Preemptive Rules and the Scope of Defensive Rights

Yitzhak Benbaji

Forthcoming in BJ Strawser, Ryne Jenkins and Mike Robillard (eds.) *Who Should Die?: Liability and Killing in War* (Oxford University Press,).

This essay advances a morality of defensive harm, which I call "Rule-SD." Rule-SD resolves in a new way two types of difficult cases. It entails that if certain conditions are met, a defender has the right to kill a man who is innocently falling on her, if this is necessary for her survival. Moreover, Rule-SD yields the "free competition resolution" in some symmetrical cases; it implies that, two people who innocently threaten each other, whose position vis-à-vis each other is identical in every factual and (therefore) normative respect, might have a right to kill each other if necessary for their survival.

Rule-SD's core claim is that, a defender's right of self-defense might arise from a "preemptive rule" rather than from facts about the liability of the attacker. In cases I call "one vs. one (or less) self-defense circumstances" (or sometimes "the designated circumstances"), the defender might be subject to a rule that permits self-preference; the rules states that, if certain conditions are met, a defender might treat the innocent attacker (or threatener) as if he were liable to necessary defensive harm. The rule further allows her to disregard some of the interests of innocent bystanders. (The permission is restricted. The rule permits self-preference as long as it does not foreseeably or intentionally harm bystanders by using force against them, cheating them, stealing from them, or any other illicit (but well defined) means.) Being subject to the rule that permits "restricted self-preference," the defender is exempted from the duty to ascertain that the attacker is indeed liable as well as from taking into account non-violent, indirect harms to bystanders.

The normative basis of Rule-SD is twofold. First, Rule-SD appeals to a Razian conception of rules as preemptive reasons.[1] Under the designated circumstances, the rule that permits restricted self-preference is normally justified in the sense that the defender might follow it, instead of deliberating on the merits of the case and responding to the first order reasons this deliberation would have exposed. In particular, the rule grounds a reason to follow it *and* a reason not to be guided by the first order reasons which the rule mediates. The other normative basis that underlies Rule-SD is the liability view of defensive harm, which has been elaborated in the last two decades, primarily by Jeff McMahan.[2] Put very simply: Attacker A is liable to intentional killing if killing A is a necessary and proportionate means to avert an unjustified threat T, and if A is *sufficiently* responsible for T; incidental harm to bystanders might be justified if it is proportionate, i.e., if it is a lesser evil.

Now, Rule-SD establishes the preemptive rule which allows the defender's restricted self-preference in one vs. one circumstances on the following *empirical* generalizations. First, threatening individuals are usually (but *not* always) at least minimally culpable for the threat they pose, hence, the liability view entails that these individuals are typically liable to defensive harm. Moreover, usually, if the defensive harm to the threatener is necessary, and the incidental harm to bystanders is nonviolent and indirect (like causing loss to the attacker's dependents), the defensive harm is proportionate. Second, ordinary defenders facing one vs. one circumstances are under stress; their deliberative capacities are limited. Moreover, exempting them from the duty to make sure that their attacker is liable to defensive harm and from taking into account non-violent indirect harms to bystanders will enable them to be more effective in promoting their self-interest. That is, in most cases, the exemption will allow them to be more effective in doing the right thing.[3]

* Acknowledgments… The research was supported by The Israeli Research Foundation, grant number 304/15

[1] Joseph Raz, *Practical Reason and Norms* (Princeton: Princeton University Press, 1990), 38ff, and Raz, *The Morality of Freedom* (Oxford University Press, 1986), 39ff.

[2] Jeff McMahan, "The Ethics of Killing in War," *Ethics* 114 (2004): 693–733; Jeff McMahan, "Self-Defense and the Problem of the Innocent Attacker," *Ethics* 104 (1994): 252–90, and elsewhere.

[3] As BJ Strawser points out, we might worry that the risk of doing wrong here is significant enough such that it might warrant a cautionary principle against the kind of preempitve rule that I advocate. So

Rule-SD thus implies that if the following epistemic conditions are met, an ordinary defender facing the designated circumstances is subject to a preemptive rule that permits restricted self-preference: *first*, the defender does not know whether the attacker or threatener is liable to killing and she does not know whether the non-violent indirect harm to bystanders is proportionate; *second*, the defender justifiably believes that making sure that the attacker is liable to killing and that the incidental harm to bystanders is a lesser evil would be difficult and costly for her.

I noted that Rule-SD resolves innocent-threat cases in a novel way. The following is a typical innocent threat case.[4]

> *Innocent Threat*: A defender realizes that a man is falling on him. Unless the falling man is blown to pieces, he will crush the defender to death. The man's falling is faultless and involuntary.

Existing moral theories offer at least three different resolutions of innocent threat cases. According to "free competition theories" both sides lose their right not to be attacked by each other but, nevertheless, they retain their right of self-defense: the defender might kill the falling man, whereas the falling man can use a gun, should he have one handy, to prevent the defender from killing him.[5] Second, according to the "fair procedure theories," in cases where the unavoidable harm is indivisible and the parties are morally equal, the parties ought to find a fair mechanism for resolving the conflict.[6] Innocent threat cases provoke a third

---

just how much "most" actually is will matter. It would also matter how difficult and costly it usually is to make sure that an attacker is liable.

[4] Cf., Judith J. Thomson, "Self-Defense," *Philosophy & Public Affairs* 20 (1991). I discuss innocent threat and symmetrical cases in Yitzhak Benbaji, "A Defense of the Traditional War-Convention," *Ethics* 118 (2008): 464-95. The present essay offers a novel rule-based morality of harm that implies different resolutions of these cases than the ones offered in the former publication. This change results from the thought-provoking discussion in Victor Tadros, *The Ends of Harm* (Oxford: Oxford University Press, 2011), 197-216.

[5] Benbaji, "A Defense of the Traditional War-Convention," Nancy Davis, "Abortion and Self-Defense," *Philosophy & Public Affairs* 13 (1984): 175–207, 192–93, Jonathan Quong, "Killing in Self-Defense," *Ethics*, 119 (2009), 507-537.

[6] The argument advanced in Gerhard Øverland, "Contractual Killing" *Ethics* 115 (2005): 692-720 suggests such a fair procedure solution. For related discussions, Susanne Burri, 'The Toss-Up Between a Profiting, Innocent Threat and His Victim', *The Journal of Political Philosophy* 23 (2015): 146-165. Bradley J. Strawser, *The Bounds of Defense: Killing, Autonomy, and War*, forthcoming manuscript.

response, according to which the defender has a right to defensively harm the falling man while the falling man has *no* right to fight back.[7]

I noted that Rule SD offers a novel resolution to symmetrical cases. In those cases the vital interests of the parties are in conflict and their position vis-à-vis each other is identical in every factual and (therefore) normative respect. While the "free competition" view is not self-evidently obvious to all even in a symmetrical case like *Bear*, many believe it to be the only plausible resolution of this type of cases:

> *Bear*: You and I are running away from a bear that is chasing us. The bear is faster and more powerful than both of us, but it needs only one of us for dinner, so either you or I will be the bear's food, but not both. Whether you or I survive depends on who is faster. We do not have to outrun the bear to ensure our survival; all we need to do is to outrun each other.[8]

The competition ought to be fair: I can put on my own running shoes to gain speed, but cannot toss yours into the fire to slow you down. *Mutatis mutandis*, the same is true of you.

