governed by different restrictions. There is, for example, no duty to retreat from a confrontation with a CA, particularly if the CA has invaded a domain where one has a special right to be, such as one's home, even if one could retreat in complete safety. By contrast, there is always a duty to retreat from a confrontation with an IA when retreat promises a probability of self-preservation that is at least equal to that offered by self-defensive killing. A second difference is that the proportionality restriction governing self-defense against an IA is stronger than that

22. In developing a theory of corrective justice, Jules Coleman argues that excuses based on absence of agency can defeat liability to compensate a Victim for injuries caused whereas excuses based on absence of culpability cannot. Thus if one were to develop an *ex ante* theory of preventive justice corresponding to the *ex post* theory of corrective justice (a possibility that I will explore in Sec. VI), it would discriminate morally between the IA and the Innocent Projectile since the latter, but not the former, would have an excuse that defeats liability. Coleman, however, treats the Hypnotized Attacker as an Innocent Threat rather than an Innocent Attacker, claiming that "moving one's body under a hypnotic trance is . . . not something one does." Thus he would deny that the comparison between the Hypnotized Attacker and the Innocent Projectile supports the claim that self-defense against an IA has the same justification as the self-preservative killing of an Innocent Projectile. Although I cannot take up his challenge here, I would argue against the view that absence of agency in the causation of harm defeats responsibility and therefore liability in a way that absence of culpability cannot. See Jules Coleman, *Risks and Wrongs* (Cambridge: Cambridge University Press, 1992), p. 260. Also see pp. 261–66 and 334–35.

governing self-defense against a CA. If, for example, one could be certain of avoiding being killed by an IA by killing him but could alternatively reduce the risk of being killed to an almost negligible level by incapacitating rather than killing him, one might be required to forgo the option of killing. No such requirement would apply if the Attacker were a CA. Finally, while the right to kill a CA clearly extends to third parties, it is not obvious that it is permissible for third parties to intervene on behalf of the Victim of an IA. While there is some tendency to grant third parties the right to kill an IA, the only cases in which commonsense morality unequivocally endorses this are those in which the Victim is someone to whom the third party is specially related—for example, a spouse or child.[23]

This suggests that, while a plausible account of the right of self-defense will have to contain elements not found in JUSTICE, it will also have to incorporate JUSTICE's claim that the Attacker's culpability is a crucial part of the justification for self-defense in certain cases. In short, it may seem that the best comprehensive account of the right of self-defense must be a hybrid theory incorporating the central elements of both JUSTICE and some other theory. This, however, may not be unproblematic. For it may be the case that considerations of justice not only fail to support but actually oppose the killing of an IA in self-defense. For the IA is, like the IB, morally innocent. Thus, unless there is something about the IA that makes killing him in self-defense relevantly unlike killing an IB in self-preservation and nullifies the injustice we find in the latter, then JUSTICE itself may oppose the self-defensive killing of an IA, since JUSTICE is concerned with preventing culpable action from harming the morally innocent. If JUSTICE were to oppose the self-defensive killing of an IA in the way that it opposes the self-preservative killing of an IB, then not only would this cast doubt on the theory, since most of us believe that self-defense against an IA is justified, but it would also mean that JUSTICE could not be coherently combined with a theory that permits self-defense against an IA. The hope for a comprehensive hybrid theory would be undermined.

The challenge, then, is to find a relevant difference between the IA and the IB that renders the permissibility of killing the former in self-defense compatible with the general impermissibility of intentionally killing the latter in self-preservation. This challenge is more formidable than it may seem. For the intuitive view is that, for it to be justifiable intentionally to attack or harm a person, there generally

23. George Fletcher disagrees. He writes that "an adequate theory" of self-defense "would permit third parties to intervene against (and not for) the psychotic aggressor" (and, by implication, against other sorts of IA). See his "Proportionality and the Psychotic Aggressor: A Vignette in Comparative Criminal Theory," *Israel Law Review* 8 (1973): 367–90, p. 375. I will return to this in Sec. VI.

must be some fact about her or her action that renders her liable to attack or harm.[24] A self-preservative attack on an IB, however, addresses only a wholly "external" fact about her—namely, the fact that she occupies a position in the causal structure such that killing her will be instrumental in averting a threat to the agent. But the relevant fact about the IA—that he is engaged in an attack—seems similarly "external" since his moral innocence absolves him of all personal responsibility for it. As George Fletcher puts it, discussing a case in tort law in which a defendant has been the innocent cause of harm, "it is true that the defendant . . . 'acts' so as to cause harm, but if the act is excused, the harm is no more attributable to the defendant than it is to the victim. . . . Excused harms are circumstantial, for they derive not from the responsible choice of the defendant, but the defendant's and plaintiff's being thrown together by circumstance."[25]

To see how similar the IA and the IB can be, consider again what I called Causes. In addition to Culpable Causes, there are also *Justified Causes,* whose past justifiable action is causally responsible for a present threat of harm that would wrong the Victim, and *Innocent Causes,* whose past action, which was unjustified but fully excused, is causally responsible for a present threat of harm that would wrong the Victim. Relative to the present threat caused by his past action, the Innocent Cause is both innocent, since he is not culpably responsible for the threat, and a Bystander, since he is now no part of the threat. And, because he is an IB, most people would strongly condemn killing him in self-preservation. If, for example, the miner in our earlier example had caused the collapse of the shaft in a way that was fully excused (i.e., the collapse was the unforeseen result of action that was not malicious, reckless, or negligent), then he would be an Innocent Cause whom it would be wrong to kill to get the oxygen tank. But notice that the only difference between an Innocent Cause and an IA is that the former's innocent, threat-creating action lies in the past while the latter's is occurring at present. And that seems far too insubstantial a difference to support the view that killing the IA is justified while killing the Innocent Cause is not.[26]

24. Compare Thomas Nagel's claim that "whatever one does to another person intentionally must be aimed at him as a subject, with the intention that he receive it as a subject. It should manifest an attitude to *him* rather than just to the situation." See his "War and Massacre," *Philosophy and Public Affairs* 1 (1972): 123–44, p. 136.

25. George P. Fletcher, "The Search for Synthesis in Tort Theory," *Law and Philosophy* 2 (1983): 63–88, esp. pp. 67–68. (There is a tension between this passage and Fletcher's view, cited earlier in n. 23, of the justifiability of self-defense against an IA.) Coleman, p. 263, makes a similar point about Innocent Threats whose excuse is absence of agency.

26. An IB is normally distinguished from an IA or an Innocent Threat by virtue of the claim that each of the latter is, while the former is not, part of a threat to the agent. Noam Zohar has, however, argued that there is no sharp distinction between

Unless we can find a defensible means of discriminating morally between killing an IA and killing an IB, we face a dilemma. One option is to stand firm in rejecting the idea that a plea of self-preservation can make it permissible intentionally to kill an IB. In that case we will have to accept a radically attenuated right of self-defense that does not permit the killing of an IA. The alternative is to accept the permissibility both of killing an IA in self-defense and of intentionally killing an IB in self-preservation.

