Draft: Do not quote or cite. Read charitably.

Provocateurs

Kimberly Kessler Ferzan*

Iago taunts Othello with false claims that Desdemona is unfaithful. As we know, Othello kills Desdemona. But if Iago had the ability to stop Othello, it seems clear that he should harm Othello or harm himself to save Desdemona. Although his actions provoked Othello's conduct, Iago now not only may, but must, act so as to stop the harm to an innocent.[1]

Things, however, are far more complicated when the provocateur provokes harm to himself. Consider one judge's hypothetical:

> Imagine a funeral ceremony with hundreds of mourners for a widely respected African-American civil rights leader. A white supremacist appears at the church and begins shouting nonthreatening, racial epithets. Enraged mourners rush the person, who pulls out a concealed gun and kills several of them.[2]

Unlike Iago, who must act, here, our intuitions seem to support the opposite. The white supremacist is not permitted to kill the enraged mourners.

And then, consider the movie *Death Wish* in which Charles Bronson sets himself up to appear as a vulnerable victim so that others will try to harm him and he may act in self-defense.

Across jurisdictions, the white supremacist's and Bronson's claims of self-defense will likely fail. For example, according to the Model Penal Code: "The use of deadly force is not justifiable if … the actor, with the purpose of causing death or serious bodily harm, provoked the use of force against himself in the same encounter…"[3] Indeed, most jurisdictions have broad forfeiture doctrines. They do not require a purpose to cause death or serious bodily harm, but just bodily injury, nor do they limit the provocation to the same encounter.[4] Generally, when one intentionally provokes another (a person whom I will call

* Professor of Law, Rutgers University, School of Law – Camden; Associate Graduate Faculty, Rutgers University, New Brunswick, Philosophy. I thank the participants at the NYU Criminal Law Theory Colloquium (Aaron Simowitz, Lauryn Gouldin, Chad Flanders, Marc DeGirolami, Dan Markel, Mike Cahill, and Tony O'Rourke) for their perceptive comments and feedback on the draft mansucript, and Larry Alexander, Doug Husak, Brian Little, and Joe Snee for helpful discussions of the topic.

[1] *See infra* section I.A.

[2] Sate v. Riley, 976 P.2d 624, 631 (Wash. 1999)(en banc)(Talmadge, J. concurring).

[3] Model Penal Code § 3.04(2)(b).

[4] *See, e.g.,* Ala.Code 1975 § 13A-3-23; Ga. Code Ann., § 16-3-21; State v. Richardson, 670 N.W.2d 267, 278 (Minn. 2003)(provoking "the difficulty"); Mt. St. 45-3-105; N.H. Rev. Stat. § 627:4; U.C.A. 1953 § 76-2-402.

"the respondent"), he is barred from using deadly force to defend himself from the attack that he provoked.

The provocateur is just one example of a more general question as to what to do with *actio libera in causa* cases. *Actio libera in causa* is the name given by German theorists to those cases where the defendant causes the conditions of his own defense.[5] For example, if Albert wants to kill Betty but is afraid he will not get up the courage to do so, he may take a hallucinogen at t_1 that will render him violent at t_2 when he knows he will be alone with her. It seems clear that Albert has purposefully killed Betty despite the fact that his mental state is temporally disconnected from his actus reus. If Carl recklessly starts a fire at t_1 that requires a firebreak at t_2, it seems clear that he should purposefully destroy the property at t_2 to create the firebreak, but he bears some responsibility for so doing.[6]

Criminal statutes that deny the defense in *actio libera in causa* cases do so in problematic and haphazard ways. Consider the following problems Paul Robinson noted with current formulations. *Proportionality:* Some statutes provide that an individual who provokes a fist fight loses the right to employ deadly force when his insult is met with gunfire.[7] *Strict Liability:* The term "provoke" could be interpreted in some jurisdictions to depend only on the effect on the respondent, such that, to use Robinson's example, one can provoke one's neighbor (and thereby lose one's defensive rights) by painting one's house a color that incites one's neightbor.[8] *Culpability Mismatches:* Some jurisdictions do not take into account the differences in the culpability in causing and the culpability of the later act.[9] We may doubt that a person who negligently starts a fire should be held responsible for purposeful destruction of property when he creates a firebreak.

Fixing statutes, however, requires that we understand the substantive principles at work, and theorists have offered us different accounts of the underlying rationale for the treatment of these cases.[10] It has often been assumed that we can solve all *actio libera in causa* cases the same way. Indeed, Paul Robinson extols the benefit of his approach in offering one unified approach across defenses.[11] The reasoning seems to be that what is good for necessity or intoxication is good for provocateurs. The purpose of this paper is to show that that assumption is wrong.

---

[5] The label comes from the Germans. *See* Claire Finkelstein & Leo Katz, *Contrived Defenses and Deterrent Threats: Two Facets of One Problem*, 5 Ohio State J. Crim. L. 479-504, 480 n.2 (2008). It originally applied only to cases in which the defendant rendered himself irresponsible, but has since been generalized to other defenses. *Id.* at 482.

[6] *Actio Libera in Causa* technically includes only those who purposefully contrive the conditions. *Actio Illicita in Causa* is the term used when individuals are only aware, but do not intend, to create the conditions. *See id.* n.3. For our purposes, nothing turns on this.

[7] Paul H. Robinson, *Causing the Conditions of One's Own Defense: A Study in the Limits of Theory in Criminal Law Doctrine*, 71 Va. L. Rev. 1-63, 13 (1985). Despite the fact that Robinson wrote this paper over three decades ago, little has changed with respect to the problems of statutory formulation.

[8] *Id.* at 5-6.

[9] *Id.* at 18.

[10] *E.g.,* Joachim Herrmann, *Causing the Conditions of One's Own Defense: The Multifaceted Approach of the German Law*, 1986 B.Y.U. L. Rev. 747-767; Robinson, *supra* note ___; Alexander (this issue); DeGirolami; Finkelstein and Katz.

[11] Robinson, *supra* note ___. *But see* Herrmann, *supra* note ___ (noting the complexities in German law that Robinson's unified approach runs roughshod over).

This paper will focus on provocateurs because they present two particularly interesting problems. The first is to explain how it is that provocateurs can lose their defensive rights without grounding the respondent's right to react. That is, when the white supremacist provokes the enraged mourners, they still do wrong by attacking him. This is to be contrasted with self-defense where the initial aggressor's attack simultaneously forfeits the aggressor's rights and grounds the permissibility of the defender's response.[12] My aim therefore is to understand what is going on with respect to the moral terrain – how and why do provocateurs lose these rights? Although my argument is largely consistent with existing law, my goal is not to rationalize it but merely to use existing legal standards as some evidence of our reflective judgments about these cases. The white supremacist seems to be a compelling case for why some provocateurs cannot fight back. The goal is to understand why.

The second puzzle takes us into the depths of *Death Wish* and cases like it. It is not altogether clear how to determine whether the provocateur's action is culpable, or blameworthy, or impermissible at t_1. One may know he is inciting a bully by leaving one's home. One may know she is encouraging rapists by dressing scantily. But we are free to leave our homes and dress immodestly. Knowledge that one might have to fend off the bully or the rapist does not render this conduct culpable or impermissible. And even with the person who purposefully engages in conduct so as to kill his attacker, why is the question, "Is it permissible to provoke this person so he will attack me so I can kill him?" rather than "Is it permissible to provoke a person so he will attack me so that I may *justifiably* kill him?" That is, if all we have at t_2 is a dead respondent who impermissibly tried to kill the provocateur, then what has the provocateur done wrong?[13]

Ultimately, I will claim that the actions of both provocateurs and their cousins, initial aggressors, alter the moral landscape at t_1. We have normative powers by which we can change rights and duties, for example we can alter our property rights with gifts, permissions, and abandonment. Provocateurs also change the moral relationship with the respondent. By consciously creating the unjustifiable risk of inciting the respondent to attack him, the provocateur forfeits his right to defend against such an attack at t_2. Necessity and intoxication do not change the rights and duties between two parties. Aggressors and provocateurs do.

As to the second question, I will argue that the *Death Wish* cases present a problem akin to entrapment. With respect to both provocation that is intended to allow the provocateur to use force in response and police behavior thought to constitute entrapment, the problem is not that the provocateur or police behave in a way that exculpates the respondent/defendant. Rather, the problem is that there is something seemingly problematic about significant alterations of another's circumstantial luck. We worry that there is something unfair about being forced to confront challenges that would otherwise not occur, challenges that bring out the worst in us. I will not answer the boundaries for police conduct. With respect to provocateurs, however, I argue that their efforts to play both God and state render

[12] For ease of exposition, I am bypassing the nuances in self-defense. My only aim here is to draw the contrast.
[13] Finkelstein and Katz pose this question in the context of comparing *Actio Libera in Causa* to deterrent threats. If one endorses what they dub the *Backward Induction View* that it is impermissible to threaten what it is impermissible to do, then by the same reasoning, the permissibility of self-defense at t_2 would justify the provocateur's action at t_1. *See* Finkelstein and Katz, *supra* note __, at 500-501.

them moral and political vigilantes that destroy rule of law values. With respect to private citizens, any act meant to alter another's circumstantial luck for the purpose of sitting in judgment upon it is impermissible. Accordingly, at least when there is a reasonably well functioning state, they are not permitted to provoke impermissible conduct so that they may act as state.