In contrast, it is the fair procedure intuition that governs the resolution of another symmetrical case, *Flotsam*:

> *Flotsam*: You and I are trapped on a sinking piece of flotsam and soon we will be submerged in the high sea where we will drown almost immediately. The flotsam can support the weight of either one of us, but not both.[9]

Here, the parties should find a fair mechanism for resolving their conflict, for instance tossing a coin.

Rule-SD supports free competition in innocent threat cases where both parties are subject to the self-preference rule. However, unlike free competition theories, under Rule-SD, restricted self-preference is permissible only if the epistemic conditions are met: the defender does not know whether the attacker is liable, and is justified in assuming that it would be costly and difficult to gain this piece of knowledge.

---

[7] For example, Judith J. Thomson, "Self-Defense," *Philosophy and Public Affairs* 20 (1991): 283-310; Helen Frowe, "Equating Innocent Threats and Bystanders," *Journal of Applied Philosophy* 25 (2008): 277–90.

[8] Tadros, *The Ends of Harm*, 209-11. Note that this is *not* the classic case of ducking harm as defined in C. Broose and R. A. Sorensen, "Ducking Harm," *Journal of Philosophy* 85 (1988): 115-134.

[9] Tadros, *The Ends of Harm*, 203.

Rule-SD draws a distinction between *Bear*-like and *Flotsam*-like cases in which the epistemic conditions are met. Restricted self-preference (and therefore, free competition) is permissible for *uncertain* deliberators in *Bear* but *not* in *Flotsam.* In both cases, we do not threaten each other. We are threaten by "nature"; we are bystanders vis-à-vis the threat. In *Flotsam*, however, free competition involves *use of force*; I should push you into the water in order to survive. Hence, the restricted self-preference rule, which does not allow use of force against bystanders, does not apply to this case. In *Bear*, self-preference meets the restrictions: by outrunning you I do not use force against you. I merely indirectly and non-violently harm you. Hence, I am governed by a rule that permits self-preference.

I will proceed as follows. In sections (1) and (2), I present and criticize two alternative accounts of defensive harm: a partialist account of the right of self-defense recently elaborated by Jonathan Quong, and the "autonomist" analysis of symmetrical cases recently developed by Victor Tadros. In Section (3), I present Rule-SD and offer a novel analysis of innocent threat cases. Section (4) enriches Rule-SD and then uses it to resolve *Bear* and *Flotsam*.

1. The Partialist Account of Defensive Harm

Rule-SD is an impartialist morality of defensive harm: the rights a defender possesses are grounded in the agent-neutral reasons that apply to her. According to "partialism," the main rival of this view, impartialism overlooks a crucial element of the morality of defensive harm: egocentric reasons ground the moral permissions we have in various types of circumstances, in particular, self-defense circumstances.[10] In this section I briefly present a partialist account of the right to defensive harm, I then show that partialism faces a deep difficulty.

As I understand it, partialism asserts that since "egocentric reasons" have moral weight, "egocentric behavior" might be permissible or justified in some circumstances. I should therefore say more about these concepts. A simple example of an egocentric reason is the fact that the pain I am experiencing is bad for me. This fact is a *prima facie* reason for me

[10] A version of partialism is developed in Jonathan Quong, "Killing in Self-Defense," *Ethics*, 119 (2009), 507-537.

to eliminate this pain.[11] Now, if I deserve the pain, my reason to eliminate it is *merely* egocentric. In such a case, others have no reason to relieve my pain and might even have a reason to prevent me from ridding myself of it.[12]

My reason is egocentric in virtue of the fact that its specification requires a pronominal back reference to me: that the pain is bad *for me* is a reason *for me*—but for no one else—to rid myself of it. Similarly, I have a reason to defend myself from a just attack, even if I am a culpable attacker, and therefore liable to defensive killing. My self-interest grounds an agent-relative reason to defend myself, while others have an agent-neutral reason to prevent me from doing so.

The notions of agent-neutrality and agent-relativity allow for a characterization of impartiality and impartial behavior, on the one hand, and egocentric behavior, on the other. An impartial agent cares about her interests to the extent that her interests are impartially important. In contrast, an egocentric (or partial) agent cares about protecting and promoting her own interests more than she cares about protecting and promoting strangers' interests, merely because of her special relation to her own interests. Obviously, egocentric agents are able to overcome their self-love. An agent acts impartially if she is guided solely by agent-neutral reasons; the mere fact that an action might also promote her interests rather than someone else's may not have had any impact on her choice. An agent exhibits egocentric behavior if her action is motivated also by egocentric reasons; her action is *strongly* egocentric if it is solely motivated by such reasons.

The partialist morality of defensive harm attaches moral significance to egocentric reasons: suppose I am threatened by a person who is falling on me by no fault of his own. Partialists would usually deem the self-interest I have in my own life sufficient to grant me

[11] The term "agent relativity" was coined in T. Nagel, *The Possibility of Altruism* (Princeton: Princeton University Press, 1970). My characterization of it is drawn from P. Pettit "Universality without Utilitarianism," *Mind* 96 (1987), 74–82. For illuminating discussion, see, Michael Ridge, "Reasons for Action: Agent-Neutral vs. Agent-Relative," *The Stanford Encyclopaedia of Philosophy* (Winter 2011 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2011/entries/reasons-agent/>.

[12] See, Christine M. Korsgaard, "The Reasons we can Share: An Attack on the Distinction between Agent-Relative and Agent-Neutral Values," *Social Philosophy and Policy* 10 (1993): 24-51.

permission to kill the falling man in self-defense; *I* possess a right of self-defense, but my right does not extend to anyone else. In other words, my right of self-defense does not imply a right of other-defense: a third party has no right to kill the falling man in my defense.

Partialism must address questions such as the number of innocent attackers the defender is permitted to kill in order to survive and whether defensive killing is permissible when killing the attacker will yield fatal consequences for many others. In solving this problem, they assume that agent-neutral reasons can be weighed against agent-relative ones. If the (morally weighted) costs the defender would bear had she avoided harming the attacker are sufficiently low, the agent-neutral reasons against harming the attacker outweigh the agent-relative reasons in favor of such harm. Indeed, in innocent threat cases, agent-relative values are sufficiently weighty to break the tie: the defender has the right of self-defense—she has the right to prefer herself—merely because of the importance of her life to her.[13]

The difficulty partialism faces is simple: Why is it that the preference that people have for their own well-being does not ground a similar set of permissions when killing bystanders is necessary for their survival?[14] Why is it impermissible to kill the bystander in *Innocent Bystander*?

> *Innocent Bystander:* A defender realizes that a man is falling on him and is about to crush her. The defender can block the falling man by using a bystander as a human shield, in which case the bystander would be crushed to death.

Partialists might respond as follows: the agent-relative values involved in *Innocent Threat* are just as weighty as the agent-relative values involved in *Innocent Bystander*. Yet, despite the moral reason one has to prefer one's own life, self-preference in *Innocent Bystander* is impermissible because manipulatively killing the bystander is more seriously wrong than the eliminative killing of attackers/threateners.