## IV. THE PERSONAL PARTIALITY ACCOUNT

We may distinguish several types of justification for self-defensive violence. There are, first, justifications that focus on facts about the Attacker. These are *target-centered accounts* of the right of self-defense. Second, there are accounts that focus on facts about the Victim. These are *agent-centered accounts*. And there are other accounts that emphasize considerations beyond those involved in the immediate conflict between Attacker and Victim—for example, *impersonal* accounts that justify self-defense only on those occasions when it promises to produce the greatest good, or *conventionalist* accounts that appeal to principles people could rationally agree to adopt or the adoption of which would best achieve certain aims. I will sketch an account of the latter type in the final section.

JUSTICE is a target-centered account, since it takes the Attacker's moral culpability to be crucial. Since, however, the IA is morally innocent, it may be that self-defense against him is more likely to be justifiable in agent-centered rather than target-centered terms. The most initially attractive agent-centered account appeals to the claim that people are entitled, at least with regard to certain types of choice, to give priority to their own interests and values over those of other people. Virtually all of us accept some view of this sort. We do not believe that we are always morally required to give the interests of all other people the same weight that we give our own. We believe, in short, that, in certain contexts, people are permitted to act on the basis of a certain degree of personal partiality.

The permissibility of partiality is, however, limited in scope in two ways. First, the degree of partiality that is permitted on those occasions when it is legitimate is limited. The additional weight one may assign to one's own interests or values may not exceed a certain

---

being and not being part of a threat. His argument focuses on the case of the *Innocent Obstructor,* someone who nonculpably (and perhaps passively) impedes one's escape from a threat. Is such a person a part of, or causally involved in, the threat one faces? Zohar's argument reinforces the suggestion that the gap between the IA and the IB is insubstantial. See Zohar, "Collective War and Individualistic Ethics: Against the Conscription of 'Self-Defense,'" *Political Theory* (in press).

specified amount. Second, there are certain choices—for example, those concerning university admissions or the assignment of grades—that partiality may not permissibly influence. It has been suggested, however, that at least some situations in which self-defense is possible are ones in which the agent may act on the basis of a limited degree of partiality.[27] If, for example, a person faces a potentially lethal attack from an IA, considerations of justice do not favor either party. Assuming that any difference in the comparative value of the two lives that are at stake does not greatly favor the preservation of the IA, the Victim may then be entitled to give priority to the preservation of her own life, simply because she values it more. Call this account of the right of self-defense the Personal Partiality Account (henceforth PARTIALITY).

As this example suggests, PARTIALITY offers a reason for thinking that self-defense against an IA is permissible. It also has the further advantage that it supports another intuitively plausible claim: namely, that the moral reason that the Victim has to resist the IA is also available to the IA as a justification for resisting the Victim's counterattack. Thus, while the IA is merely excused for his initial attack, he is, on this view, justified in defending himself against the Victim's counterattack. Recall, furthermore, the earlier suggestion that the justification for killing an IA in self-defense and that for killing an Innocent Projectile in self-preservation should be the same. It is another virtue of PARTIALITY that it justifies each in the same way, by appealing to the legitimacy of the agent's preference for her own life.

There are, however, objections to this account. One derives from the fact that there are cases in which the harm that the Victim's self-defensive action would cause to the Attacker would exceed the harm that the Victim thereby avoids by more than the amount that is permitted by legitimate personal partiality. In these cases, PARTIALITY cannot sanction self-defensive action. While this may or may not seem acceptable in the case of self-defense against an IA, it is clearly unacceptable in the case of self-defense against a CA. This objection might be overcome by combining PARTIALITY with JUSTICE to form a hybrid account. While PARTIALITY would provide justification in cases involving IAs and Innocent Threats, JUSTICE would cover cases involving CAs and Culpable Causes. In the latter cases, partiality would either reinforce considerations of justice (e.g., by supporting the Victim's right of self-defense against a CA) or be nullified by the agent's culpability (thus one type of case in which an agent clearly may not appeal to personal partiality to justify self-defensive action is that in which a CA faces a justified counterattack by his Victim).

27. See Davis, pp. 175–207.

The hybrid account would, moreover, have an explanation of why self-defense against an IA is governed by stricter restrictions than self-defense against a CA (namely, that the former is justified by considerations of partiality while the latter is justified by considerations of justice). Yet the restrictions that PARTIALITY imposes on self-defense against an IA may seem excessively strict. According to PARTIALITY, one may kill an IA in self-defense provided that the expected harm one causes to the IA does not exceed, by an amount determined by the extent of permissible partiality, the expected harm to oneself that one thereby averts. According to PARTIALITY, therefore, the probability that the IA's attack will otherwise prove lethal must pass a certain threshold in order for self-defensive killing to be justified; and there is also a limit to the number of IAs that one may kill in self-defense. It is not obvious, however, that the proportionality limit on self-defense against an IA coincides with the degree of permissible partiality. Suppose, for example, that one must kill ten IAs to prevent oneself from being killed. Most people appear to believe that it would be permissible to kill all ten; but it is not clear that partiality permits one to value one's own life at more than ten times the value of each IA's life.[28]

The fundamental objection to PARTIALITY, however, is that it is unclear how it can justify killing an IA in self-defense without also justifying killing an IB in self-preservation. The defender of the theory will of course want to claim that the former is the sort of case in which partiality is permitted while the latter is not. But the problem is that PARTIALITY itself provides no ground for distinguishing between the two cases. If there are considerations that give scope to partiality in the former case but not the latter, then these considerations, and not PARTIALITY, will provide the deep account of the permissibility of self-defense against an IA. That account remains to be discovered.

Some who believe that PARTIALITY itself provides the deep account of the permissibility of self-defense against an IA might be willing to accept that the theory also justifies killing an IB in self-preservation. In that case, however, the theory could not be coherently combined with JUSTICE since, even if the latter does not condemn self-defense against an IA, it surely condemns killing an IB as a means of self-preservation. (It is no use claiming that considerations of justice would simply override partiality in the case of an IB, for then they would do so in the case of an IA as well.) But considered on its own, without

28. The degree of permissible partiality is probably not fixed or invariant but varies with the nature of the case. It may be, for example, that a greater degree is allowed in distributing benefits when not all can be benefited than in determining whom to save from harm and whom to allow to suffer harm when not all can be spared. And a greater degree may be allowed in the latter case than in cases involving the redirection of a threat. And so on.

the ability to discriminate morally between the CA and the IA but with the implication that it is permissible intentionally to kill an IB, or perhaps numerous IBs, in self-preservation, PARTIALITY is wholly implausible.

## V. DEONTOLOGICAL CONSTRAINTS

As a rule, partiality seems to have no justificatory effect whatever with respect to an act that falls within the scope of a deontological constraint. As most theorists understand it, commonsense morality has constraints against both doing harm and intentionally harming. Nevertheless, it is widely accepted that certain acts of doing harm to the innocent are somehow not subject to a constraint provided that the harm is, in the relevant sense, unintended. The most familiar example is the act of the *Tactical Bomber* who drops his bombs with the intention of destroying a military facility while foreseeing that, to his regret, the bombs will also unavoidably kill a small number of innocent civilians living nearby. (If he could destroy the facility without harming the civilians, he would do that instead.) The possibility that some acts of this sort, that do harm without intending it, are not subject to a constraint suggests a way of distinguishing morally between killing an IA in self-defense and killing an IB as a means of self-preservation. This strategy does not seek a relevant difference between the IA and the IB themselves but instead argues that there is a relevant difference between the two modes of agency. For recall that the presumption against shifting harms is primarily a presumption against shifting harms by certain forms of agency. Thus it seems weaker in cases of harming the innocent by the redirection of a threat and may not apply at all to ducking harm. It is therefore possible that, even though both the self-defensive killing of an IA and the self-preservative killing of an IB involve harming the innocent, the latter involves a mode of agency that is specially objectionable while the former does not. If so, that could explain why partiality may help to justify the killing of the IA but not the IB. The absence of a constraint would provide the negative part of the justification, giving partiality a justificatory role, while partiality itself would then provide the Victim (though not disinterested third parties) a positive reason for shifting potential harm from herself to the IA.