This paper proceeds as follows. Part I will survey three approaches to *actio libera in causa* in the literature, those of Larry Alexander, Paul Robinson, and Marc DeGirolami, all three of which presuppose a single solution to the problem, and I will argue that all three answers are ill-suited to solve the provocateur problem. Part II argues that the reason we cannot use a "one size fits all" approach is because self-defense is not best understood as simply part of a lesser-evils analysis. Rather, what initial aggressors do is to become liable to the force used against them. Part II argues that we need something akin to the liability principle to explain why provocateurs cannot fight back. Part III looks for this forfeiture concept in the doctrines surrounding the provocation mitigation defense, but ultimately concludes that an understanding of provocateurs and *actio libera in causa* cannot be derived from an understanding of when respondents are entitled to mitigation, even under a partial justification view of provocation. Part IV sets out the positive claim that the reason why provocateurs lose their defensive rights is that they cannot complain and defend against a risk that they themselves impermissibly created. Part IV also argues that this forfeiture doctrine requires subjective appreciation of this risk, and that the creation of the risk must itself be unjustifiable. Part V explores within the context of the unjustifiability criterion, the question of whether the later potential justifiability of the defensive conduct can render the risk justifiable at t_1. Part V concludes that what is going on is that provocateurs, in a way akin to entrapment, impermissibly alter another's circumstantial luck in order to induce the crime. Because this is something which they lack standing to do, they threaten rule of law values in a way that renders their conduct at t_1 impermissible.

I. Why Other Approaches to Actio Libera in Causa Don't Solve the Provocateur Problem

A. Alexander's Solution to the Non-Problem of Actio Libera in Causa

In accord with Larry Alexander's contribution to this symposium, I think we can and should, for the most part, analyze *actio libera in causa* cases in two stages.[14] The question is when does the actor consciously disregard a substantial and unjustifiable risk. If at t_1 Ben sets a fire to kill Joe, but has also risked harm to countless others, then at t_1 he is responsible for this culpable action.[15] In addition, by unleashing this risk, Ben is now under a duty to prevent the harm from occurring when possible, and so, he has additional culpability for the duration of his failing to rescue. Indeed, Alexander and I have previously

[14] Larry Alexander, *Causing the Conditions of One's Defense: A Theoretical Non-Problem* (this issue).

[15] As Alexander and I argue in our book, results do not matter so the offense is complete at the moment the risk is unleashed. Larry Alexander and Kimberly Kessler Ferzan, with Stephen J. Morse, Crime and Culpability: A Theory of Criminal Law ch. 5 (2009).

bitten the bullet that if Ben shoots Joe at t_1, but can save Joe at t_2, but does not, then Ben is guilty of two crimes – one for the risking and one for the omitting.[16]

This approach resolves the question of what to do when Iago provokes Othello to kill Desdemona. Iago increases the risk of harm to Desdemona (and for those who care about results, bears responsibility for her death) because he does two things. First, he influences Othello's reasons. He gives Othello a reason to act that did not exist before. Second, he influences Othello's rationality. He appeals to Othello's anger and jealousy.[17] Iago unleashes the risk at t_1, and at t_2, he has a duty to rescue Desdemona if possible. (There is a rather large puzzle embedded in the deceptively simple claim to which I cannot do justice here, but will also allude to later. This is the question of the degree of responsibility we bear for our increasing the risk that others will do wrong.[18])

---

[16] *Id.* at 242.

[17] Although because provocateurs "provoke," we might think they are conceptually limited to those who do cause anger, it is important to note that either reason-giving or rationality-influencing behavior may be individually sufficient for blameworthiness. In the first case, someone who solicits a hit man, or joins a conspiracy, increases the risk of harm to a victim by giving others reasons to harm her without impairing those others' rationality. In the second case, one might, for example, involuntarily intoxicate someone who is about to drive home. Such conduct does not give the driver new reasons for action, but still increases the harm to others and is thus blameworthy. Doctrinally, the criminal law, because of its view that voluntary human actors cut causal chains, offers different accounts for these two sorts of behaviors. The first are dealt with through complicity, conspiracy, and other doctrines that do not assume the actor causes the result. The latter are dealt with by assuming that when one acts through an innocent or irrational person, there is no voluntary human action cutting the causal chain. This clear doctrinal split is exactly what gets complicity into trouble when it attempts to deal with Iago, as he both contributes to reasons and rationality. *Cf.* Glanville Williams, Criminal Law: The General Part 391 (2d ed. 1961)(offering a theoretical account that allows Iago to receive greater punishment than Othello, *contra* the criminal law's general requirement that accessorial liability is derivative of the principal's).

Of course, if we abandon the idea that humans somehow have contra-causal freedom that cuts causal chains, we can avoid such doctrinal anomalies. For those of us who think that results do not matter, it is as simple as asking whether someone is culpable at t_1 for increasing the risk of harm to others by giving reasons *or* intoxicants. And, even for those who think that results matter, a cleaner analysis is available. For instance, Michael Moore thinks that we do not need a doctrine of accomplice liability, as causing someone to have a reason to act is still causing. Michael S. Moore, Causation and Responsibility: An Essay in Law, Morals, and Metaphysics 299- 301 (Oxford and New York: Oxford University Press 2009).

[18] Alexander and I reject that purpose is a requirement for complicity (*see* Alexander and Ferzan, *supra* note __, ch. 2), but even Michael Moore, given his views accomplice liability as superfluous, has problems limiting the responsibility of agents who give reasons to others. That is, if voluntary actions by others do not break causal chains and thereby limit our responsibility, then our actions seem to have potentially wide-ranging and problematic effects. A.P. Simester and Andreas von Hirsch, Crimes, Harms, and Wrongs 58, ch. 5 (2011); *see also* Heidi M. Hurd, *Is It Wrong to Do Right When Others Do Wrong? A Critique of American Tort Law*, 7 Legal Theory 307-340 (2001). If Alex reads *The Girl with the Dragon Tattoo* and then decides to engage in horrendously sadistic acts against women, why isn't Stieg Larson culpable? (Even if we think the First Amendment does some of the work, it can't do it all. After all, you aren't permitted to cause folks to trample over others by yelling "Fire" in a theater.) Lord Jauncey suggested in the *Brown* case that one reason to prohibit the consensual sadomasochistic behavior of adult men was that it might encourage similar conduct against children. Brown, [1194] 1 AC 212, 246.

To answer this question, the Moore/Duff debate is instructive. In urging contra Moore that accomplice liability is not superfluous, Duff argues that the distinction between purpose and knowledge is not a difference in degree but

This position then entails that if Iago creates a destructive robot programmed to kill Desdemona, or places hallucinogens in Othello's cup and feeds him lies, then Iago actually has a duty to rescue Desdemona to prevent the peril that he has created. One may even think that if Othello is about to shoot Desdemona, then Iago ought to throw himself in the bullet's path.[19] The criminal law seems to capture this sort of continuing duty in its requirement in both solicitation and conspiracy cases that defendants, to avail themselves of an abandonment defense, must not only abandon the criminal purpose but also thwart the success of others.[20] Once you light a fuse, it is your responsibility to stomp it out.

All of this is well and good because thus far we have only been dealing with the bullet that the provocateur has to take to prevent harm to others. The more complicated question is how to translate this principle when we are dealing with the relationship between the provocateur and the respondent. If at t1 the provocateur engages in conduct that provokes the respondent to try to harm him, he has curiously made himself both agent and victim. He has unleashed an unjustified risk of harm to himself!

---

a difference in kind. R.A. Duff, *Is Accomplice Liability Superfluous?* 156 U. Pa. L. Rev. Pennumbra 444, 450-51 (2008). Where knowledge is concerned, Duff argues that one may simply argue that what another person will do is not one's business such that one's prospective responsibility requires consideration of that consequence. Thus, argues Duff, a doctor may argue that although prescribing contraceptives may facilitate underage intercourse, that is not a factor she should consider, all that she needs to consider is the appropriate medical treatment of her patient. In contrast, when one acts with the purpose of encouraging the action, then, to Duff, one makes it one's business.

Moore rejects that the intervening-cause doctrine can serve to limit our obligations as Duff would have them limited. Although one may be entitled to expose oneself to risks because of one's own strong liberty interest, Moore denies that it follows for exposing others: "There is no moral privilege: to dress an attractive woman in provocative garb when I know rapists will find her in isolated situations; no privilege to send a jogger across Central Park at night to buy me a pack of cigarettes; no privilege to stack *your* flax too close to the tracks of a railroad I know to have inadequate spark arrestors." Moore, *supra* note ___, at 293.