But this response seems to fail; the prohibition on self-preference seems to be much more straightforward than partialism would like us to believe: a defender has *no moral reason*

[13] Quong, *Killing in Self-Defense*, 514-23.
[14] Cf., Tadros, *The Ends of* Harm, p. 209. Tardros raises this difficulty regarding Quong's partialist distinction between doing and allowing.

to prefer himself over the bystander in *Innocent Bystander*. Indeed, a proponent of this view, Jonathan Quong, realizes that the agent-relative value can ground a moral right of self-defense only if it can be objectively, rather than subjectively, specified. He insists that "there is an objective sense in which each person may permissibly attach much greater weight to their own life in comparison to the lives of others".[15] He would acknowledge that a sadist who genuinely needs to cause pain to someone has *no* moral reason to cause pain to anyone. Acting on such a reason is malicious, precisely because of the nature of sadistic needs.[16] But, I submit, the same might be true of many cases of eliminative killing: I have no *moral* reason to kill you even if I genuinely need something you own and I cannot attain it unless I kill you. Partialists cannot explain this conviction.

Consider partialism's implication in a less trivial case: a bank robber, who initially was a potential attacker but is now a liable defender; as things stand, the guard of the bank threatens the robber's life. Partialism asserts that if necessary for his survival, the robber has a moral reason to kill the guard. Partialism means by this that the robber has *some* moral reason to kill the guard. It does not follow that this reason wins out – the moral reasons against killing the guard outweigh the reason for killing him.

But even this weak claim seems false: the fact that the robber has an interest in his own life— an interest that constitutes the robber's agent-relative reason to attach greater weight to his own life— doesn't seem to give the robber any *moral* reasons at all to kill the guard. For, the very fact that the robber is liable to killing seems to imply that even if he has a reason to attach greater weight to his own life, in these circumstances, he has no *moral* reason to do so. The prohibition on killing bystanders in self-preservation is similar: despite the partialist claim to the contrary, there is *no* moral reason that speaks in its favor.

Quong might argue in response that egocentric reasons need not always have moral weight—they need not weigh against every other conflicting moral consideration—for it to be the case that these types of reasons sometimes have moral weight. Sometimes – under

[15] Quong, *Killing in Self-Defense*, 517

[16] Note, though, that sadism is not *intrinsically* bad. If a person meets his sadistic needs by watching horror movies, we might have a moral reason to provide him with such movies.

clearly delineated circumstances considerations of the type "…it would be me…" are morally weighty. Alas, this line of argument might be effective, only if Quong can offer a plausible explanation as to why in some circumstances egocentric reasons are "suddenly" weightless.

The morality I advance in Section 3—Rule-SD—maintains that egocentric reasons as such are morally insignificant. It therefore concurs that under self-defense circumstances, the special relationship between the defender and her interests is morally irrelevant. Yet, Rule-SD implies that while partialism is wrong, the verdicts it advances in innocent threat cases are mostly correct: ordinary uninformed defenders have the right to exhibit egocentric behavior—i.e., to kill their attackers if necessary—even where they could have easily known that their attackers are not liable to defensive harm, but justifiably believe that they cannot gain this piece of knowledge.

Beforehand, I will discuss *Bear* and *Flotsam*, the symmetrical cases that motivate Rule-SD, and the autonomist resolution of these cases.

2. Symmetrical Cases and the Autonomist Morality of Harm

It should come as no surprise that some symmetrical cases are resolved by free competition and others by fair procedure. The following *Crash* cases are an example:

> *Crash*$_1$: Following the instructions of a negligent inspector, two agents are driving two trains toward each other on the same track. A total-loss crash where both drivers are killed is bound to happen, unless one driver stops the other by using deadly force. Cooperative fair procedure is impossible as the drivers cannot communicate with each other. A non-cooperative randomization—where one of the drivers, or both, flip a coin in order to decide whether to use lethal force against the other driver—would increase the chance that both drivers will avoid action and end up dead. In general, then, anything other than mutual defensive action will increase the likelihood that the drivers will be killed.

Certainly, free competition is permissible since it is the only way to prevent the suboptimal outcome in which both drivers are killed. Compare, however, *Crash*$_1$ to *Crash*$_2$:

> *Crash*$_2$: Like *Crash*$_1$, except that non-cooperative randomization would increase the chance that one of the drivers survives, as mutual defensive action is most likely to cause the drivers to kill each other.

The key normative consideration in each *Crash* case is aggregation. The same outcome-based considerations that support a free competition resolution in *Crash*$_1$ justify a fair procedure

resolution in $Crash_2$. Any morality of defensive harm, according to which the defender's choice is *prima facie* justifiable if it minimizes harm to innocents, easily distinguishes between $Crash_1$ and $Crash_2$. In contrast, the symmetrical cases under discussion, *Bear* and *Flotsam*, resist an aggregation-based analysis; instead, in these cases the key normative consideration is fairness in the distribution of harm of a *given* size.

Commonsense morality distinguishes between *Bear* and *Flotsam* for reasons related to self-ownership. While both of us have an equal claim on the flotsam, I have an exclusive claim on my person and on my physical and mental abilities. As far as I do not exercise violence against you, use anything that belongs to you, or interfere with your affairs by coercion, deception, etc., it is permissible for me to take advantage of my powers and the things that I own in order to secure my survival. Therefore, I am allowed to run as fast as I can in *Bear*, but not to push you in *Flotsam.* Another commonsensical difference between these cases relates to the killing/letting die distinction. While your death in *Bear* is not the usual case of letting die (as by outrunning you I actively divert the bear to you), this is no killing either. I do not directly attack you. But, in freely competing on the flotsam, we directly attack each other. The feeling that a coin flip is needed emerges from the commonsense morality requirement to avoid a direct attack.

Why do these differences matter?[17] Probably, they somehow imply that, as an autonomous agent, I am free to outrun you but I am not free to push you from the raft. Victor Tadros' systematic analysis of "the means principle" is an attempt to elucidate this autonomist reasoning. For Tadros, the means principle is violated where one incorporates others into one's plans and projects in a harmful way, or where one manipulates others to serve one's own ends.[18] This principle asserts that using a person as a means to another's end is fundamentally wrong because it disrespectfully offends her standing as an independent

[17] For a related autonomist argument for the distinction, see Fiona Wollard, "If This Is My body…: A Defence of the Doctrine of Doing and Allowing," *Pacific Philosophical Quarterly* 94 (2013): 315–341.
[18] Tadros, *The Ends of Harm*, 140.

person.[19] The use of one's internal resources (viz., one's own body and talents) in pursuance of one's goals is permissible, unless this use interferes with another person's autonomy.

The means principle underlies some of the killing/letting die contrasts: harmful action usually interferes with the victim's autonomy, while harmful omission does not. If one harmfully attacks another person, or if one harmfully uses another person's vital resources, one fails to treat her as an autonomous person who is entitled to set for herself her own projects and plans. But, if one is not subject to the duty to help, letting another person die does not amount to disrespecting him.