Let us assume that there is a special constraint or presumption against harmful intentional killing. Sometimes the killing of an IB in self-preservation does not violate this constraint. Suppose, for example, that one discovers a bomb in the fourth-floor room one is in. If it is about to explode and one throws it out the window, thereby foreseeably killing the gardener working in the bushes below, one's act may be wrong but not because it involves intentional killing. In other cases, however—for example, if one seizes an IB to use as a shield against a bullet or (as in the notorious case of *Regina v. Dudley*

*and Stephens*)[29] if one eats an IB to avoid starvation—killing the IB does violate the constraint. Let us say of the latter cases that they involve killing an IB as a *means* of self-preservation. The argument suggested above may be expressed by claiming that, whereas killing an IB as a means of self-preservation always violates a constraint, so that the appeal to partiality has no power to justify it, killing an IA in self-defense need never violate the constraint and hence may be defended by appeal to partiality.

The claim that the killing of an IA in self-defense need not be an intended effect follows from what is becoming the orthodox understanding of an intended means.[30] According to this view, what an agent intends as her means may (at least sometimes) be specified by the narrowest description of the relevant effect that is compatible with the agent's believing the effect, under that description, to be causally instrumental to the achievement of her end. In the case of self-defense against an IA, for example, the Victim may believe that only the incapacitation of the IA is causally necessary for self-defense; thus she may intend to incapacitate but not to kill the IA, even if she foresees that incapacitating him will involve killing him. This view is not a contemporary innovation but has formed the basis for the views of many Catholic theorists, including Aquinas, about the right of self-defense even against a CA. Many of these theorists claim, for example, that killing in private self-defense is permissible, but only when the killing is a side effect of action intended only to incapacitate the Attacker or ward off the attack.[31]

This ground for discrimination between killing an IB as a means of self-preservation and killing an IA in self-defense fails, for two reasons. First, the only cases involving the self-defensive killing of an IA that it treats as relevantly different from killing an IB as a means of self-preservation are those in which the Victim does not intend to kill the IA. But most of us believe that it is permissible to attack an IA with the intention of killing him if killing him is necessary in the circumstances for effective self-defense. Thus this argument will appeal only to those in the Catholic tradition who believe that intentional killing is always objectionable, even in self-defense.

29. 14 Q.B.D. 273 (1884).

30. See, e.g., John Finnis, "Intention and Side-Effects," in *Liability and Responsibility*, ed. R. G. Frey and Christopher W. Morris (Cambridge: Cambridge University Press, 1992), pp. 32–64; and Quinn.

31. See, e.g., G. E. M. Anscombe, "War and Murder," in her *Collected Philosophical Papers*, vol. 3, *Ethics, Religion, and Politics* (Minneapolis: University of Minnesota Press, 1981), pp. 51–61, p. 54; Robert L. Phillips, *War and Justice* (Norman: University of Oklahoma Press, 1984), pp. 44–46; and John Finnis, Joseph Boyle, Jr., and Germain Grisez, *Nuclear Deterrence, Morality, and Realism* (Oxford: Oxford University Press, 1987), pp. 310–18. Aquinas's view is found in *Summa Theologiae*, II, ii, q. 64, a. 7.

Second, and more important, the narrow conception of an intended means that underlies the claim that the killing of an IA need not be intended also makes it possible that the killing of an IB need virtually never be an intended means even of foreseeably lethal self-preservative action. For in virtually all cases, it is not the *killing* of the IB that is causally required for self-preservation but, rather, some more narrowly described effect that is itself then causally or logically sufficient for the IB's death. For example, in seizing an IB as a shield, one may intend only that her body should stop the bullet from penetrating one's own body. One need not intend that she be killed; thus, if she were to be wearing a bullet-proof vest and were therefore to survive, none of one's intentions would thereby be frustrated. Self-defense and self-preservation, therefore, can be entirely symmetrical with respect to intention.

Because most acts that we describe as killing an IB as a means of self-preservation need not, in fact, have the killing (or the death) of the IB as a strictly intended effect, we must, if we believe these acts are nevertheless subject to a special constraint, revise our understanding of the constraint. This, however, must be the task of a different work.[32] For present purposes, what is important to note is that, unless there is some difference in the mode of agency that we have yet to notice, we must expect the revised constraint to apply equally to killing an IA in self-defense.

A possible difference might be that, while those acts that are naturally (though perhaps inaccurately) described as killing an IB as a means of self-preservation involve *using* the IB for one's own purposes, killing an IA in self-defense does not use the IA, just as killing a CA does not use him. Yet the reason that the Victim cannot be said to use the IA is simply that the concept of using is such that, whatever one does to a person in addressing a problem, one cannot be said to be using that person if the person is himself the problem. Thus the claim that the agent uses the IB but not the IA simply restates the fact that the IA constitutes the threat to which self-defense is a response while the IB is not a part of the threat to which self-preservative killing is a response. It does not explain why this difference should be morally relevant.

## VI. THE UNJUST THREAT ACCOUNT

As we have seen, the ORTHODOX VIEW takes this difference between the IA and the IB to be morally fundamental. We have also seen, however, that the ORTHODOX VIEW is untenable. It may be possible, though, to introduce refinements that will yield a view that is plausible.

32. The approach I favor is briefly indicated in n. 5.

In particular, one might modify the claim that it is the fact that a person is engaged in causing harm or poses a threat of harm that makes it permissible to attack or kill him so that it is instead being engaged in morally unjustified harmful action, or posing an unjustified threat, that compromises a person's immunity, making it permissible to attack or kill him. A view of this sort has, for example, been advanced by Elizabeth Anscombe, who, couching her view within a definition of innocence, writes that "what is required, for the people attacked [e.g., in self-defense] to be non-innocent in the relevant sense, is that they should themselves be engaged in an objectively unjust proceeding which the attacker has the right to make his concern; or—the commonest case—should be unjustly attacking him."[33] I take it that the force of the word "objectively" is to emphasize that one can engage in unjust action without culpability. Thus, on this view, it is the fact that the IA's action is unjustified that constitutes the critical asymmetry between him and his Victim, making it permissible to shift inevitable harm away from the Victim to him. A similar view prevails in the law of torts, where objective fault on the part of an injurer is normally sufficient, even in the absence of culpability, to make the injurer liable to make reparation for harms he has caused.