I think we can grant some points to each side of this debate. First, in recognizing that culpability is a comparison of the risk one believes she is imposing as compared to her reasons for acting, we can agree, with Duff, that criminal behavior is a very bad reason for acting, such that most purposeful actions will be culpable. Whatever may permit us to outweigh harm done by others, when our very reason is to put that harm is motion, we are likely culpable for it. (Still, even these actions might be justified.) With Moore, however, we can also agree that just because it is not one's reason for acting does not mean that one is always permitted to disregard it. (Indeed, Duff might not disagree here. He might say that our prospective responsibilities are complicated.) Indeed, even if Iago did not tell Othello that Desdemona was unfaithful so he would kill her, but rather, just to torment Othello, but Iago consciously disregarded the risk that Othello would kill Desdemona, it seems that he ought to be held liable. What is needed, and I cannot provide a full account here, is how to account for when one imposes an unjustifiable risk and under what conditions we are permitted to exclude certain considerations from the calculation or what sorts of interests will always weigh heavily in a calculation.

[19] Victor Tadros, The Ends of Harm: The Moral Foundations of Criminal Law 53 (Oxford and New York: Oxford University Press 2011)("Given that I have created the threat through my own wrongful action, I must bear the burden of averting it.").

[20] MPC §§ 5.02; 5.03(6).

Although he typically would be under a duty to stop the harm to others, it seems that the exact opposite is true, which is that he is not permitted to intervene – he may not defend himself.

Alexander simply denies this. He claims that an actor may act at t_2; he does have a right to self-defense. The culpability is complete at t_1. However, it is not remotely clear why this is so. The t_1-t_2 culpability approach is failing to capture something. What it is missing is the fact that what the provocateur does at t_1 is to alter the normative relationship between the provocateur and the respondent.

B. Robinson's Unified Approach to Causing the Conditions of One's Own Defense

Paul Robinson's approach fares no better. Robinson proposes a single principle to apply across different defenses.[21] He argues that

> Where the actor is not only culpable as to causing the defense conditions, but also has a culpable state of mind *as to causing himself to engage in conduct constituting the offense*, the state should be [sic] punish him for causing the ultimate justified or excused conduct. His punishment, however, is properly based on his initial conduct of causing the defense conditions with his accompanying scheming intention, not on the justified or excused conduct he subsequently performs.[22]

According to Robinson, we take the provocateur's culpability at t_1 and link it to the result caused at t_2, but grant the provocateur a justification defense at t_2. This is intelligible (though I don't endorse the approach) for something like necessity. If Joe starts a fire negligently at t_1, and creates a firebreak at t_2, then the two can be combined such that although there was a purposeful harm at t_2, a crime with a negligence mens rea is that for which Joe will ultimately be held liable.

However, with respect to provocateurs, it seems rather odd to say that purposeful engagement in conduct at t_1 renders one criminally culpable for murder because a dead body is caused at t_2, but one is justified for the act of killing at t_2. Why say that the action is wrong at t_1 but right at t_2? Rather, we might think that the provocateur's act at t_1 alters his moral relationship with the respondent in a way that renders defensive force impermissible at t_2. In other words, it seems far more plausible to claim that the act the provocateur is not permitted to do *just is* engage in the killing at t_2. However, if provocateurs cause the conditions of their own defense and by that act forfeit the right to defend themselves then we need to understand how or why that happens, and Robinson's approach to other *actio libera in causa* cases is not going to get us there.

C. DeGirolami's Approach to Culpability and Justification

Marc DeGirolami advances a third approach to the puzzle, arguing that "created culpability" alters the later t_2 act in such a way that it is no longer justified.[23] DeGirolami's argument is that when a defendant has a culpable hand at t_1 in creating the situation at t_2, we must reassess whether the t_2 conduct is

[21] Robinson, *supra* note __, at 26.
[22] *Id.* at 31.
[23] Marc O. DeGirolami, *Culpability in Creating the Choice of Evils*, 60 Ala. L. Rev. 597 (2009).

socially valuable such that it ought to be deemed justified. Unfortunately, I think this approach while gesturing in the right direction ultimately fails as both as an approach for necessity or as an answer for provocateurs.

As argued above, I think it is clear that the later action is morally required. To the extent that DeGirolami wishes to reserve the term "justification" for praiseworthy acts, then there is some truth to the matter that those who owe a duty for having created peril are not performing justified actions. That simply depends upon whether one should group the morally right with the morally obligatory. Still, it seems clear that the t_2 act must be done and must be afforded a defense.

With respect to provocateurs, DeGirolami does little explain why the culpability at t_1 changes the nature of the conduct at t_2. I agree that it does, but we need a fuller account than the one at which DeGirolami gestures. I do not agree that culpability in causing can simply eliminate the justificatory nature of the later act because of something about *justifications*. Rather, it is how the t_1 act relates to the t_2 act that offers a substantive understanding of the later t_2 act.

II. The Use of Defensive Force by Provocateurs and Initial Aggressors

As discussed in the introduction, the claim that I am making is that provocateurs forfeit defensive rights, just as initial aggressors do, and that this forfeiture explains why initial aggressors and provocateurs are not permitted to engage in defensive force. In this section, I want to make a few general points about the nature of this claim, and then I want to sharpen our focus by distinguishing between aggressors and provocateurs.

A. Self-Defense is Not (Always)[24]a Lesser Evils Justification

As noted, Robinson viewed one virtue of his approach as its single approach to the problem across doctrines.[25] I do not think this works. The reason it does not work is because it presupposes that self-defense is a choice of evils defense and therefore that its analysis should follow the same process as necessity. However, it is extraordinarily doubtful that self-defense may be fully explained by a lesser-evils analysis. Consider the problems with this position. First, one must still give an account of why the culpable aggressor's life is discounted.[26] After all, why isn't the balance between the aggressor and the defender simply a draw? And indeed, how do we balance these lives if, in all other respects, the culpable aggressor has more value for society (a doctor working on the cure for cancer) than his

[24] I leave open the possibility that when there are innocents on both sides, numbers and balancing matters. Larry Alexander, *Self-Defense, Justification, and Excuse*, 22 Phil. & Public Affairs 53-66, 61 (1993) "the most plausible moral theory underlying common intuitions about self-defense is one that would be sensitive to, among other things, (1) number of deaths, (2) relative moral fault, (3) fair allocation of risks and incentives, and (4) nonappropriation of others.").

[25] Robinson, *supra* note __, at 26.

[26] Sanford H. Kadish, *Respect for Life and Regard for Rights in the Criminal Law*, 64 CAL. L. REV. 871, 882 (1976); David Wasserman, *Justifying Self-Defense*, 16 PHIL. & PUBLIC AFF. 356, 358 (1987) ("Unfortunately, the analogy begs the critical question of *why* it is a lesser evil to kill the aggressor.").

innocent victim (a criminal law theorist)?[27] Second, the consequentalist view seems to indicate that if many culpable aggressors attacked a lone innocent defender, there would be a point at which the balance would tip in favor of the aggressors.[28] But that simply cannot be right. You get to kill as many bad guys as threaten you.[29]

Moreover, consequentialist accounts are problematic even if we include the value of having a more general rule of self-defense. First, such a rule seems too narrow. If we simply want to deter violence, we might prefer a far broader rule, allowing for retaliation or other "punitive" acts.[30] Second, and more importantly, the approach seems to lose the importance of the relationship between the defender and the aggressor. The defender's act is not justified because the aggressor aims to harm her, but because of some greater societal value. And, this view also leads to the conclusion that a given defender is not justified in a case in which the action would not deter others' aggression. But, we would think that a culpable aggressor may still be killed in these instances.

To put this point another way, to view self-defense as serving some broader societal goal is to make it contingent.[31] Because the law will not be in a position to calculate in every instance, we will have a broad rule prohibiting aggression. However, in any individual case, it may be that that rule is over inclusive and the defender should not have the right to self-defense. However, it seems extraordinarily odd to think that the permission to use self-defensive force is always dependent upon the right consequences.

B. Liability Cases of Self-Defense Alter the Normative Relationship Between Aggressor and Defender

Although self-defense is a justification within criminal law, I take some instances of self-defense to ask questions prior to our categorization or understanding of defenses themselves (and certainly prior to the debate over the nature of justification).[32] If Jane takes a computer that is on the table, whether she has even committed an offense depends on whether the laptop was abandoned; or Jane was given permission to use it by Fred, the owner; or Fred gave the laptop to Jane; or Fred had not given Jane permission to use it. That is, to understand when one person has harmed another's legally protected interest, well, you need to know to whom the interest belongs.

---

[27] Wasserman, *supra* note ___ , at 359 ("The law permits the aggressor's life to be taken even if his survival is linked to other, innocent lives: a victim is entitled to lull an aggressor even if his killing is sure to provoke widespread bloodshed, or even if the aggressor is on the brink of discovering a cure for cancer or a solution to African famine.").