In *Bear*, my running away is not disrespectful, despite being very harmful. This is because, first, the parties whose interests conflict with each other maintain no normatively significant special relationship (such as, say, parent to child.) Second, the stakes are high; I could have rescued you, but only at a great cost to myself. Finally, I am not under duty to bring about an equal distribution of chances to survive even if such an outcome is fairer and therefore impersonally better. In short, while by running as fast as I can I harm you, I do not use you or interfere with your autonomy in any way; thus, so the argument goes, I do not treat you disrespectfully. Or as Tadros puts it:

> Suppose that morality required me to toss a coin in *Bear*. That would require me, were I to lose, to sacrifice myself to save you. … But if that were so, morality would require one of us, the person who loses the toss, to make himself available as a means to save the other at the cost of his life. This would threaten the idea that each person is independent of every other in determining which projects and goals, amongst those that are valuable, to value and pursue.[20]

If I am required to agree to a fair procedure in *Bear* and sacrifice my life if I lose, then, according to Tadros, I am forced to be used as a means for realizing fairness. A morality that

[19] This idea is deeply related to the non-consequentialist moralities of T. M. Scanlon, Arthur Ripstein, and Jay Wallace. T. M. Scanlon, *What We Owe To Each Other* (Cambridge, Mass: Harvard University Press, 1998). Arthur Ripstein, "Authority and Coercion" *Philosophy and Public Affairs*, 32 (2004), 2-35 and *idem* "Beyond the Harm Principle" *Philosophy and Public Affairs*, 34 (2006), 215-245, R. Jay Wallace, "The Deontic Structure of Morality" available at https://philosophy.berkeley.edu/file/2/Deonticstructure-final.pdf (last visited, Nov. 16, 2014). These versions of non-consequentialism analyzes morality in terms of bipolar rights and duties. For the seminal analysis see, Ernest Weinrib, *The Idea of Private Law* (Cambridge, Mass: Harvard University Press, 1995), Michael Thompson, "What is it to Wrong Someone? A Puzzle about Justice" in J. Wallace, P. Pettit, S. Scheffler, and Smith (eds.), *Reason and Value*, 333-38 and Stephen Darwall, *The Second-Person Standpoint* (Cambridge, Mass: Harvard University Press, 2006).

[20] Tadros, *The Ends of Harm*, 210-11.

requires such a sacrifice is coercive and, as such, inconsistent with the means principle; it fails to recognize its addressee's autonomous personhood.

Thus, the autonomist framework, in symmetrical cases, is consciously insensitive to considerations of fairness. Suppose that I was paired with a particularly slower runner, or I was paired with a handicapped person who can barely limp along, much less run. In those cases the other runner was eaten by the bear due to brute bad luck; the difference between me and him is obviously unfair. Why should the brute luck fact of whoever happens to be faster be that upon which our lives hinge? No good answer is forthcoming, and still, according to the autonomist resolution of *Bear*, I am allowed to use my person and resources in order to advance important interests, even if this is unfair.

Still, Tadros's autonomist resolution of *Bear* and *Flotsam* faces a straightforward difficulty: he rightly observes that, by running as fast as I can, I do not use you as a means, but, I still use you in avoiding the bear. I run as fast as I can in order to divert the bear to you: had you not been around, my running away would be pointless. I adopt the running strategy precisely because I hope that you are slower than I am. This plan seems to amount to a violation of the means principle.

The description offered here does not contradict Tadros's observation that a requirement to self-sacrifice might also violate my autonomy. The idea that in exceptional cases autonomy rights can conflict with each other is familiar.[21] In order to complete the argument against the fair procedure resolution in *Bear*, Tadros should explain why this resolution is inferior to free competition.

Tadros might answer that free competition in *Bear* is non-violent and does not directly involve the other. But this answer is incomplete without explaining why direct involvement and the use of force matter. The fact that I exercise no violence in outrunning you seems irrelevant, from the perspective of the autonomist (means-principle-based) morality Tadros employs.

[21] Jeremy Waldron, "Rights in Conflict". *Ethics* 99 (1989): 503–519.

3. A preemptive rule-based morality of self-defense

3A. A Preemptive Rule Morality of Harm

The theory I advance in this section—Rule-SD—is a three-layered morality of defensive harm. It is composed of (*i*) a theory of the rights of fully informed defenders (and fully informed third parties); (*ii*) prospectivism: a decision theoretic theory of the rights possessed by fully rational but uncertain, incompletely informed, defenders; (*iii*) a Razian theory of "normally justified" rules that mediate (and therefore preempt rather than undermine) the reasons that apply to ordinary (i.e., less than fully rational) uninformed defenders.

Consider the first layer. A fully informed defender is justified in exercising defensive force if and only if in doing so he minimizes and fairly distributes the harm the attacker is about to inflict. Thus, if harming the attacker enforces fairness in the distribution of harm, the attacker is liable to defensive harm. I will follow Jeff McMahan in assuming that the attacker's degree of responsibility for a threat determines how the harm he is about to cause ought to be distributed. If the attacker is fully culpable for the threat, it would be fair if he would bear all of the necessary harm; if his responsibility for the threat is minimal, it would be fair if the harm were distributed more equally.[22] As an impartialist position, Rule-SD asserts that people who constitute a threat by, say, faultlessly falling on the defender—viz., they constitute a threat but not through their agency—are not liable to defensive harm; a fully informed defender has no right to kill the falling man. Probably, if one of the two must die, they ought to flip a coin in order to decide who will survive. Similarly, if in *Bear* and *Flotsam* the parties are fully informed they should decide who is to survive by a fair procedure.[23]

---

[22] These propositions are endorsed in Philip Montague, "Self-Defense and Choosing among Lives", *Philosophical Studies*, 40 (1981), 207-219 and in his "Self-Defense, Culpability, and Distributive Justice", *Law and Philosophy*, 29 (2010), 75-91; McMahan develops two different versions of this view in Jeff McMahan, "Self-Defense and the Problem of the Innocent Attacker", *Ethics* 104 (1994): 252-290 and in his "The Ethics of Killing in War", *Ethics* 114 (2004): 693-733 Jeff McMahan, "The Basis of Moral Liability to Defensive Killing," *Philosophical Issues* 15 (2005): 386–405, and other places. Compare, Michael Otsuka, "Killing the Innocent in Self-Defense," *Philosophy & Public Affairs* 23 (1994): 74–94. Suzanne Uniacke, *Permissible Killing: The Self-Defence Justification of Homicide* (Cambridge: Cambridge University Press, 1994).

[23] As Susanne Burri points out to me, my presentation leaves many questions open. First, the Otsuka argument against fair procedure in the innocent threat case: if I act to kill the falling man, I am responsible for my actions, whereas he is not responsible for falling towards me, hence there is an asymmetry that means the harm should befall me (see his "Killing the Innocent in Self-Defense"). Second, what if fairness and minimization conflict: I can either inflict significant harm on the falling

The second layer of the morality of harm addresses the problem of uncertainty, by overcoming the fact-relative morality/evidence-relative morality distinction.[24] Let me illustrate through *Falling Man with Defender's Uncertainty*:

> *Falling Man with Defender's Uncertainty*: A man *y* is faultlessly and involuntarily falling on a defender, *x*; unless *y* is blown to pieces, he will crush *x* to death. The defender *x* falsely but justifiably believes that *y* is culpable for the threat he is posing. That is, *x* justifiably believes that the falling man *y* is liable to defensive killing: she has evidence that indicates an 85% chance that falling man is liable.