This account of the right of self-defense is target-centered, since it makes the unjustifiability of the Attacker's action the factor that compromises his immunity. It is, moreover, not just an account of self-defense but also covers certain cases of killing in self-preservation. It offers a justification, for example, for killing an Innocent Projectile. For the threat that the Innocent Projectile poses is unjustified in the sense that he would be culpable for it were he not excused by virtue of the fact that it is not the result of his agency. Thus this account, which we may call the Unjust Threat Account (henceforth, UNJUST THREAT, satisfies the presumption noted earlier that the justification for killing an IA in self-defense should also apply to the killing of an Innocent Projectile in self-preservation. It also appears to have the implication, which will be welcomed by many, that the permission to kill an IA or an Innocent Threat extends not only to the potential Victim but also to third parties. For, being target-centered, it appears to provide an agent-neutral reason for acting. Finally, while UNJUST THREAT justifies the defensive killing of an IA and the self-preservative killing of an Innocent Projectile, it appears not to provide any justification for killing an IB, at least as a means of self-preservation.

There is an immediate objection to this view, which is that it fails to justify self-defense in certain cases in which it seems justified.

33. Anscombe, "War and Murder," p. 53. Notice that an IA is not innocent in Anscombe's sense of the term.

Consider, for example, the case of the Tactical Bomber mentioned in the previous section. The Tactical Bomber is a Justified Attacker since, although he wrongs the civilian victims of the raid by harming them, his action is nevertheless morally justified; indeed, it may be morally required. Most of us, however, believe that, provided they are fully morally innocent, the civilians would be justified in killing the Bomber in self-defense. Except perhaps in conditions of extremity, they are not required to sacrifice themselves to facilitate his justified action. Yet, because his action is justified, UNJUST THREAT does not permit the civilians to engage in self-defense against him.

There is a further view that is closely related to UNJUST THREAT that appears to deal more satisfactorily with this case. According to this view, which has been defended by Judith Thomson, the right of self-defense is a corollary of the possession of other rights. An Attacker (or a Threat) who is about to violate the right of another thereby loses the right not to be harmed in whatever way is necessary to prevent the violation (provided that the harm is necessary and proportionate). It is this fact about the Attacker—his lacking a right—that makes self-defense permissible. Thus, even though this account focuses initially on the rights of the Victim, it is in fact target-centered. Call it the Rights-based Account (henceforth, RIGHTS).[34]

According to Thomson, rights are not absolute; that is, they may sometimes be justifiably infringed.[35] Thus it seems possible both that the civilians have the right not to be killed and that the Bomber is justified in doing what will kill them. If so, then even though the Bomber's action is justified, it threatens to violate the civilians' rights; therefore, according to RIGHTS, he loses his right not to be killed and they are justified in killing him in self-defense. Yet, as Thomson develops it, RIGHTS also holds that whenever a Victim is justified in killing an Attacker in self-defense, it is impermissible for the Attacker to counterattack in self-defense, since he lacks a right not to be attacked.[36] Thus RIGHTS implies that, while the Bomber is justified in bombing the military facility, he is not permitted to defend himself against the civilians who try to kill him in order to stop him. It is, however, implausible to suppose that, while the civilians retain their rights of self-defense against the Bomber, his justified action causes him to forfeit his right not to be killed and hence also his right of self-defense against them.

34. Judith Jarvis Thomson, "Self-Defense," *Philosophy and Public Affairs* 20 (1991): 283–310. The necessity and proportionality restrictions mentioned parenthetically in the text are suggested on pp. 301 and 302–3.

35. Judith Jarvis Thomson, *The Realm of Rights* (Cambridge, Mass.: Harvard University Press, 1990), chaps. 3 and 4.

36. Thomson, "Self-Defense," pp. 304–5.

This shows, I think, that Thomson has failed to develop the most plausible version of RIGHTS. While she holds (roughly) that one loses a right whenever one threatens a right, it is more plausible to claim that one loses a right only when one *unjustifiably* threatens a right. If one justifiably threatens a right, one retains one's own. On this view, the Bomber retains his right not to be killed. Yet, just as the civilians need not lose a right in order for it to be permissible for the Bomber to do what will kill them, so the Bomber need not lose a right in order for it to be permissible for the civilians to kill him in self-defense. In short, the Bomber is justified in dropping his bombs; but, because this will infringe the rights of the civilians, they are justified in trying to protect their rights by killing him in self-defense; since his action is justified, however, he does not lose his rights; hence he is also justified in counterattacking in self-defense.

Even if it is revised in this way, RIGHTS remains vulnerable to objections. Its central claim is that self-defense is justified because the Attacker, by threatening to violate (or infringe) the right of another, thereby forfeits his own right not to be attacked. In order to have an acceptably wide range of application, however, it must hold that one can violate a right not only without culpability but even without agency. Thus Thomson claims that both the IA and the Innocent Projectile are potential violators of rights. Yet there is reason, deriving from Thomson's own analysis of rights, for thinking that only morally responsible agents can violate rights. According to Thomson, a person's having a right just is for certain others to be morally constrained in certain ways.[37] But a moral constraint can apply only to the action of a responsible agent. Neither a falling boulder nor a charging tiger can be subject to a moral constraint; thus neither can violate a right.[38] Since a Nonresponsible Attacker is no more a moral agent than a tiger, and since an Innocent Projectile is no more an agent than a falling boulder, it seems that at least some IAs and some Innocent Threats cannot violate rights and hence cannot forfeit them. If so, RIGHTS cannot have the desirable range of application Thomson attributes to it.

RIGHTS also depends on a sharp distinction between rights violators and Bystanders. But since a violator need be neither culpable nor an agent, not only are there cases in which it is unclear (and perhaps indeterminate) whether a person is a violator or a Bystander,[39] but

37. Thomson, *The Realm of Rights*, p. 77.

38. Thomson claims both that responsible agency is not necessary for the violation of a right and that third parties may act in other-defense to protect rights that the Victim is unable to protect. If we add to these the claim that animals have rights, it follows that one is always permitted (and perhaps sometimes obligated) to prevent carnivorous animals from killing their prey.

39. Compare n. 26.

also there are cases in which the theory seems to get the classification wrong. Suppose, for example, that we are all in the proverbial overcrowded lifeboat, having all clambered aboard more or less simultaneously. Unless one of us leaves or is ejected, the boat will sink and we will all die. I would like to heave you overboard but only if that would be morally permissible. What does RIGHTS imply about this? While at first glance you have the hallmarks of an IB, it is also true that the weight your body adds to the boat threatens my life; so perhaps you are an Innocent Threat, a potential violator of my right not to be killed. Of course, I stand to you in the same relation in which you stand to me. Thus the question is: are we all potential violators of one another's rights, so that each forfeits his or her right not to be killed by the others, or are we all IBs who are equally menaced by external circumstances but retain our rights vis-à-vis one another? Morally speaking, only the latter is plausible; but, if being a potential rights violator requires neither culpability nor agency, then it is hard to deny that each threatens to violate every other occupant's rights. (Thomson might reply that, while each constitutes a threat to the others, none has a right not to be threatened in that way. But to see that this is implausible, imagine that one occupant is a misanthrope with a life jacket who got on board in the hope of sinking the boat. That person clearly threatens the others' rights. Yet Thomson denies that whether or not a person is a violator of rights can depend solely on whether or not he or she acts culpably.)

A third objection to RIGHTS emerges when we consider that self-defense is in all cases subject to a requirement of minimal force—namely, that one must not cause a greater harm if one can defend oneself equally effectively by less harmful means. Thomson agrees.[40] But notice what this means. It means that what rights an Attacker forfeits, and therefore what rights he has, depend on what options the Victim has. Suppose, for example, that one is threatened by an IA wielding a knife. If one is a master of the martial arts who can easily disarm and incapacitate the IA without killing him, then one may not dispatch him with a revolver. If, by contrast, one can effectively defend oneself only by killing him, then one may use the revolver. Thus whether or not the IA has a right not to be killed depends on something as contingent as how skillful the Victim is in the martial arts.