[28] *Id.*

[29] Kadish*, supra* note ___, at 882 ("For surely the rule allows one attacked to kill all his attackers no matter how numerous they may be.").

[30] *Id.* at 883 (noting that deterrence would support retaliation); Wasserman, *supra* note , at 360 (noting deterrence cannot explain retreat or proportionality).

[31] Kadish, *supra* note ___, at 883 (noting that the "argument rests on the contingent fact that justifying deadly defensive force will, in the long run, save more lives by deterring deadly assaults").

[32] For a survey of the debate, see Kimberly Kessler Ferzan, *Justification and Excuse*, in The Oxford Handbook of Philosophy of Criminal Law (Deigh and Dolinko, eds., OUP 2011).

Liability cases of self-defense are similar to these other usages of normative powers.[33] Because liability is an independent ground of permissibility, let me spend a minute defining the scope of the claim about self-defense. Ultimately, because provocateurs alter their rights, examination of liability-based instances of permissible self-defense is the place to start.

In analyzing self-defense, the first question is to figure out what we mean by "self-defense." As a matter of ordinary language our use of the term is extensive – I kill the Villainous Aggressor who tries to kill me in self-defense; I kill the rabid dog that is about to bite me in self-defense;[34] indeed, I even destroy your television that is flying at me during a tornado in self-defense.[35] Rather than assume that we ought to offer one normative justification for self-defense – one that will necessarily run roughshod over important nuances if it is to offer an account that includes villains and televisions-- I think it is best to think of self-defense as a type of defense, where different tokens (or subtypes, at least) may be normatively justified for different reasons.[36]

It is thus useful to distinguish between defender-centered theories of permissibility and aggressor liability-centered theories of permissibility.[37] Within the self-defense literature, theorists struggle with how to explain why it is permissible to kill a culpable aggressor, someone intends to cause you harm; an innocent aggressor, someone who will engage in a voluntary act that will harm you but lacks a culpable mental state because of mistake, immaturity, insanity, and the like; and an innocent threat, someone whose body will cause you harm but the bodily movement was involuntary such as a push.

Elsewhere I have argued that liability is its own interesting conceptual and normative path to permissibility, and the reason it is permissible to kill culpable aggressors is because they are liable to defensive force.[38] (Importantly, I have not maintained that liability is a necessary requirement for permissibility.) The "liability" formulation belongs to Jeff McMahan: "At least part of what it means to say that a person is *liable* to attack is that he would not be *wronged* by being attacked, and would have

---

[33] *Accord* Vera Bergelson, Victims' Rights and Victims' Wrongs: Comparative Liability in Criminal Law 110 (2009) ("If we were to define the principle of conditionality of rights in Hohfeld's terms, it would be characterized as the victims' power to change the balance of rights (in the broad sense) between themselves and the perpetrators."). Bergelson explains self-defense in terms of this conditionality: "If you try to kill me, you violate your duty to me and thus lose moral parity with me. That loss of moral parity reduces your right to inviolability and allows me to disregard it to the extent necessary to protect my right to life." Id. at 105.

[34] Indeed, in the case of Jimbo on *South Park*, the claim, "It's coming right for us!" allowed evasion of all hunting regulations. (South Park, Season 1, Episode 3, Volcano).

[35] One reader suggested to me that he would not deem this "self-defense" but rather "self-preservation." But this only proves my point. Arguing about ordinary language usage and then theorizing from there is not particularly useful. Eugene Volokh uses the term self-defense for the killing of bacteria in one's own body. Eugene Volokh, *Medical Self-Defense, Prohibited Experimental Therapies, and Payment for Organs,* 120 HARV. L. REV. 1814-15 (2007). I reject the usage. But what does that prove?

[36] As suggested by VICTOR TADROS, CRIMINAL RESPONSIBILITY 117 (2005); Jeff McMahan, *Self-Defense and the Problem of the Innocent Attacker*, 104 ETHICS 252, 256 (1994).

[37] This suggestion seems to originate in McMahan, *supra* note 4. McMahan has continually distinguished these concepts in his work. It was made quite explicit by Helen Frowe, *A Practical Account of Self-Defence,* 29 LAW & PHIL. 245 (2010).

[38] Kimberly Kessler Ferzan, *Culpable Aggression: The Basis for Moral Liability to Defensive Killing* (forthcoming Ohio State Journal of Criminal Law).

no justified complaint about being attacked."[39] McMahan further explains that being liable to attack just is having forfeited one's right not to be attacked.[40] Elsewhere, I have argued that a culpable aggressor forfeits his moral complaint against defensive force being used against him.[41] We can see the trappings of the loss of the right in the following possible implications: the number of aggressors does not matter; the aggressor is not entitled to compensation in tort for harms inflicted; third parties may aid the defender but not the aggressor; and the aggressor may not fight back. On the other hand, if one kills an innocent aggressor, it is certainly more debatable as to whether numbers do not matter or whether third parties should aid them or you.[42] If it is permissible to kill an innocent aggressor, the permissibility is not grounded in the aggressor's liability, as he has done nothing to forfeit his rights, but rather there is some other reason why you are permitted to infringe his right.[43] (Similarly, one may turn the familiar runaway trolley but not because the lone individual is liable to be killed. You are just permitted to infringe his right.)[44]

C. Distinguishing Provocateurs From Aggressors

Although both provocateurs and initial aggressors alter the underlying moral landscape, they do so in different ways. The important contrast to draw then is between the liability inherent in self-defense and the loss of rights that stems from the provocateur's act. We can put innocent aggressors and innocent threats to the side because the permissibility of harming them is not grounded in their liability. Rather, if one is permitted to kill innocent aggressors or innocent threats, the rationale will be something akin to an agent-relative permission to prefer one's life to others'.[45]

What distinguishes (culpable) aggressors from provocateurs? As a matter of ordinary language, we might start with something as simple as: provocateurs provoke and aggressors aggress. But that is not particularly helpful. Rather, I think the distinction lies in how the normative relationship is affected by the actor (be she a provocateur or aggressor) and the respondent. Namely, what aggressors do, but provocateurs do not, is engage in behavior that renders them liable to defensive force.

When an aggressor attacks her victim, by say, pointing a gun at her and saying, "I am going to kill you," she forfeits her right against the defender using force aimed at preventing the threatened harm from

---

[39] JEFF MCMAHAN, KILLING IN WAR 8–9 (2009).
[40] *Id.* at 10.
[41] *See* Ferzan, *supra* note __.
[42] Larry Alexander, *Self-Defense, Justification and Excuse*, 22 PHILOSOPHY AND PUBLIC AFFAIRS 53, 62 (1993).
[43] Ferzan, *supra* note __; *see also* Vera Bergelson, Victims' Rights and Victims' Wrongs: Comparative Liability in Criminal Law 76 (2009)(distinguishing culpable and innocent aggressors and arguing that the latter's rights are overridden).
[44] Dressler rejects a forfeiture theory for self-defense, arguing that it is over and under inclusive. Joshua Dressler, *Rethinking Heat of Passion: A Defense in Search of a Rationale*, 73 J. Crim. L. & Criminology 421, 454 (1982). It is under inclusive, he claims, because it cannot explain innocent aggressors. I agree that this is true, but as noted above, reject that we need one theory for self-defense. Dressler argues that it is over-inclusive because it would seem to allow the defender to kill the aggressor even when it is unnecessary. I demur. We alter our rights and duties in fine-grained ways. I can give you permission to use my car only on Tuesdays when it is snowing. When one forfeits one's right against defensive force, one forfeits one's right against *defensive* force.
[45] Alexander and Ferzan, 136-141; Jonathan Quong, *Killing in Self-Defense*, 119 ETHICS 507, 516–19 (2009).

coming to fruition. When a provocateur says, "I slept with your husband," the respondent may engage in force – and the wrongfulness of that force is something to be explored below -- but it seems clear even at this point in the discussion that we can say that the force that is used is not intended to be defensive force.

Jurisdictions conflate these actors but they are importantly distinct.[46] Because aggressors start the fight, their forfeiture of defensive rights naturally flows from how their behavior changes the normative relationship between aggressor and defender. [47] If A impermissibly uses deadly force against D, then D's use will be permissible and A's response will not be. (Even theorists who believe there can be conflicting justifications are not going to see A's action as justified here.) The only necessary tweaks then come when (1) there is a difference in proportionality between A's use of nondeadly force and D's return with deadly force and (2) A has ceased to aggress and specifying the conditions under which he regains his defensive rights. (I'm not saying these are complex tweaks, just that they do not affect the central case.) Provocateurs, on the other hand, need more specific rules because their conduct does not ground the permissibility of the respondent to act (he still acts wrongly) and because they are less able to unring the provocative bell. ("I've stopped attacking you!" is an easier claim to give normative force to then "Sorry I pissed you off on purpose! Takesy backsies.")[48] When an aggressor stops an attack, there is no need to defend. When a provocateur incites anger and rage, there is no way to undo the damage.