Consider a decision theoretic presentation of this case. Defender has only two options, {*Self-Preference, ~ Self-Preference*}, whose possible values are presented as follows:

Innocent Threat	*pr*	Good Outcome	*pr*	Bad Outcome
Self-Preference	.85	1 = U(Defender survives thanks to killing a liable Attacker)	.15	-1 = U(Defender survives due to killing an innocent Attacker)
~ Self-Preference	.15	1 = U(Defender dies because she did not kill an innocent Attacker)	.85	-1 = U(Defender is killed because she did not kill a liable Attacker)

Since the Attacker is innocent,

> U(~*Self-Preference*) = U(Defender dies because she did not kill an innocent Attacker) = 1
>
> U(*Self-Preference*) = U(Defender survives due to killing an innocent Attacker) = -1.

As U(*Self-Preference*) < U(~*Self-Preference*), a fully informed Defender ought not to kill the falling man in self-defense.

---

man, or allow him to inflict a small harm on me? Does minimization take priority over fair distribution? These and other questions should not concern us here.

[24] For the distinction between fact-relative morality and evidence-relative morality, Derek Parfit, *On What Matters* (Oxford: OUP, 2011) chap.7 and the discussion of those distinctions in Tadros, *The Ends of Harm*, chap. 10. It is challenged in Frank Jackson, "Decision-Theoretic Consequentialism and the Nearest and Dearest Objection," *Ethics* 101 (1991): 461-482, and Michael J. Zimmerman, *Living with Uncertainty: The Moral Significance of Ignorance* (Cambridge: Cambridge University Press, 2008).

What about the uncertain Defender? Suppose that as far as Defender ought to know, the probability that Attacker is liable is .85; that is, the probability that the good effect of *Self-Preference* will come about is .85 (or, more formally, *pr*(Good/*Self-Preference*) = .85) and therefore *pr*(Bad/*Self-Preference*) = .15. If so,

U_e(*Self-Preference*) = *pr*(Good/*Self-Preference*)·U(Good) +*pr*(Bad/ *Self-Preference*) ·U(Bad) = .85·1+.15· (-1) = .7

U_e(~*Self-Preference*) = *pr*(Good/~*Self-Preference*)· U(Good) + *pr*(Bad/~*Self-Preference*)· U(Bad) = .15 - .85. = -.7

While the actual value of ~*Self-Preference* is greater than the actual value of *Self-Preference* (U(*Self-Preference*) < U(~*Self-Preference*)), the expected value of *Self-Preference* is greater than the expected value of ~*Self-Preference* (U_e(~*Self-Preference*) < U_e(*Self-Preference*)).

What should the uncertain Defender do? A Parfit style treatment asserts that defensively killing the falling man in *Falling Man with Defender's Uncertainty* is impermissible in the fact-relative sense because the falling man is not liable to defensive harm. Now, *x*, the defender, is incompletely informed, and *x*'s belief that the attacker is liable has been formed on the basis of adequately researched evidence. If such a research results in a belief that the attacker is most probably liable to harm, defensive killing is permissible in the "evidence-relative sense." For Parfit, the fact-relative justification and the evidence-relative justification are both valid and irreducible to each other.

In contrast, prospectivism (viz., the second layer of Rule-SD) offers a single concept of the rightness of action that follows from a general theory of uncertainty in ethics, developed by Frank Jackson and in more detail by Michael Zimmerman.[25] The fully rational yet uncertain deliberator ought to maximize *expected* value.[26] Thus, if U_e(~*Self-Preference*) <

[25] Jackson: "I hereby stipulate that what I mean from here on by 'ought,' …was the ought most immediately relevant to action, the ought which I urged it to be the primary business of an ethical theory to deliver. When we act we must perforce use what is available to us at the time, not what may be available to us in the future or what is available to someone else, and least of all not what is available to a God-like being who knows everything about what would, will, and did happen" ("Decision-Theoretic Consequentialism and the Nearest and Dearest Objection," p. 472).

[26] In "In Dubious Battle: Uncertainty and the Ethics of Killing," (unpublished ms.) Seth Lazar offers a "deontologized" construal of Frank Jackson's decision theoretic consequentialism. (Lazar does accept the fact-relative/evidence-relative distinction.)

U_e(*Self-Preference*), *x*'s self-preference is permissible, even if it involves the killing of a non-liable person.

To understand the difference between prospectivism and the Parfit style approach to uncertainty, consider a fully informed third party *z*, who knows that the falling man *y* is not liable to killing; *z* can prevent *x* from killing *y* by neutralizing *x*'s gun (rather than by directly harming *x*). In doing so, *z* allows *y* to fall on *x* and crush her to death. Under the Parfit style approach, *x*'s killing is permissible merely in the evidence-relative sense, so by neutralizing *x*'s gun *z* can prevent an action which is in fact impermissible. Thus, *z*'s intervention has two consequences: *y* would faultlessly (and therefore not impermissibly) crush *x* to death and *x* would fail to impermissibly kill innocent *y*. Since *z* can prevent *x*'s fact-relative wrongdoing by neutralizing *x's* gun (rather than by killing or wounding *x*), *z* ought to intervene.

In contrast, according to prospectivism, in the case just described, *z* has no reason to prevent *x*'s defensive action and thereby *z* has no reason to allow *y* to fall on her and crush her to death. After all, *x*'s defensive killing is permissible since it maximizes expected moral utility. Hence, her status has not been compromised due to this killing. True, had *x* been fully informed, killing *y* would have been a violation of her duty. Since *x* is not fully informed, her killing is permissible.

The distinction I offer above between the Parfit style and the prospectivist treatments of uncertainty does not rest on the false assumption that if it is permissible for *x* to *v*, then it is not permissible for *z* to interfere with *x*'s *v*-ing. In many cases where *z* knows something that *x* does not, it is permissible for *z* to prevent *x* from doing what would be *x*'s mistake had *x* been fully informed. The point here is different: *x* does not become *liable* to *z*'s harmful action in virtue of killing a non-liable man, if her killing is permissible (according to prospectivism). That is, she does not become liable merely in virtue of the fact that had she been fully informed, she should have acted differently. Unlike the Parfit approach, which renders *x's* action wrongful in the fact relative sense, prospectivism implies that *x*'s defensive killing is permissible in every sense, and infers from this fact that *x* cannot become liable to harm in virtue of permissibly killing non-liable *y*.

The second layer of Rule-SD entails the possibility of conflicting moral permissions. To see why consider:

> *Falling Man with Defender's and Attacker's Uncertainty*: Like Falling Man with Defender's Uncertainty, but in addition, the falling man, *y*, has a gun, and can kill *x* before *x* kills him; *y* justifiably but falsely believes that *x* knows that *y* is not liable: they both have evidence that indicates an 85% chance that the other party is liable.

Where both fully rational parties justifiably believe each other to be liable to defensive harm, both are justified in defending themselves against each other: despite being innocent, both lost the right not to be attacked by each other, but retain the right of self-defense.[27]

3B. The Third Layer: A Preemptive Self-Preference Rule

The third layer of Rule-SD utilizes Joseph Raz's distinction between first-order and second-order reasons. The first order reasons that apply to ordinary defenders in central self-defense cases might not be the only reasons they have. Ordinary defenders might be subject to a rule that functions as a second order reason, or, more particularly, as a preemptive reason. Saying that a person is subject to a rule that instructs/permits *v*-ing might mean that this person ought to/might *v*, instead of deliberating on whether *v*-ing is supported by the balance of the first-order reasons that apply to her in the circumstances.