This is a strange conclusion if one assumes the traditional conception of moral rights as protective barriers that individuals possess by virtue of their having certain natural properties, such as certain capacities for experience and action, and which ground and explain certain

40. Thomson, "Self-Defense," p. 301.

duties of others. But it is less strange if we recall Thomson's view that a person's rights are nothing more than a set of moral constraints on others. The question what rights a person has can be answered only be specifying the ways in which others are constrained in their treatment of her. How do we know, for example, that the Victim of an attack by an IA has a right not to be killed by him? We know this because the IA's attack is (by definition) impermissible. A more revealing question, however, is how we know what right the IA forfeits. Thomson deals with the simple case in which it is necessary to kill the IA to prevent him from killing the Victim. In this case, the IA threatens the Victim's right not to be killed and forfeits his own right not to be killed. But IAs do not always or necessarily forfeit this right. Whether or not an IA forfeits a right and which right or rights he forfeits depend on how the Victim is constrained—that is, on, inter alia, how the constraints of necessity, proportionality, and minimal force apply in the particular case. If, for example, killing is unnecessary, disproportionate, or excessive, then the IA retains his right not to be killed.

It is revealing how the foregoing reasoning proceeds. It is critical to the justification of self-defense against an IA offered by RIGHTS that the IA should, by threatening the Victim's rights, forfeit certain rights (i.e., those that would otherwise forbid harming him in the ways required by self-defense). How do we know that this happens? Answer: because there is no constraint that forbids the Victim to attack or harm or kill him (depending on the case). How do we know there is no constraint? RIGHTS provides no answer. For we cannot, without circularity, answer that there is no constraint because the IA lacks a right. There is, in fact, no independent theory of what constraints there are or what rights people have, nor any independent theory of forfeiture (e.g., one that makes forfeiture a function of culpability, or culpability that would result in wrongful harm). In short, the claim that the Victim is under no constraint not to attack or harm or kill the IA, and hence that it is permissible for her to attack or harm or kill him, is entirely ungrounded. What RIGHTS offers is an ingenious exercise in begging the question. According to RIGHTS, the IA in effect forfeits the right not to have done to him whatever we judge, on the basis of considerations that are wholly unspecified by the theory, that it is permissible for the Victim or third parties to do to him.

So, even if RIGHTS seems to deal acceptably with the case of the Tactical Bomber, it must be rejected. Another option is to make what may seem an ad hoc revision to UNJUST THREAT so that it stipulates that the presumption against shifting inevitable harms is defeated, other things being equal, not only in cases in which a person unjustifiably poses a threat but also in cases in which a person is justified in posing a threat that will nevertheless wrong a person by harming her (or infringe her rights, where a right is understood in the traditional

way as the *source* of a constraint). The added element in the revised account permits the civilians to kill the Bomber in self-defense since the latter's action, though justified, threatens to harm them in a way that would wrong them (or infringe their rights). Whether it permits the Bomber to fight back as well is unclear. Clearly the Bomber cannot defend himself against the self-defensive action of the civilians on the ground that their action is unjustified. Thus, if he is to be justified in counterattacking in self-defense, it must be because the civilians' action, though justified, threatens to harm him in a way that would wrong him (or infringe his rights). It is difficult to say whether this is plausible. It is not, I believe, obviously implausible.

This revised version of UNJUST THREAT (henceforth, UNJUST THREAT 2) is closely related to the theory of corrective justice advanced by Jules Coleman as the best foundation for the liability rules of tort law. That theory holds that injurers have a duty to repair losses (i.e., harms) they have caused either by means of wrongdoing (i.e., action that is objectively unjustified) or by wronging (i.e., action that is contrary to rights).[41] In neither case is culpability required for liability. Thus an act of wrongdoing may be fully excused and an instance of wronging may even be fully morally justified; yet the injurer is liable as a matter of justice since the harm he has inflicted is a wrongful one. The similarity between this theory and UNJUST THREAT 2 is in fact so close that one may be tempted to regard the latter as merely an *ex ante* version of the theory of corrective justice—a corresponding account of *preventive justice*. This, however, is not quite right. For it is not true in every case in which an injurer has an *ex post* duty to repair a harm he has caused that the victim has an *ex ante* permission to prevent the harmful action. One of Coleman's own examples illustrates this point. Suppose that a diabetic has nonnegligently lost his supply of insulin and that, to prevent himself from lapsing into a coma, he takes some of another person's insulin without her permission, leaving her enough to meet her own needs.[42] This is a case in which, while the diabetic has a duty *ex post* to repair the loss his action imposes on the owner, it would be wrong for the owner to prevent the diabetic from wronging her by taking her insulin without her consent. Thus one cannot infer the permissibility of preventing a harm from the fact that the harm wrongs the victim and imposes on the injurer a duty *ex post* to compensate the victim for the harm.[43]

41. Coleman, pp. 324–26, 332, and 361.

42. Ibid., p. 282.

43. Coleman claims that the ground of the diabetic's duty to repair the owner's loss is that his taking the insulin infringes her right, albeit justifiably. Because he assumes the owner retains her right to the insulin, he also assumes that she would be acting within her rights if she were to exclude the diabetic's use of the insulin, though he

UNJUST THREAT 2 licenses shifting harms in certain cases involving Attackers and Threats but never, it seems, does it provide a justification for intentionally harming a Bystander. (This, of course, is its principal attraction: that it distinguishes between the self-defensive killing of an IA or the self-preservative killing of an Innocent Threat and the self-preservative killing of an IB.) But, like UNJUST THREAT and RIGHTS, it attributes no significance to culpability. It does not, therefore, have the resources to justify killing a Culpable Cause (e.g., the culpable miner) in self-preservation. This, I believe, is a serious, perhaps fatal, limitation.

Although Thomson develops RIGHTS in a way that seems to make the categories of rights violator and Bystander mutually exclusive, it may be possible for RIGHTS to justify the self-preservative killing of a Culpable Cause despite the fact that he is a Bystander. For the assumption that the two categories are mutually exclusive may be a mistake. It might be argued that, by killing the Culpable Cause now, one prevents him from violating one's rights through his past action. That Thomson accepts that one can now prevent past action from violating a right in the future is suggested by the fact that she accepts that an individual who in the past set in train a sequence of events that will lead to the death of an innocent person can now prevent himself from becoming a murderer by intervening to stop the sequence of events.[44] If, similarly, killing a Culpable Cause now can prevent him from violating one's rights through his past action, then RIGHTS can justify killing the Culpable Cause in self-preservation. For, although he is a Bystander, he is also a potential rights violator who therefore forfeits certain rights of his own.