---

[46] *See, e.g.,* People v. Barnard, 567 N.E.2d 60 (1991) ("mere words may be enough to qualify one as an initial aggressor"). There is even some slippage in discussions, see e.g., Robinson, CC, at 6 ("Rather than using the term 'provoke,' some jurisdictions deny self-defense if the actor is the 'initial aggressor.'").

[47] *Cf.* Herrmann, *supra* note __, at 750-751 (noting the problems with assimilating provocateurs to aggressors, as provocateurs are not guilty of attempts). There are two classes of cases that lie at the provocateur/aggressor border. One group involves Inchoate Aggressors, those whose conduct does constitute an attack but whose conduct is not sufficient for the defender to respond because of additional requirements for the defender, such as imminence. The other group includes Culpable Apparent Threats. These individuals intend to make the defender believe that they are attacking but who lack the actual ability to carry out the threat (for example, robbing a store with an unloaded gun). I would cast both of these actors on the aggressor side, as opposed to deeming them provocateurs. *See* Kimberly Kessler Ferzan, *Culpable Aggression: The Basis for Moral Liability to Defensive Killing* (forthcoming Ohio State Journal of Criminal Law). Specifically, Inchoate Aggressors meet my condition 1(a) and Culpable Apparent Threats meet condition 1(b).

[48] To further complicate matters, an actor could be both an aggressor with respect to non-deadly force and a provocateur with respect to deadly force.

Interestingly, in many jurisdictions, if one intentionally provokes with the purpose of creating a defense, one cannot recover one's right to defend by withdrawal; whereas, if one otherwise provokes or aggresses, one can withdraw and recover the right to defend. Kansas' statute (Ks. Stat. §21-3241) is representative:

> The justification described in sections 21-3211, 21-3212, and 21-3213, is not available to a person who:
>
> (1) Is attempting to commit, committing, or escaping from the commission of a forcible felony; or
>
> (2) Initially provokes the use of force against himself or another, with intent to use such force as an excuse to inflict bodily harm upon the assailant; or
>
> (3) Otherwise initially provokes the use of force against himself or another, unless:

III. Provocateurs, Provocation, and Provocation Mitigation

If we agree that the question at issue is rights forfeiture (or specification, or conditionality, depending on one's view of rights),[49] then there are still questions we need to answer (1) does the provocateur forfeit rights and (2) by virtue of what is the rights forfeiture accomplished? If we are trying to understand a principle of rights forfeiture for provocateurs, the natural starting point is with the literature on when the respondent is entitled to mitigation for legally adequate provocation. Ultimately, I will argue that this is a theoretical dead end, but an instructive dead end nevertheless.

A. Provocateurs and Provocation

Who is a provocateur? First, let us make one important distinction. A respondent may be provoked by an act without the person who committed that act being deemed a provocateur.[50] That is, we must distinguish, "provocation" from "provocateur." An act can incite someone toward violence, even if the person who committed that act was not aware of its inciting properties. For instance, imagine that Ed is sleeping with Sally, but does not know, nor does he have any reason to know, that Sally is married. If Sally's husband Stan finds the two in bed together and shoots Ed, Ed's actions certainly provoked Stan's response. However, Ed, it seems, should not count as a provocateur as he was unaware of the fact that his actions could even have that effect.[51]

Notice that this yields that even if Stan would be entitled to mitigation for provocation, Ed is not a provocateur. At least, given that the question is when do provocateurs forfeit rights, it seems that even though Stan is provoked, we should not consider Ed a provocateur for our purposes. I think we would need a positive argument about why strict liability as to inciting a deadly affray should cause one to lose defensive rights.[52] I doubt that one will be forthcoming.

B. Provocateurs and Provocation Mitigation

> (a) He has reasonable ground to believe that he is in imminent danger of death or great bodily harm, and he has exhausted every reasonable means to escape such danger other than the use of force which is likely to cause death or great bodily harm to the assailant; or
>
> (b) In good faith, he withdraws from physical contact with the assailant and indicates clearly to the assailant that he desires to withdraw and terminate the use of force, but the assailant continues or resumes the use of force.

[49] Following David Rodin, I think nothing turns on how we conceive of rights for these purposes. See David Rodin, War & Self-Defense (Oxford and New York: Oxford University Press 2002), p. 74.
[50] I owe this point to Mike Cahill.
[51] A similar hypothetical is employed in Kadish and Schulhofer, Criminal Law and Its Processes, 5th ed. p. 845.
[52] *Cf.* Robinson, CC, *supra* note __, at 6 (noting that if provoking does not require fault, one could lose one's self-defensive rights if one's neighbor is provoked by the color one paints her house).

Indeed, in understanding provocateurs, I think it is also critical to notice that those conditions that render an actor a "provocateur" need not be identical to those conditions that entitle an actor to a provocation defense. A provocateur may provoke an individual into engaging in conduct that the individual is still not permitted to do.

Consider State v. Riley, in which the Washington Supreme Court devoted paragraphs (of dicta) to the question of whether "mere words" would be sufficient to constitute provocation such that an "initial aggressor" instruction ought to be given when a defendant claims self-defense.[53] That is, the defendant claimed that his use of words about the victim being only a "wanna be" in his gang did not rise to the level of warranting an initial aggressor instruction (and violated his First Amendment rights).[54] Because there was evidence suggesting that the defendant did more than just verbally taunt the victim, the court did not need to decide the mere words issue. Nevertheless, the court stated, "we hold that words alone do not constitute sufficient provocation."[55] One of the court's primary arguments for this position was that:

> such a rule would effectively permit violence by a "victim" of mere words, contrary to the underpinnings of the initial aggressor doctrine. As noted, the initial aggressor doctrine is based upon the principle that the aggressor cannot claim self-defense because the victim of the aggressive act is entitled to respond with lawful force. For the victim's use of force to be lawful, the victim must reasonably believe he or she was in danger of imminent harm. However, mere words alone do not rise to reasonable apprehension of great bodily harm.[56]

As should be clear, the problem is that by linking provocateurs to aggressors, the court misses the fact that the reason why a provocateur is not entitled to respond cannot be the same as the reason why the aggressor is not entitled to respond. Indeed, but-for the fact that initial aggressors can regain defensive rights (by withdrawing), there would be no need for an initial aggressor instruction. The aggressor uses unjust force; the defender responds with just force; and therefore, the aggressor is not entitled to respond with just force. This is just the upshot of the clear conceptual linkage between aggression and self-defense. Provocateurs are importantly distinct in that there behavior does not—contra the court's view – *justify* the violent response. (Even to the extent that provocation is seen as a partial justification, the behavior does not fully justify the respondent's use of force, which still makes these actors different than initial aggressors.)

---

[53] 976 P.2d 624 (Wash. 1999)(en banc).

[54] Washington's aggressor instruction is as follows:

> No person may, by any intentional act reasonably likely to provoke a belligerent response, create a necessity to act in self defense and thereupon use, offer or attempt to use force upon or toward another person. Therefore, if you find beyond a reasonable doubt that the defendant was the aggressor, and the defendant's acts and conduct provoked or commenced the fight, then self-defense is not available as a defense.

Id. at 627 (citing II Wash. Pattern Jury Instructions; Crim. 16.04 (2d ed. 1994)). The instruction conflates aggression and provocation -- provocateurs "provoke" "belligerent" responses but aggressors "commence" fights.

[55] *Id.* at 628.

[56] *Id.* at 629.

Notably, Judge Talmadge, writing a concurrence, sees exactly the problem. He rejects the majority's position that

> No matter what one says, no matter how provocative, no matter what the circumstance of the provocation, a speaker may always assert self-defense if attacked by the person the speaker provokes into attacking, and the State is never entitled to an aggressor instruction. The majority's rule defies human nature. [57]

The judge then offers hypothetical cases to prove his point. Here is his hypothetical that I employed in the introduction:

> Imagine a funeral ceremony with hundreds of mourners for a widely respected African-American civil rights leader. A white supremacist appears at the church and begins shouting nonthreatening, racial epithets. Enraged mourners rush the person, who pulls out a concealed gun and kills several of them. At his trial for murder, he argues self-defense.[58]

Judge Talmadge then notes, "Under the majority's reasoning, because he used only words to provoke the attack, the white supremacist was not the aggressor and the State is not entitled to an aggressor instruction."[59]

At this point, we should have at least clearly established, as noted above, that provocateurs are not initial aggressors. But I think we can draw a second implication as well, which is that what we are interested in is the type of conduct that may cause the provocateur to lose rights, and not all such losses will themselves be sufficient to give the respondent the right to go on the attack.

C. Provocateurs' Rights and the "Partial Justification" Theory of the Provocation Defense

If our aim is to understand when provocateurs are culpable and when they forfeit rights, we have now eliminated two dead ends. One is that we cannot understand provocateurs simply by analyzing acts that are provoking. Second, we cannot understand provocateurs by analyzing when a respondent is entitled to a provocation defense.