A rule grounds a valid preemptive reason if it is "normally justified." A rule that instructs or allows its addressee to *v* is normally justified if following the rule—instead of deliberating on whether *v*-ing is supported by the balance of the first-order reasons—would enhance the conformity of the addressee to these reasons. Typically, rules are normally justified in circumstances in which the cost of comprehensive deliberation is high, the deliberative circumstances are poor, and the action supported by the rule would also be the most likely result of a suitably idealized comprehensive deliberation. [28]

---

[27] The second layer of Rule-SD cuts across the consequentialists/non-consequentialists division; both camps might agree that under circumstances of self-defense, fairness in the distribution of harm is of crucial importance. Both camps might agree that the uncertain defender should maximize the chances that fairness in the distribution of harm would be realized.

[28] A preemptive reason is a second-order reason to exclude some first-order reasons from deliberation. Such reasons neither override nor conflict with first-order reasons. Rather, they determine which considerations are to be excluded from the calculation of the balance of first-order reasons. The rule to which Rule-SD subjects ordinary defenders is in fact, a 'protected reason': a protected reason to φ at *t*

The third layer of Rule-SD asserts that the permissions involved in some central self-defense cases emerge from the rule to which the defender is subject, rather than from the first-order moral reasons that apply to her according to propectivism (the second layer). This core claim can be illustrated by *Falling Man with Defender's Uncertainty.*

According to prospectivism (of the second layer), if *x* is uncertain but fully rational, *x* has the right to act in self-defense if *x* justifiably but falsely believes that *y* is liable. In such a case U(*Self-Preference*) < U(~*Self-Preference*)—because Defender *x* kills an innocent threatener *y*; nevertheless, U_e(~*Self-Preference*) < U_e(*Self-Preference*), the killing is permissible.

The third layer takes this idea a step further. Suppose that as a result of *x*'s limitations, *x* ought to follow a preemptive rule that permits self-preference, instead of carrying out the procedure recommended by prospectivism. If so, *x* has the right to act in self-defense, even where U_e(*Self-Preference*) < U_e(~*Self-Preference*). Thus, a fully informed third party *z* has no reason to allow *y* to crush *x* to death by forcing inaction on *x*, even if, had *x* been fully rational, *x* could have known that *y* is not liable.

The third layer is based on an empirical speculation about ordinary defenders. This speculation is composed, in turn, of three empirical generalizations regarding cases in which the defender is threatened by one attacker or by a wild animal ("the designated circumstances"). Typically, under the designated circumstances, (1) the attacker/threatener (if there is one) is at least minimally culpable for the threat he poses and therefore liable to

---

is the combination of (i) a first-order reason to φ at *t* , and (ii) an preemptive reason to disregard some of the first-order reasons that bear on the choice at *t*. See, Joseph Raz, *The Morality of Freedom* (Oxford University Press, 1986), 39ff. For an illuminating discussion of Raz's views see Scott Shapiro, "Authority" in Scott Shapiro and Julius Coleman (eds.) *The Oxford Handbook of Jurisprudence and Philosophy of Law* (Oxford University Press, 2004)).

Raz's conception of the normal justification of rules has been criticized from various angles. See a summery in Shapiro Section IV (which includes a clear presentation of Michael S. Moore, "Authority, Law and Razian Reasons," *Southern California Law Review* 62 (1989): 866-867). Cf. Chaim Gans, "Mandatory Rules and Exclusionary Reasons," *Philosophia* 15 (1986): 373-396, and William A. Edmundson, "Rethinking Exclusionary Reasons" *Law and Philosophy* 12 (1993): 329-43.

The Razian analysis of preemptive rules employed here might be replaced without losing much. For interesting alternatives, see, e.g., Stephen Perry, "Second Order Reasons, Uncertainty and Legal Theory," *Southern California Law Review* 62 (1989): 932-957 and Fred Schauer, *Playing by the Rules: An Examination of Rule-Based Decision-Making in Law and in Life* (Clarendon Press, 1991): 88-93. I do not try to defend the Razian view of authority here, nor do I try to defend the modifications of his views, offered by Perry and Schauer.

defensive harm; (2) the deliberative circumstances are poor, so that the ordinary defender is under stress (i.e., not fully rational). Her deliberation would significantly deviate from ideal deliberation. In fact, ordinary people have a terrible track record in calculating likely consequences and correctly inferring the decision theoretic morality that governs stressful circumstances. In following a well-designed rule instead of directly maximizing expected moral value they are more likely to imitate ideal deliberators. And, (3) the defensive actions of a defender under stress are more effective if she exhibits strong egocentrism, viz., disregards any agent neutral reasons that apply to her.

If the first generalization is true, most one vs. one cases are such that the fully informed defender has a right to harm the attacker in self-defense. It follows that according to prospectivism fully rational but uncertain defenders have a right to defensively harm their attackers, due to the high probability that the attackers are liable to defensive harm. Indeed, under the empirical generalization adopted here, rational but not fully informed defenders may permissibly defend themselves in some falling man cases since mostly, their false belief that the falling man is liable to harm will be supported by the defender's evidence.

Enter the second and third generalizations: ordinary deliberation on the merits of one vs. one self-defense cases is usually both very different from ideal deliberation and costly. Ordinary defenders are under stress and, therefore, their beliefs as to the liability of the attacker are biased. Furthermore, allowing defenders to exhibit egocentric behavior rather than respond to the agent-neutral reasons that apply to them is more likely to lead them to do (what Rule-SD's second layer defines to be) the right thing. A defender who invests all of his available resources in defending her interests, but invests nothing in making sure that defending her interests is the correct thing to do (i.e., supported by the balance of agent neutral reasons), is actually more likely to do the right thing.

Rule-SD infers from these generalizations that the defender is subject to the "self-preference rule":

> The rule of restricted self-preference in self-defense: A defender can promote her own interests by harming the attacker in one vs. one (or less) self-defense cases,

> provided that (a) the defender is morally innocent vis-à-vis the threat imposed on her; (b) has no immediately available or obvious evidence that indicates that the attacker is not liable to defensive harm; (c) (to be completed.)[29]

Arguably, if the generalizations listed above are all true, *and* ordinary defenders can easily identify one vs. one self-defense cases and follow the self-preference rule, the rule is normally justified. It mediates the first-order reasons that, according to prospectivism of the second layer, apply to ordinary defenders in these circumstances. As such, the self-preference rule grounds a preemptive reason: allowing the defender to follow it and exempting her from attending to first order agent-neutral reasons would enable the defender to promote her interests more effectively, thus be more successful at doing the right thing to do. This permission to exhibit strongly egocentric behavior is likely to enhance the defender's conformity to the agent-neutral reasons that apply to her in the designated circumstances. Therefore, the self-preference rule grounds a permission to exhibit egocentric behavior (viz., to disregard agent-neural reasons) without abandoning the conviction that morality is intrinsically related to impartiality. Thus, although egocentric reasons are of no moral significance, the agent is permitted to exhibit egocentrism because it is likely to bring about the best outcome impartially considered.