This suggests that RIGHTS has an important advantage over UNJUST THREAT 2. But even here the theory's promise is illusory. For it extends the range of justification too far. If it is possible for present action to prevent past action from violating a right in the future, and if one can violate a right without culpability, then RIGHTS will also justify the self-preservative killing of an Innocent Cause—for example, it would justify killing the miner to get his oxygen tank even if his past action had accidentally caused the collapse of the shaft in a

---

concedes that "right holders may unreasonably or wrongly insist upon enforcing their rights," in which case it may be reasonable "to subject the right holder to moral, if not legal liability, for the consequences of the exclusion" (p. 301). I believe, by contrast, that, because it would be wrong, in the circumstances, for the owner to exclude the diabetic from the use of the insulin, she cannot have the right of exclusion. The diabetic's need causes her right to lapse. Yet he clearly owes her compensation. If this cannot be because he infringes her right (which has lapsed), perhaps it is because he deprives her of it. Her loss of the right itself is a serious unmerited loss.

44. Thomson, "The Trolley Problem," in her *Rights, Restitution, and Risk* (Cambridge, Mass.: Harvard University Press, 1986), pp. 98–99.

way that was not malicious, reckless, or negligent. Indeed, since RIGHTS can justify the self-defensive killing of a Justified Attacker (e.g., the Tactical Bomber) and the self-preservative killing of a Justified Threat, it can, if extended to justify the self-preservative killing of a Culpable Cause, also justify the self-preservative killing of a Justified Cause—for example, it would justify killing the miner to get his oxygen tank even if he had caused the collapse of the shaft as an unforeseen result of justified action intended to save hundreds of trapped miners. Since Innocent Causes and Justified Causes are both IBs, RIGHTS now justifies certain instances of killing an IB as a means of self-preservation. This, I assume, is unacceptable.

The failure of RIGHTS to distinguish appropriately between Culpable Causes on the one hand and Innocent and Justified Causes on the other suggests that, even if UNJUST THREAT 2 could be suitably revised so that it too could justify the self-preservative killing of a Culpable Cause (perhaps on the ground that this would prevent unjustified action—albeit past action—from causing harm), it too would then also justify the self-preservative killing of Innocent and Justified Causes. For, like RIGHTS, UNJUSTIFIED THREAT 2 attributes no significance to culpability and it is only the culpability of the Culpable Cause that distinguishes him from the Innocent Cause and the Justified Cause. If we believe, as most of us do, that there is a morally important difference between these cases, then we must accept both that culpability is significant and therefore that neither RIGHTS nor UNJUST THREAT 2 can provide a satisfactory account of the permissibility of shifting harms in self-defense or self-preservation.

A further deficiency of UNJUST THREAT 2 that is traceable to its failure to recognize the significance of culpability is that, while it justifies self-defense against both IAs and CAs, it cannot justify or explain the differences between the restrictions that apply to the two types of defense. A comprehensive account of the rights of self-defense and self-preservation must, it seems, recognize the significance of culpability. One possibility might be to conjoin UNJUST THREAT 2 with JUSTICE (both of which are target-centered accounts) to form another hybrid theory—call it HYBRID. According to HYBRID, the fact that someone poses an unjust threat or justifiably poses a threat of harm that would wrong someone is sufficient to lower moral barriers to harming him; if he is culpably responsible for a threat to someone, that lowers the relevant barriers to an even greater extent. Together these claims support intuitively plausible conclusions about the cases considered so far. The self-defensive killing of an IA and the self-preservative killing of an Innocent Threat are supported by appeal to the fact that the targets pose an unjustified threat; the self-defensive killing of a Justified Attacker (such as the Tactical Bomber) is supported by appeal

to the fact that his action threatens to harm someone in a way that would wrong her; and the self-defensive killing of a CA and the self-preservative killing of a Culpable Cause are supported by appeal to the fact that the targets are culpably responsible for threats that can be averted only by killing them. HYBRID thus rejects the view that there is a single, unified foundation for the right of self-defense and proposes instead that self-defense is differently justified in different cases. This has obvious advantages. For example, because the different justifications may carry different restrictions, HYBRID can explain why the restrictions that apply in some cases are more stringent than those that apply in others.

What HYBRID does not justify is as important as what it does justify. It provides no justification for killing an IB as a means of self-preservation, or for killing a Just Attacker in self-defense, or for killing either an Innocent Cause or a Justified Cause in self-preservation. (The assumption in the latter case is that UNJUSTIFIED THREAT 2 cannot justify self-preservative violence against a Cause, since a Cause poses no threat. However, HYBRID's other element—JUSTICE—does justify the self-preservative killing of a Culpable Cause.)

Earlier the question arose whether JUSTICE could be coherently combined with a theory, such as UNJUST THREAT 2, that justifies the self-defensive killing of an IA. The worry is that JUSTICE may itself condemn the killing of an IA as unjust. I believe, however, that this doubt can be dispelled. JUSTICE holds that, when either the guilty or the innocent must suffer harm, it is permissible, other things being equal, to ensure that it is the guilty party that is harmed. In the conflict between an IA and his potential Victim, neither party is guilty. JUSTICE, therefore, is silent. The Victim may, of course, become guilty in attacking the IA if her action is wrong, but another theory is required to generate the condemnation and thus the claim that the Victim is culpable. If, therefore, JUSTICE is paired with UNJUST THREAT 2, which implies that the self-defensive killing of the IA is permissible, then JUSTICE has no ground for condemning it. Thus UNJUST THREAT 2 and JUSTICE appear to be compatible.

HYBRID gets as close as any theory I can think of to organizing our intuitions under a modest set of general principles. It is, however, not without problems. The first of these may not be serious; indeed, some would not regard it as a problem at all. This is that, while HYBRID justifies self-defense against an IA, it does not seem to permit a self-defensive counterattack by the IA.[45] I have noted that it is unclear

45. Among those who believe that the IA has no right of self-defense are Thomson, "Self-Defense," p. 304; Alexander, p. 1179; and George Fletcher, "The Right and the Reasonable," *Harvard Law Review* 98 (1985): 949.

whether UNJUST THREAT 2, and thus HYBRID, can justify self-defense by a Justified Attacker (e.g., by the Tactical Bomber against the civilians). There is even less reason to believe that it can justify self-defense by an IA. For there is more reason to suppose that the Justified Attacker is wronged if he is killed to prevent his justified action than there is to suppose that the IA is wronged if he is killed to prevent his unjustified action. But, if the IA is not wronged by the Victim's self-defensive action, then HYBRID provides no justification for self-defense by the IA.

Still, if the civilians are justified in trying to kill the Bomber in self-defense while the Bomber is also justified in trying to kill them in self-defense, this shows that two parties can each be justified in trying to kill the other in self-defense. I believe that, while this is clearly controversial, the conflict between the IA and his Victim is of this sort. The claim that an IA may permissibly be killed in self-defense is hard enough to justify; the further claim that the IA forfeits or loses his own right of self-defense would be even harder to justify. And it is intuitively implausible. The IA and his Victim are both innocent victims thrown together by circumstances for which neither is morally responsible. Even if the IA loses his immunity to attack, he retains a right of self-defense.[46] If HYBRID cannot accommodate this view, this counts against the theory.