We may ask, however, whether the theoretical understanding of the provocation defense could shed some light on the question of how provocateurs forfeit rights. To this point, I have eschewed the use of justification and excuse, and I have also argued that provocateurs need not act in a way that grounds a provocation defense for the provocateur to lose rights. Yet, in this section, I intend to look specifically at provocation and even the question of whether it is a justification or an excuse. Let me explain.

With self-defense, the reason why a defender is entitled to act is because an aggressor is *liable* to defensive force. Provocation, a defense which reduces murder to manslaughter, does not look like this because the provocateur is not liable to defensive force. Yet to ask the question whether the provocateur "had it coming" requires further analysis of the provocateur's rights, and the best analysis

[57] *Id.* at 630.
[58] *Id.* at 631.
[59] *Id.*

of these rights exists within the literature that says that the responder is *partially justified* in using force against the provocateur. The questions are whether that is true and if so, why it is so. To skip to the conclusion, I will ultimately claim that the partial justification view cannot explain why provocateurs cannot fight back.

The first thing to note is that one need not adopt one rationale for when provocation mitigates. Indeed, Vera Bergelson argues that some instances of provocation are partial excuse *and* partial justification and other instances are simply partial excuse.[60] So, we might have two different accounts of the mitigation. Moreover, the goal here is not to rationalize existing doctrine but to get to the heart of the underlying moral questions.

Notably, there seem to be cases where our intuitions favor killing the provocateur. In the movie, *Gladiator*, Commodus tells Maximus about how Maximus' wife and son were murdered, reporting that his son "squealed like pig when he was nailed to the cross," and his wife "moaned like a whore when she was ravished again and again and again..." When Maxium kills Commodus, the audience does not think, "Well, there is an instance of legally adequate provocation. His rationality was diminished and he is certainly entitled to a partial excuse." *The audience applauds*. No one argues that Maximus should go to jail at all. When we cheer for Maximus we do so because Commodus is getting what he *deserves*, not because of the taunting words but because of Commodus' role in the commission of the underlying acts. There are two notable features about the case (1) a crime has been committed which deserves substantial punishment and (2) there is a corruption of the legal order such that we cannot expect the law to administer justice.

Indeed, more generally, the argument that provocation is a partial justification must be grounded in the fact that the provocateur deserves at least some of the harm he receives. Although the original grounds for the provocation defense were based upon those acts that threatened a man's honor, such that he might be said to be "defending" his honor, the "required" retaliation should not be considered any sort of true self-defense.[61] Rather, the man's actions are meant to demonstrate that he is the sort of individual who views this conduct as a sufficient affront such that retaliation is proper, and he demonstrates his manhood by responding in this way. To update the example a bit, if a mother saw her child brutally murdered in front of her and did not respond with rage and violence, we might question her motherhood status, but this is not to say that the conduct is designed to defend her motherhood

---

[60] Bergelson also notes that the label "provocation" masks different claims, though she carves up the pie differently than I:

> [T]he current law of provocation shelters under its roof two conceptually different defenses. The first, "true" provocation, is a mix of partial justification and partial excuse. Its justificatory capacity stems from the responsive character of the perpetrator's attack on the provoker who was the initial aggressor and violated some important legal rights of the perpetrator….
>
> The other defense, which, for the sake of clarity, may be called the "heat of passion," is purely excusatory. It applies in the circumstances when the perpetrator does not have a legal right that the victim act in a certain way but has some moral claim or legitimate expectation that the events that caused his loss of control not happen.

Bergelson at 89.

[61] Jeremy Horder, Provocation and Responsibilty 24-26 (1992).

status as much as the very act is a demonstration of it. Indeed, as Horder later notes, the aim is not defensive but retributive, and the reason why mitigation and not exculpation was appropriate was because the punishment was too severe.[62]

If part of what grounds the provocation defense is that the provocateur gets what he deserves, and this rationale exists irrespective of the emotional state of the actor, [63] one may question why there is also a requirement that the respondent be in a particular emotional state. Although I need not answer this question (as I am not attempting to justify the provocation defense but rather to see how we are to understand the provocateur's loss of rights), Berman and Farrell argue that "partial excuse and partial justification are necessary and sufficient conditions for provocation manslaughter."[64] They reason that although mitigating reasons flow from both excuse and justification reasoning, neither alone is sufficient to warrant a departure from murder.[65] More importantly for our purposes here is the fact that sometimes provocateurs act in such a way that they deserve injury, and it is this feature of desert that reduces the wrongfulness of the respondent's behavior.

Notably, if this desert feature is what renders the respondent's conduct less wrongful, the questions of when and whether the provocateur forfeits defensive rights are still left unexplained. Except in *Gladiator* cases – that is cases with clear failures of rule of law and extraordinary desert where we seem to have no problem with vigilantism or the death penalty – what a respondent does is to inflict excessive punishment in violation of rule of law values. So, the desert account cannot fully explain why the provocateur loses *defensive* rights. Hence, even understanding why provocation is a partial justification cannot explain why the provocateur is not entitled to fight back against the residual unjustified force.

IV. Provocateurs and Actio Libera in Causa

With a better understanding of provocateurs – one that distinguishes them from aggressors and from the provocation's mitigation provisions– we may now turn directly to provocateurs and what I view as the two *actio libera in causa* puzzles for provocatuers. The first puzzle is what explains why provocateurs can't fight back. The second puzzle is whether in determining whatever those conditions are at t1 that lead the provocateur not to fight back, we are to consider that sometimes the provocateur views himself not simply as killing but rather as killing *justifiably*. That is, is the provocateur allowed to take into account the later impermissibility of the respondent's behavior in determining whether he may engage in conduct at t_1?

A. Provocateurs and Rights

[62] *Id.* at 52.
[63] Mitchell N. Berman and Ian P. Farrell, *Provocation Manslaughter as Partial Justification and Partial Excuse*, 52 William and Mary Law Review 1027-1109, 1069 (2011).
[64] *Id.* at 1034.
[65] *Id.* at 1065. Berman and Ferrell employ the conceptual landscape of reasons, rather than rights, arguing that desert gives the respondent a reason to act that renders the respondent's behavior less wrongful, rather than arguing that the provocateur has lost any rights. *Id.* at 1088-96. Notably, Berman would be unlikely to construe this as forfeiture in any event. Mitchell N. Berman, *Punishment and Justification*, 118 Ethics 258-290 (2008).

Let's start with cases in which we will assume the provocateur is not justified at t_1 in provoking. (We will reach the question next of how to determine when the provocateur's action is justified.) How is it that the provocateur alters his rights?

Unfortunately, the literature on this question has been tethered to the question of whether the respondent is entitled to a provocation defense. As discussed above, these are two separate questions. For instance, in analyzing provocation, Bergelson offers the "conditionality of rights" principle, a principle that covers self-defense, provocation, and consent. But this is too broad.[66] Normative powers are likely to be more complex than one test.

Does the provocateur consent to be harmed? Typically, a person who consents to a harm waives a moral complaint. His consent transforms the conduct.[67] However, even if the provocateur wants to incite the respondent to hit him, the provocateur wants (perhaps he does not succeed, we shall see….) to keep his moral complaint.[68] That is, he claims the respondent still wrongs him by hitting him and that is why he wants to be able to fight back. Our understanding of consent is also that it is more voluntary than this; that is, when one consents to harm, one acquiesces to the contact that will occur. But this seems far too strong for the provocateur. We don't think, for example, that Charles Bronson is acquiescing to being mugged.

Along similar lines, another approach would be to claim that the provocateur essentially harms himself because he acts through the respondent.[69] Even if I were a fan of the relevance of causation to criminal responsibility (which I am not[70]), this approach is still problematic. The problem is that the respondent is still culpably responsible for the attack. And so, we still do not have an account of what excludes the action of the provocateur vis-à-vis the respondent. Unlike tort law, where the responsibility pie is limited to the damages the plaintiff suffers and therefore may be divided in proportion to responsibility, criminal law has room to hold more than one person fully accountable for their actions.[71] Both principal and accomplice can go to jail for the offense. The problem then is why, if the provocateur is "harming himself" and the respondent is "harming the provocateur," the provocateur still loses standing to defend against the respondent's action.[72]

The answer actually lies at the surface of the problem. Provocateurs forfeit defensive rights for the very simple reason that provocateurs start the fight. They initiate fights without initiating aggression. They are actually stealth aggressors. Although initial aggressors start fights by throwing first punches, stealth

---

[66] *Cf.* Alon Harel, *Victims and Perpetrators, The Case against a Unified Theory of Comparative Criminal Law*, 8 Buff. Crim. L. Rev. 489, 490 (2005)(arguing one principle is too broad to be useful); Kenneth W. Simons, *The Relevance of Victim Conduct in Tort and Criminal Law*, 8 Buff. Crim. L. Rev. 541 (2005)(noting the complexities and rejecting a single principle of comparative liability).

[67] Heidi M. Hurd, *The Moral Magic of Consent*, 2 Legal Theory 121 (1996).

[68] *Cf.* Herrmann, *supra* note ___, at 749-50 (rejecting the German view that a person who provokes the attack does not intend to defend himself).