Prospectivism and the preemptive rule morality of the third layer is neutral between consequentialism and non-consequentialism. Whatever the reasons that apply to the defender in the designated circumstances—they might follow from facts about consequences (as per the consequentialist formulation) or from facts about autonomy and respect (as per non-consequentialism)—subjection to the self-preference rule usually enhances the defender's conformity with these reasons.[30]

---

[29] The third part of the rule concerns harm inflicted on bystanders. I will develop it in the next section.

[30] Having said that, I tend to read Rule-SD as a consequentialist doctrine sensitive to both aggregate welfare and fair distribution. Despite appearances to the contrary, Rule-SD is not an instance of the common version of rule consequentialism ("An act is wrong if it is forbidden by the code of rules whose internalization by the overwhelming majority of everyone everywhere … has maximum expected value ...." (Brad Hooker, *Ideal code, real world a rule-consequentialist theory of morality*. (Clarendon Press and Oxford University Press, 2000), p. 32). According to Rule-SD, the following the rule is permissible because subjection to this rule is likely to enhance compliance of the *individual* with the demands of act consequentialism, in each circumstance she faces. The rule's belonging to the "ideal code" has no role in determining the rightness of her action, according to Rule-SD.

3C. Rule-SD vs. Partialism, or Four Payoffs of Rule-SD

Four interesting implications of Rule-SD are immediately apparent. First, like partialism, it allows egocentric behavior in *some* innocent threat cases. In one vs. one circumstances Defender is entitled to disregard (some of the) agent neutral reasons that apply to her. Prospectivism further implies that even a non-violent intervention of a third party, whose goal is preventing defensive killing, has no moral justification, despite the fact that the falling man is not liable to this defensive killing.

Second, like partialism, the self-preference rule permits defensive harm in some of the rare cases in which a defender under stress could have easily known that the attacker is totally innocent and therefore not liable to defensive harm. Suppose an ordinary defender does not know that the attacker is innocent, and, that she could easily gain knowledge about the liability of the attacker. This defender *might be* subject to the self-preference rule as well in case she does not know that she can gain the knowledge about the nature of the threat. Then, she might follow the self-preference rule instead of deliberating on the merits of the case. Being subject to this rule, she is exempted from the duty to assess the probability that her attacker is liable to defensive harm. Thus, under prospectivism of the second layer, killing an attacker whose innocence can easily be known is indistinguishable from killing an innocent bystander. Yet, Rule-SD insists that given the normal justification of the self-

---

Now, Rule-SD might be read as an instance of indirect consequentialism. *Direct* consequentialism is "usually construed as holding that an act's moral permissibility depends on a comparison of that act's consequences with the consequences of alternative acts open to the agent. *Indirect* consequentialism judges an act permissible if and only if it accords with motivation, dispositions, rules and kind of conscience that are favoured by the consequentialist assessment."… (Brad Hooker, "Impartiality, Predictability and Indirect Consequentialism," in Roger Crisp and Brad Hooker (eds.), *Well-Being and Morality: Essays in Honour of James Griffin* (Oxford University Press, 2000), p. 130.) I prefer a different reading of Rule-SD. It asserts that the self-preference rule to which Rule SD subjects ordinary defenders functions like future directed decisions: "Future-directed decisions are … tools for the non-manipulative, intrapersonal division of deliberative labor over time. A future-directed decision to ϕ gives rise to a defeasible exclusionary reason to ϕ. This reason is grounded on the default authority that is normally granted to one's prior self as an 'expert' deliberator" (Luca Ferrero, "Decisions, Diachronic Autonomy & the Division of Deliberative Labor," *Philosopher's Imprint* 10 (2010): 1-23, at p. 7). Subjection to the self-preference rule is a way to overcome a predictable limitations at the time of action. In light of those limitations, the agent should decide to follow a normally justified rule, because this decision maximizes the expected value of the future choice at the time of the decision. For the limited, biased defender, the way to maximize expected value is to subject herself to a rule, *before* limitations and biases get hold on her. Accordingly, deliberators ought to follow a rule, only if they justifiably believe that deliberation on the merits of the case at the time of action won't bring about better results.

preference rule to which ordinary defenders are subject, defensive harming against a person who could have been known to be an innocent attacker can at least in principle be permissible.

Thirdly, unlike partialism, Rule-SD does not permit a defender to kill attackers or threateners who are already known to her to be innocent. Further, the rule does not exempt a defender who knows that deliberation on the merits of the case is easily doable from conducting it. Rather, it merely exempts a stressed defender that lack these pieces of knowledge from conducting an investigation into the attacker's liability. Thus, Rule-SD yields an asymmetry between fully rational but uninformed defenders and uninformed *ordinary* defenders.

Fourth, Rule-SD supports the intuition that egocentric reasons can only function as tie-breakers in one vs. one self-defense cases. This is because the empirical generalizations that justify a rule permitting egocentric behavior in these cases are untrue of more complicated cases. Suppose the costs the defender would bear should she avoid killing the attacker are sufficiently low. In such a case, the benefit she secures by defensively killing an attacker is less likely to justify self-preference. As a consequentialist would put it, even if the attacker is fully culpable for the threat he poses, an outcome whereby the attacker is killed and the defender's minor interest is protected is worse than an outcome in which the attacker stays alive and the defender suffers minor harm. The non-consequentialist articulation of Rule-SD might analyze this situation in different terms, but would reach the same conclusion.

Suppose the aggregate costs the defender needs to inflict in order to survive are higher than they are in one vs. one cases. Consider, for example, a modification of *Innocent Threat* where the defender is threatened by a large number of trapped elevator passengers. As opposed to the culpability/innocence of attackers, in this self-defense case the number of falling persons is a visible feature of the case, such that even defenders under stress can grasp it. Hence, treating a self-preference rule as a preemptive reason is less likely to lead the defender to make the right choice, impartially considered.

Self-preference is likely to be impartially justified in one vs. one cases also because, usually, an ordinary defender is certain that she is innocent vis-à-vis the threat imposed on her, but has no easy way to determine the degree to which the attacker is responsible for the threat he is posing. This asymmetry yields only minor impact on the first-order moral reasons in one vs. many self-defense cases. Put, again, in terms of consequences, an outcome in which the innocent defender is killed and ten minimally culpable attackers survive might be preferable to an outcome in which the ten are killed and the single defender survives. Before following the self-preference rule, defenders should try to make sure that they are indeed facing a one vs. one self-defense case governed by this rule.

4. Innocent Threats/Innocent Bystanders, Bear/Flotsam and the Preemptive Rules Morality of Harm

Free competition in *Flotsam* involves use of force while free competition in *Bear* does not. This difference seems insignificant from standard consequentialist perspectives. Moreover, this difference seems insignificant from deontological perspectives that centralize the means principle. This section argues that properly enriched, Rule-SD explains the moral significance of violence in these cases, and thus successfully draws a plausible distinction between versions of *Bear* and *Flotsam*.