A second possible problem concerns HYBRID's implications for other-defense or third-party intervention in a conflict between an IA and his Victim. UNJUST THREAT 2 is a target-centered theory; hence it presumably grounds an agent-neutral reason for killing the IA rather than an agent-relative one (as it would if it were agent-centered). If the reason it offers is agent-neutral, then the reason should extend to third parties. Thus UNJUST THREAT 2, and hence HYBRID, presumably justifies intervention by disinterested third parties against an Attacker or Threat whenever self-defensive or self-preservative action would be justified. This seems implausible. I assume, for example, that HYBRID implies that the civilians may kill the Tactical Bomber in self-defense. Since, however, the Bomber is a Justified Attacker, it seems implausible to suppose that third parties also have the right to kill him. But suppose, for the sake of argument, that they do. If we also assume—what I take to be true—that the Bomber himself has a right of defense against the civilians, then it follows that third parties are also permitted to kill

46. Recall that these remarks are confined to cases in which it is inevitable that either the IA or the Victim will be killed. There are other cases in which the IA may not engage in self-defense (e.g., when the nonlethal harm he would cause would be significantly greater than that which he would avert). But there are also similar cases in which the Victim may not defend herself against the IA.

the civilians in defense of the Bomber. Nor does the absurdity end here. If the Bomber retains his right of self-defense against the civilians, then presumably he may also defend himself against third parties who intervene on behalf of the civilians. And, if other-defense is justified whenever self-defense is, then third parties may intervene on his behalf. In that case, third parties would be justified in trying to kill other third parties who were justifiably trying to kill the Bomber. (It would also be a mistake, though a less glaring one, to suppose that, while third parties are permitted to intervene on behalf of the civilians, they would not then be justified in intervening to defend the Bomber. For this suggests a moral asymmetry between the Bomber and the civilians that simply is not there. The only plausible view is that third parties are not permitted to intervene on behalf of either.)

If HYBRID implies an agent-neutral justification for self-defense, then it also implies the permissibility of third-party intervention against an IA. Many will find this plausible. I have doubts. Again, self-defense by the Victim is difficult enough to justify; the claim that others may intervene on her behalf is even more controversial. Admittedly, our intuitions here are rather fluid. It may seem, for example, that third parties who are specially related to the Victim may intervene on her behalf. Similarly, however, it may also seem that third parties specially related to the IA may intervene to defend him against the self-defensive action of the Victim. Or consider variations in numbers. Suppose, for example, that a single IA threatens ten Innocent Victims. It seems clearly permissible for a third party to intervene on behalf of the Victims. But now imagine that the numbers are reversed, so that ten IAs threaten a single Victim. While it may be permissible for the Victim to kill all ten, I find it implausible to suppose that a disinterested third party would also be permitted to kill all ten.[47] Yet HYBRID seems to imply that the Victim's justification extends equally to third parties.

This case, in which ten IAs threaten a single Victim, may be one in which all third-party intervention is disallowed. For, if the Victim were about to kill all ten IAs, I doubt that it would be permissible for a third party to kill her to prevent it. This shows, among other things, that our intuitions do recognize a moral asymmetry between the IA and his Victim; for, while a third party may kill a single IA to defend ten Victims, he or she may not kill a single Victim to defend ten IAs.

47. As developed by Thomson, RIGHTS implies not only that a Victim may kill any number of IAs to defend her own life but also that third parties may kill any number of IAs (or any number of Innocent Threats) to defend her. This seems very implausible.

(The asymmetry also shows up in the fact that, while most of us believe that a single Victim may kill ten IAs if that is necessary for self-defense, it is less clear that an IA may kill ten Victims if that is necessary to defend himself from their counterattack.)

While we clearly do believe that such an asymmetry exists, the main doubt about HYBRID is whether it really provides a satisfactory explanation of it. HYBRID justifies self-defense against an IA on the ground that he poses an objectively unjustified threat. Initially at least, his Victim, like the IB, poses no threat to anyone. This, according to HYBRID, is the basis of the asymmetry. (Moreover, if the Victim engages in self-defense, she then poses a threat, but one that is justified. Hence an asymmetry remains even if she comes to pose a threat.) But, since the IA is wholly absolved of moral responsibility for the threat he poses, it is not obvious that the simple fact that he poses it is sufficient to rebut the presumption against shifting harm, making it permissible to kill him. Admittedly, in some cases this fact seems sufficient to justify shifting a harm. If, for example, a Victim can either acquiesce in suffering a significant but nonlethal harm at the hands of an IA or, by inflicting a harm half as severe on the IA, reduce the severity of the harm she will suffer by half, then she seems clearly justified in harming the IA. And the fact that he threatens her unjustifiably provides a reasonably plausible explanation of why she may force him to share the inevitable harm. But, when the harm is severe and cannot be divided, the mere fact that the IA's action (or the Innocent Threat's movement or location) is objectively unjustified may seem insufficient to override the presumption against doing harm in order to avoid allowing harm to occur.

## VII. JUSTIFICATION AND EXCUSE

If we are not altogether satisfied that HYBRID has located a compelling difference between the IA and his Victim, or between the IA and the IB, perhaps we should consider the possibility of resolving the dilemma stated at the end of Section III by accepting that the killing of an IA in self-defense should be evaluated in the same way as the killing of an IB as a means of self-preservation. If we are to assimilate one to the other, I think we should conclude that killing an IA in self-defense is wrong in the same way that killing an IB as a means of self-preservation is. For it is even more repugnant to common sense to suppose that the killing of an IB as a means of self-preservation is normally justified. If the killing of an IA is in fact wrong, then perhaps JUSTICE alone provides a comprehensive account of the right of self-defense after all.

Even, however, if the killing of an IA is wrong, it does not follow that a Victim who kills an IA in self-defense is culpable, or deserves

blame or punishment. For it might be that killing an IA is, while unjustified, nevertheless excused.[48] The usual defense of this position appeals to the fact that the decision to kill an IA in self-defense is taken under extreme duress, so that there is a weak sense in which self-defensive action is compelled and hence involuntary. As George Fletcher notes, "Stressing the element of involuntariness is but our way of making the moral claim that he [the Victim] is not to be blamed for the kind of choice that other people would make under the same circumstances."[49]

Fletcher has, however, argued against the view that self-defense against an IA is excused in the following way:

> It would follow that third persons, unrelated to the defendant [i.e., the Victim], would incur criminal liability if they intervened on his behalf. And why shouldn't they? If he is in the wrong, why should anyone have a right voluntarily to intervene on his behalf? Indeed, the implication would be, that according to the German and Soviet theory of self-defence, the aggressor [i.e., the IA] would acquire a right of defence against the defendant's wrongful (*rechtswidrig*) resistance. And third parties would have a derivative right to intervene on his behalf. All of these implications conflict with our sense of justice in the situation. If anyone is to be assisted it is the party struggling to save his life against the psychotic [i.e., Innocent] aggressor.[50]

The idea here is that, if the IA's Victim is merely excused in engaging in self-defense, then the IA, now confronted with a wrongful counterattack, must be morally justified in engaging in self-defense against the counterattack. But that cannot be right. For merely being confronted with an unjustified attack cannot, on the view we are considering, generate a justification for self-defense. If it did, then the Victim would (contrary to the view under consideration) be justified rather than merely excused for responding to the IA's initial attack, since that attack is also unjustified, though excused. In short, the position of the IA faced with the Victim's counterattack seems to be largely symmetrical with that of the Victim faced with the IA's initial attack. Thus both the IA's initial attack and the Victim's counterattack must be wrong but excused. So, if the Victim is only excused for trying to kill in self-defense, then the IA will also be only excused for trying to kill in self-defense against the Victim's counterattack. Violence at

48. On the distinction between justification and excuse, see Kent Greenawalt, "The Perplexing Borders of Justification and Excuse," *Columbia Law Review,* vol. 84 (1984). One theorist who argues that the killing of an IA in self-defense is normally excused but seldom justified is Alexander, pp. 1177–89.