[69] *Cf.* Finkelstein & Katz, *supra* note ___, at 488 (discussing the perpetration-by-means approach).

[70] *See* Alexander and Ferzan, *supra* note __, ch. 5

[71] R.A. Duff, *Responsible Victims and (Partly) Justified Offenders*, 8 Ohio St. J. Crim. L. 209-216, 210-211 (2010).

[72] Indeed, with the bee example noted below, we do not say that the person who provokes bees to sting her is harming herself in such a way that she is not entitled to kill the bees.

aggressors start fights by inciting the other person to anger or otherwise causing the other person to do violence.

This is just a sense of "asking for it." What a provocateur does is to taunt someone in a way that he knows creates the risk of a harmful retaliatory reaction. In the same way that one does not poke tigers with sticks or swing at beehives with bats, one likewise cannot act in a way that one knows will provoke and then complain about the reaction. (Admittedly, one may defend against bees and tigers, both because of differences in the underlying normative relations and differences in the comparative value of the person and animal.)

A provocateur creates the risk that another person will harm him by engaging in conduct that he knows may produce a violent response, either because it will anger the respondent or otherwise give him a reason to engage in violent conduct. Having created this risk of harm to himself, he forfeits his moral complaint when this very risk materializes. It is true that this conduct is insufficient to justify the respondent's behavior, but sometimes, there are just two wrongdoers. And two wrongs doesn't make either of them right.

B. *What the Provocateur Must Do to Forfeit His Defensive Rights*

The next question, then, is what is required for this sort of moral forfeiture. I want to suggest here that something along the lines of recklessness should be required. With respect to provocateurs, the law typically requires something extraordinarily narrow – *purpose to provoke the affray*. Like complicity, there seems to be some implicit view within the law that one only identifies with the wrong done by another if one purposefully causes it. But, this is too narrow. Why should not any subjective mental state suffice? In the same way that one cannot complain when one purposefully provokes a tiger, it seems that one cannot complain if one consciously disregards the unjustifiable risk that one will provoke a tiger. If Andrew wants to get his friend's attention but does so by blasting a loud horn by a dog he knows hates loud noises, why would Andrew have a valid complaint against the dog's response?

Moreover, it seems that a subjective mental state ought to be required. My more general worries about negligence have been vetted elsewhere,[73] but I think we have reason to worry about a forfeiture principle that is not connected to a person knowing he is doing something impermissible. If Andrew does not see the dog, we might say to Andrew, "Well, you ought to have seen the dog" but I think it is more doubtful that we will say to Andrew, "as between you and the dog, we don't feel sorry for you because you provoked him."

Indeed, a broader provocateur doctrine would likely turn tort law on its head. If Alice negligently drives across the media and Bob has the ability to avert an accident, he is obliged to do so under the last clear chance doctrine. But if Alice is deemed a provocateur, then any injury done by Bob would not wrong

[73] *See* Alexander and Ferzan, ch. 3.

Alice and so she would not be entitled to compensation in tort. Bob, even if he is annoyed by Alice, is not permitted to stand on his rights and hit her car.[74]

Now, certainly the human respondent is not a dog, but rather, is a rational creature. But this, it strikes me, is all the more reason to require subjective appreciation on the part of the provocateur. Human relations and human anger are exceedingly complex and the respondent is still engaging in conduct that he ought not to engage in. To take an extreme example, assume someone utters a significant racial slur, completely oblivious to its inflammatory meaning, and this provokes an attack from a member of the racial group. It seems rather extraordinary to think that this act is itself going to be sufficient to warrant a moral principle preventing the negligent provocateur from defending himself against the wrongful attack.

Even if we require a subjective appreciation that one is provoking another toward violence, we would then need to ask the question of whether one loses one's defensive rights even when the risk is itself justified.[75] I think this would be too broad, as it would hold us hostage to the most despicable but predictable actions of others. We need an account that is not so broad as to forbid defending against rapists who attack the scantily clad but is not so narrow as to require violation of a legal right.[76] We need an account of unjustifiable provocation that respects the liberty interests we have in walking streets, dressing immodestly, and speaking the truth. This is ultimately then just part of the puzzle about when we are permitted to engage in actions that may encourage impermissible conduct by another.[77]

Still, there are other remarks that can be made now. First, the question of whether the provocateur's conduct created an unjustifiable risk does not affect the impermissibility of the respondent's response. He still is not allowed to attack in any event. Second, in determining whether such risks are justified, we must take into account how our consequentialist calculus is deontologically side-constrained. For instance, when Leo Katz asks about the gateway sin paradox, where one will cause more good if one commits the "gateway sin," the answer I would give is that it depends upon whether the gateway sin requires the use of another person as a means.[78] Unlike those who endorse the doctrine of double effect or other formulation of our deontological constraints, Larry Alexander and I have argued that our consequentialist calculus is constrained to the extent that one cannot use another as a mere means. [79]

---

[74] Hurd, *supra* note __.

[75] Morally, one will only forfeit defensive rights about the kind of reprisals one foresees. Thus, one would still retain defensive rights against unforeseen responses. Statutory drafting would need to get this right.

[76] Bergelson advocates legal. Bergelson, pp. 114-118. *Cf.* Douglas Husak, *Comparative Fault in Criminal Law: Conceptual and Normative Perplexities*, 8 Buff. Crim. L Rev. 523 (2005)(noting that mitigation must be based on morally, not legally, suspect behavior).

[77] *See supra* note ___.

[78] Leo Katz, *Preempting Oneself: The Right and Duty to Forestall One's Own Wrongdoing*, 5 Legal Theory 339-362 (1999).

[79] *See* Alexander and Ferzan, ch. 4; Larry Alexander and Kimberly Kessler Ferzan, *Moore or Less Causation and Responsibility*, (Criminal Law and Philosophy).

And to me, it is only in context that we can determine whether there was a gateway sin in the first place.[80]

V. The Final Puzzle: Assessing the t_2 Act at t_1

This brings us to the final puzzle. Is the provocateur entitled to consider that his response will be justified in determining the initial culpability of his conduct? We left off in the last section with how to understand whether the conduct is justified given the wide array of reasons that the individual can have for performing it. After all, if Bully threatens Victim that if she leaves her house, then he will shoot her, then Victim knows that she is consciously disregarding the substantial risk that she will need to act in self-defense if she leaves. The question then is whether leaving the house is justified, not whether leaving the house is legal.[81] The reason she may leave, even knowing that she will be attacked, is because the initial action is not culpable, even including that it may lead to a later need to kill.

Sometimes at t_1 we act so as to be able to act justifiably later and we are entitled to take this later justifiability into account in determining whether to engage in the conduct at t_1. Carl is entitled to apply for a job as an executioner because he wants to justifiably kill people. Debbie is entitled to drive an ambulance because she wants to justifiably speed through red lights. A police officer is allowed to act undercover and even encourage another to commit a crime so that a gang can be brought down. In these cases, the act at t_1 is causally linked to an act that is justifiable at t_2.

This brings us back to Bronson. Why can't Charles Bronson say that all he did was cause a bad guy to try to kill him and that allowed him to act in self-defense? This is the final puzzle. Assume a provocateur acts at t_1 in such a way so as to provoke the respondent to try to kill him so that the provocateur can kill him at t_2. Is the correct question, "May one engage in conduct so that one may purposefully kill?" or is the correct question, "May one engage in conduct so that one may purposefully and justifiably kill?" That is, can the provocateur build in the justifiability of his conduct at t_2 in determining whether his culpable and therefore impermissible at t_1?

Finkelstein and Katz present this as a question of whether "we must assess plans—or component parts of plans—in the context of the overall moral character of the entire package."[82] Certainly, we do. An action at t_1's justifiability critically depends upon its effects at t_2 and the reasons for causing these t_2 effects. What I reject here, however, is that *even taking into account that he would be acting in what is otherwise justifiable self-defense*, Bronson's conduct is justified at t_1. In my view, Bronson cannot act at t_1 because his conduct purposefully aims to usurp the role of the state, and this is an impermissible goal.

Because I have never been able to have a discussion about these cases without someone referring to a movie, let's get the three most plausible candidates on the table:

---

[80] Katz raises a number of puzzle I cannot resolve here, including the question of why we are inclined to excuse someone when character traits are brought on by voluntarily ingested drugs and cause rational but blameworthy conduct but we are inclined to blame those born with such traits.

[81] Margaret Raymond, *Looking for Trouble: Framing and the Dignitary Interest in the Law of Self-Defense*, 71 Ohio St. J. Crim. L. 287-339 (2010).

[82] Finkelstein and Katz, *supra* note __, at 502.

Shane: In this Western, Jack Palance plays Jack Wilson, a gun slinger hired by antagonist, Rufus Ryker, who is trying to force homesteaders off their land. When one homesteader approaches where Wilson is standing, it is clear that Wilson is to shoot the homesteader but to make it look "right" to witnesses. Wilson therefore proceeds to insult the South and Southerners, which provokes the homesteader to draw his gun. (Despite the homesteader's having done so somewhat slowly and pathetically, where the bullet will hit Wilson in the toe at best), Wilson, quicker on the draw, shoots the homesteader.