Consider two distinctions, and the empirical generalizations associated with them. The first is the attackers/bystanders distinction: unlike attackers, bystanders are not likely to be liable to defensive harm. The second is between violent and non-violent harms: other things being equal, violence is more likely to cause harm than other types of action. In light of these generalizations, the rule of self-preference offered above would be normally justified only if it sharply restricts the permission to self-preference. The rule allows self-preference, and exempts the ordinary defender from attending to agent-neutral reasons, only if the harm to bystanders is unlikely to be disproportionate.[31] Typically, a harm is likely to be disproportionate if it results from violent and direct involvement.

[31] Put in terms coined by Jeff McMahan, in the designated circumstances, reasons grounded in *narrow proportionality* are preempted by the self-preference rule. (Narrow proportionality determines the size of harm to which the attacker/threatener is liable in proportion to the degree to which he is responsible

> The rule of self-preference in self-defense: a defender can promote her own interests in one vs. one (or less) self-defense cases … as far as (c) she does not foreseeably or intentionally harm bystanders by using force against them, cheating them, stealing from them, or any other illicit (but well defined) means.

The restriction is designed to impede the undesirable foreseeable effects of egocentric behavior in designated circumstances, without forcing the defender to overcome his partiality.

These restrictions are *not* prohibitions. The self-preference rule does *not prohibit* killing bystanders, cheating or stealing from them. Rather, it states that *if*, in order to survive the defender must violently harm bystanders, she is not exempted from deliberating on the merits of the case. Violently harming bystanders and, especially, stealing from them might be justified by lesser evil considerations; the permission to do or allow these harms can never be grounded in the self-preference rule. Thus, Rule-SD is consistent with the view which asserts that if intentionally inflicting serious harm on bystanders is necessary for the defender's survival, the defender is permitted to inflict it, as far as she makes sure that the inflicted harm is a lesser evil.[32]

Interestingly, permissible egocentrism in the free market is restricted in the same way. Every individual is allowed to choose the most advantageous employment for whatever capital she can command. According to the free market myth, this rule enhances conformity to agent-neutral reasons: the egocentric agent is led to prefer the employment that is most advantageous to society. And yet, permissible egocentrism is constrained: as its undesirable effects are immediate and certain, robbery is prohibited even if it provides advantageous employment to some people. The same is true of enslavement. The regime restricts egocentric agents operating within the free market by ruling out clearly undesirable activities, without imposing on them a duty to be impartial.

---

for the threat he poses). For this distinction between narrow proportionality (that usually concern harms inflicted intentionally) and wide proportionality that usually concern harms inflicted foreseeably but unintentionally on bystanders, see Jeff McMahan, *Killing in War* (Oxford: Clarendon, 2009), sec. 1.3.

[32]As Ryne Jenkins points out, I am permitted to steal from innocent bystanders to defend myself. Suppose you and I are running away from the bear, and I tip over a picnic basket that I find in someone else's campsite. I thereby destroy someone else's property. But by doing so, I save both your life and mine; as the bear focuses its attention on the food from the basket. Surely this is permissible. In fact, it seems obviously preferable to the two of us running until one of us is exhausted and then mauled.

Rule-SD offers a very simple resolutions of various *Bear* like cases. They are all a one vs. less than one self-defense case, in which both parties face a wild animal threat. Initially, the parties, you and I, do not threaten each other: I am a bystander vis-à-vis the threat imposed on you and, *mutatis mutandis*, the same is true of you. Supposedly, if it is part of a plan to divert the threat to another innocent bystander, running away in *Bear* is impermissible. Under a means-principle based morality, no first-order difference exists between diverting the threat posed by the bear by running as fast as one can and diverting it by using force against the other party.

But consider a *Bear*-like case—*Bear with Uncertainty*—in which we are both subject to the self-preference rule because we are unaware of the idiosyncratic features of the case. All we initially know is that the bear threatens us. What follows from our subjection to the self-preference rule is that we are allowed to exhibit strong egocentrism as long as we respect the restrictions specified by the third part of the rule. That is, we are allowed to run as fast as we can without taking each other's interests into account, insofar as we do not use force against bystanders, cheat them, steal from them, or act in any other illicit way towards them.

True, it turns out that in running away from the bear, I violate the means principle; I divert the bear to you, and (we assume that) diverting the threat by running away is just as wrong as using force against bystanders. Yet, *Bear with Uncertainty* is one of the unusual cases governed by the self-preference rule. When facing a threat by wild animals, the right choice of the uncertain defender is most probably to run away; impartial deliberation on the merits of the case will likely amount to a waste of time (the most precious resource in such situations). Being subject to the self-preference rule, I have a preemptive reason not to respond to agent neutral reasons, so I do not have to make sure that I do not use you as a means. *Mutatis mutandis*, the same is true of you: you are allowed to run as fast as you can and in fact to prevent me from doing what I am at liberty to do. Unbeknownst to us, these permissions conflict with each other: I am allowed to divert the bear to you, and you are allowed to prevent me from doing so. Thus, Rule SD implies that I am permitted to run from the bear even *only* if it is not obvious to me that I use you in avoiding the bear.

Turn to *Flotsam*. In the relevant sense, we are both bystanders vis-à-vis the threat imposed on us; we are threatened by drowning in the deep waters rather than by each other. Perhaps, in another sense, we do threaten each other by virtue of our presence on the flotsam, which cannot carry us both. Still, the self-preference rule would treat us as bystanders because there is no presumption that a person who poses a threat by being on the *Flotsam* is minimally culpable for this threat, and therefore there is no presumption that he is liable to defensive harm. Being subject to the self-preference rule, we are allowed to exhibit strong egocentrism, as long as it does not involve using force against bystanders.

Yet, in this case, mutual egocentric behavior entails violence, which the strong partiality rule does not allow.[33] If the parties in *Flotsam* are subject to no other preemptive rule, they should resolve their conflict by deliberating on the merits of the case they are facing. Considered on its merits, according to any view that relates morality to impartiality, fair procedure seems to be the only acceptable resolution of this case.

The resolutions of *Bear with Uncertainty* and *Flotsam* offered by the preemptive rule morality of harm are consistent with the claim that, considered on their merits, the cases are indistinguishable. All Rule-SD does (and can do) is provide a second-order distinction between these cases, which implies that the parties might be subject to the self-preference rule in *Bear* like cases that involve uncertainty, but never in *Flotsam*-like cases.

5. Conclusion

Rule-SD asserts that ordinary defenders facing one vs. one (or less) self-defense circumstances are subject to a preemptive rule allowing them to defensively harm attackers and exempting them from conducting research into whether the attacker is liable to defensive killing. Rule-SD explains why uncertain, ordinary deliberators might have conflicting moral permissions in innocent threat cases and symmetrical cases like *Bear* (where *x* is permitted to

[33] As Susanne Burri noted in personal communication, it might be thought that the self-preference rule does not prohibit the parties to hang on to the flotsam, see who drowns faster, and if the other person drowns faster, I get to live, because the flotsam will then support me. However, this might seem less like free competition and more like fair procedure to determine who should live, especially because of the restraint that the parties exhibit, and their tacit agreement not to use force.

kill *y* in self-defense, while *y* is permitted to prevent *x* from doing so). Moreover, the way the rule distinguishes between bystanders and attackers and between violent and non-violent harms enables Rule-SD to draw a plausible moral distinction between different symmetrical cases like *Bear* and *Flotsam*.