49. Fletcher, *Rethinking Criminal Law,* p. 856.

50. Fletcher, "Proportionality and the Psychotic Aggressor," p. 375.

each level of escalation will be wrong, though excused. Therefore intervention by disinterested third parties on behalf of either combatant is impermissible.

Fletcher's mistake is to switch theories in mid-argument. He begins with the idea that self-defense against an IA is excused; but, when dealing with self-defense by the IA, he switches to the "German and Soviet theory," which regards self-defense against an IA as justified rather than excused. But the Victim's self-defensive attack is by hypothesis merely excused, so she incurs no liability in engaging in self-defense and is thus herself an IA.

So far, then, the view that self-defense against an IA is wrong though excused appears defensible. Possibly in the end we may be driven to accept it. But it represents a significant surrender. For what most of us believe is that the self-defensive killing of an IA is not wrong but justified, provided that certain conditions are met. We also believe that killing an IB as a means of self-preservation is normally not only unjustified but also culpable—that is, not excused. (Recall that the killing of an IB as a means of achieving political ends, even just and important ones, is condemned as terrorism.) Yet if the killing of an IA is excused on grounds of duress, then it is difficult to see why killing an IB as a means of self-preservation should not also be excused on the same grounds. For the threat to the agent is the same in both cases: death. If it is conceded that the threat of death overwhelms the will in the one case, the same must be conceded in the other as well. Perhaps it could be argued that there is this difference: that the circumstances of an imminent attack preclude the possibility of deliberation that may be present in cases in which killing an IB is necessary for self-preservation. But this, surely, is a contingent general difference and there are bound to be instances in which the opportunities for deliberation about killing an IB as a means of self-preservation are as limited as those present in the case of self-defense.

There is one other point that should be noted. Suppose that a case can be made that, in wartime, soldiers who fight in an unjust cause should generally be treated as if they were IAs rather than CAs.[51] This is because many, though not all, fight only because they have been manipulated by deception, conditioning, and indoctrination or coerced by threats and pressures that in fact most people are unable to resist. Because they act under duress or ignorance or both, they may be to some extent excused and therefore may be regarded as IAs, even though most are hardly paradigm cases of IAs. But, if we grant that for practical purposes they are more like IAs than CAs, and if we accept that defensive violence against IAs is only excused rather than

51. This issue is discussed in my "Innocence, Self-Defense, and Killing in War."

justified, then we will be committed to a pacifist position that most will be reluctant to accept.

## VIII. A CONVENTIONALIST EXPLANATION?

There are thus reasons to reject the view that the self-defensive killing of an IA is wrong. Yet we have been unable to find a fully convincing justification for the view that it is permissible. Perhaps we should try to understand what motivates us to believe so strongly in the permissibility of the self-defensive killing of an IA even though we can provide no fully persuasive defense of our belief.

One suggestion might be that one tends to identify oneself with the Victim rather than with the IA, probably because one feels that it is less likely that one will ever actually be an IA than that one will be the Victim of one. And, given the identification with the Victim, personal partiality prompts one to favor the permissibility of self-defense: one values one's own life more than that of a person, probably a stranger, by whom one may be innocently attacked. We are, of course, capable of resisting the operation of partiality—for example, we do not believe that partiality justifies killing an IB as a means of self-preservation. So there must be another factor. I believe that this other factor is our tendency to assimilate the killing of an IA to the paradigm case of justified self-defense: the killing of a CA. The justifiability of self-defense seems so obvious in the paradigm case that the intuitive sense of justification lingers even as we move some distance away from the paradigm. Our robust sense of the justifiability of self-defense against an IA may, in short, be the result of overgeneralizing from the paradigm case.

Perhaps we would be more conscious of the apparent illegitimacy of extrapolating from the paradigm case to the case of the IA if the effects of our doing so were bad. But having a rule that permits self-defense against any threatening person other than a Just Attacker probably has considerable social utility. For in practice one can seldom be certain that an Attacker is indeed fully innocent. (Although we use locutions such as "fully innocent" and "largely innocent," it is culpability rather than innocence that comes in degrees. An IA is by definition fully innocent.) And, since cases involving IAs are rare, a case in which there is uncertainty is more likely to involve a CA than an IA. Since people are unlikely to internalize a rule that permits self-defense in all cases except those in which it is certain that the Attacker is fully innocent, justice and utility may be best served if we can converge on a rule that permits self-defense in all cases except those involving a Just Attacker. (There will seldom be the same sort of uncertainty about whether an Attacker is a Just Attacker, since the Victim must normally have done something wrong in order for an attack on her to be just.) In short, the rule

permitting self-defense against an IA may simply be a convention that is acceptable because of its social utility.

If we accept this explanation, then we accept that at least some parts of our morality are not susceptible of direct or intrinsic justification. Some moral theorists—certain relativists, rule-consequentialists, and contractualists—hold that the whole of morality is like this: that its rules are justifiable only because of our having converged on them as conventions in our society, or because of the social utility of our conforming to their demands, or because we could rationally agree to accept them. Let us, for convenience, refer to this large and untidy collection of theories as "conventionalist" theories. I believe, in contrast to conventionalist theories, that much of morality can be justified independently of its acceptance or of the effects of its acceptance. Yet in certain gray areas, where our intuitions seem confused or incapable of being unified under a principle or theory that can be directly justified, it may be necessary to resolve issues that might otherwise remain indeterminate by accepting a rule that constitutes a natural point of convergence because of the beneficial effects of its acceptance. This may be the best form of defense we can give for our belief in the permissibility of self-defense against an IA.

The objection to this proposal, however, is that it seems to presuppose that the wrongness of intentionally killing an IB in self-preservation is likewise conventionally grounded. For suppose that this were not true—that is, suppose that intentionally killing an IB is wrong for reasons that are independent both of our agreeing that it is wrong and of the effects of our agreeing that it is wrong. Since we have been unable to find a fully convincing intrinsic moral difference between killing an IA and killing an IB, the view that killing an IB as a means of self-preservation is wrong for nonconventional reasons is in conflict with the claim that there is a conventional justification for killing an IA. For, unless the case of the IA is relevantly different from that of the IB, then the nonconventional objection to killing the latter will also apply to killing the former. But this would undermine the conventional justification for killing an IA in self-defense, for the conventional elements of morality cannot be allowed to conflict with the nonconventional elements. A convention is acceptable only if it is compatible with nonconventional morality. If this is right, then we must either give up the idea that there is a conventional justification for killing an IA in self-defense or else accept that the justification for prohibiting the killing of an IB as a means of self-preservation is also merely conventional. In the latter case, the moral significance of the distinction between the IA and the IB would be merely conventional.

Those of us who believe that there are elements in morality that are intrinsically rather than conventionally justified will be profoundly reluctant to accept that killing an IB as a means of self-preservation

is wrong for purely conventional reasons. For, if anything is wrong independently of the utility of the prohibition, it is the intentional killing of an IB. Hence the conventionalist defense of the permissibility of killing an IA in self-defense may seem a credible option only for those who are conventionalists about the whole of morality. The rest of us may be tempted to conclude that the problem of the IA reveals both an incoherence in commonsense morality and a lack of grounding for an important set of discriminations in the law.