Death Wish: After muggers kill his wife and severely would his daughter, Charles Bronson's character goes on a vigilante spree. On streets, subways, and parks, he poses as a victim, so that when attacked he can kill his attacker.[83]

Dolores Claiborne: Dolores Claiborne lives with her physically abusive alcoholic husband and her teenage daughter in Maine. Once Dolores discovers that her husband is molesting the daughter and that her husband has stolen all of Dolores' private savings with which to leave, Dolores plots her husbands demise. She buys him alcohol and then verbally taunts him, so that he will chase though the field on their property. Dolores, who has previously almost fallen down a hole in the field, has covered the hole. She jumps over it; he falls in; and she leaves him to die.

These movies have a great deal of background noise, however, and it is worth cleanly separating out the various factors. In Dolores Claiborne, the husband is molesting the daughter. The audience is led to believe that the wife is in a situation where she will not be able to end the abuse and save her daughter. We are told, by Dolores' employer, "Sometimes being a bitch is all a woman has to hold on to." The husband *deserves* quite a bit of punishment. The state is not going to give it to him. Although from personal conversations I know that there are those who think she did nothing wrong, that seems to go too far. After all, she gets him intoxicated, taunts him so he will attempt to kill her, and leads him on a chase so that he will fall down a hole she has covered up and she herself jumps over. But even if all of that was permissible, she still had a duty to rescue him. It is only because we believe the system will fail her that we don't condemn her. *Death Wish* likewise raises issues as to law enforcement and the rule of law. We like to know there is justice in the world (or at least at the movies).

Now, turning to *Death Wish* more directly, what should we say about Bronson's conduct? Paul Robinson clearly places the protagonist on the culpable side of the divide: "Perhaps the 'grand schemer' best known in popular culture is the vigilante character of the 1974 movie, 'Death Wish,' who deliberately takes late-night walks and subway rides to place himself in threatening situations where he could kill the muggers who tried to attack him."[84]

Recall that what provocateurs do is to affect the respondent's reasons for action, typically by creating new ones. Iago tells Othello that Desdemona is cheating on him. By his words, Iago aims to influence

[83] Interestingly, the New York Times review of the movie concluded, "it's a despicable movie, one that raises complex questions in order to offer bigoted, frivolous, and oversimplified answers." Vincent Canby, Death Wish (1974) Movie Review (July 25, 1974).
[84] PHR, CC, at 31 n.114.

Othello's behavior. It is tempting to think that there is a causal distinction that can be drawn, and we might analogize this to the entrapment defense.[85] No one finds it remotely problematic for police to merely leave a wallet out for someone to find. Creating opportunities seems unproblematic. Thus, it is tempting to say here that what one can do, and Bronson does, is to merely create an opportunity, and this is entirely permissible. But this is too quick. After all, creating an opportunity creates a reason for the respondent that did not exist before. However, this is always true. In every case, the police *cause* the crime. They always influence the defendant's reasons for action. It does not seem that a principled causal distinction can be drawn.[86]

Let us parse two distinct issues that entrapment can raise. First, we must always remember that causation is not compulsion and causation itself is not an excuse.[87] All human actions are caused. Thus, any test for entrapment that is meant to be exculpatory would require an understanding of how the police presented a defendant with "too hard a choice" that would be akin to the duress defense. It would also need to enter the fray as to whether offers should be distinguished from threats.[88] Threats and offers have different effects on the moral baseline, but may have similar effects on the agent's rationality, and the question is why to privilege the former of the latter when formulating excuses. Second, we might also think that *even if the defendant is not entitled to an exculpatory defense because his will was not (figuratively) "overborne"* there are just some things the police cannot do and thus are estopped from enforcing.[89] Particularly in the area of sentencing manipulation, we might think that police who engage in conduct just to increase the defendant's punishment are acting with unclean hands, despite the fact that the defendant is fully responsible and blameworthy for her actions. In many cases, these two theories may overlap (usually if the government's hands are dirty, it is from creating duress-type situations), but this need not be the case.

To the extent that the private citizen does overbear another's will or engage in extraordinarily distasteful behavior, he will not be able to say that his conduct is justified. If the justification in acting is to say, reveal that someone is a hardened criminal who will hurt others and stop that person, then situations that undermine this showing undermine the justifiability of the actor's conduct at t_1.

On the other hand, this does not seem true of Bronson. Taking the excuse theory first, Bronson, our provocateur, might argue that although his conduct had causal powers, it did not overbear the respondent's will. "I didn't *make* him attack me." And, using the unclean hands theory, Bronson might similarly say that nothing he did was so inappropriate so as to cause him to forfeit his own right of response.

---

[85] I thank Larry Alexander and Doug Husak for suggesting this analogy and Mike Cahill for urging me to pursue it more than my first draft did. For a general survey of the doctrine, see Ronald L. Allen, Melissa Luttrell, and Anne Kreeger, *Clarifying Entrapment*, 89 J. Crim. L. and Criminology 407-431 (1999)

[86] Dan Squires, *The Problem with Entrapment*, 26 Oxford J. Legal Studies 351 (2006).

[87] Moore, Placing Blame, p. 538.

[88] Anthony M. Dillof, *Unraveling Unlawful Entrapment*, 94 J. Crim. L. & Criminology 827-895 (2004), pp. 849-852.

[89] Along with entrapment, there is the possibility of a due process violation for outrageous government conduct. *See* U.S. v. Citro, 842 F.2d 1149 (9th Cir. 1988).

Still, the entrapment analogy yields insights for the case at hand. It seems that underlying entrapment is a concern about external actors pulling the strings on our circumstantial luck.[90] In many ways, I think that we feel "But for the grace of God go I" and so, we might worry that there are circumstances at which we would meet our price, either by threat or offer. One thus tries to manage one's life to avoid such situations. What both the provocateur and the entrapping officer do then, is to shift our opportunities in a way that may reveal the worst about us. And the fear is that anyone would be a criminal at some point. Hence, in those cases that do not exculpate the respondent, the question is whether there is something simply impermissible with modifying another's circumstantial luck in such a way as to take advantage of his shortcomings.

The police, it might be thought, are entitled to slight but not extraordinary shifts in circumstantial luck. So, if a police officer poses as a marijuana dealer, then the buyer cannot complain that she picked the police officer. And importantly, that seems to be partly driven by the fact that if the buyer did not buy from this cop, the buyer would buy from another. [91]

What courts and commentators are reaching for, it seems to me, is the idea of possible worlds and the question of whether there had been a nearby possible world where the crime still occurred versus whether the police shifted the defendant into circumstances that only arise in far away worlds (and this one, because of the police.) This would largely explain that the predisposition test in entrapment law is not really about propensity, but about this extraordinarily difficult to capture idea that we want to determine what the person's luck and actions would have been without the intervention, and the best way to do that is to look at facts about the person as evidence of the likelihood that he would have committed the offense in a nearby possible world. Moreover, this analysis also sheds light on the valid insight that entrapment exacerbates distributional inequities.[92] Drug dealers sell drugs in nearby possible worlds, but housewives don't. But it is largely? somewhat? a matter of circumstantial luck which one you are to begin with.

Having articulated the intuition that animates this analysis, let me be clear about one thing. I am not adopting the metaphysics of possible worlds. Indeed, I am extraordinarily skeptical of their usage, and our ability to run the sort of counterfactuals the test requires.[93] I am not endorsing a possible worlds answer to entrapment's problem. However, I think that this explanation gets at the root of the intuition

---

[90] *See also* Dan Squires, *The Problem with Entrapment*, 26 Oxford J. Legal Studies 351 (2006). Tony Dillof is right to note that in playing God, there is also the problem of selection. Anthony M. Dillof, *Unraveling Unlawful Entrapment*, 94 J. Crim. L. & Criminology 827-895 (2004).

[91] This circumstantial luck analysis shares the line of thinking about above-market transactions offered by legal economists. *See* Allen et al. *supra;*Richard H. McAdams, *The Entrapment Defense Defended*, *in* Criminal Law Conversations 509-511 (Robinson, Garvey, and Ferzan eds. 2009)("my defense of the entrapment defense is consequential: that without the defense, police would waste society's resources seeking valueless arrests of individuals who would offend in sting operations but not in the real world; that the system would inflict needless suffering by punishing such people; and that the unfettered power to target individuals in sting operations presents the danger of political abuse..."). The thought is that there may be little reason to deter those who would not have committed an offense in a nearby possible world.

[92] *Cf.* Squires; Louis Michael Seidman, *Entrapment and the "Free Market" for Crime*, *in* Criminal Law Conversations 493-513 (Robinson, Garvey, and Ferzan eds. 2009).

[93] Alexander and Ferzan, Moore or Less Causation and Responsibility.

that the cops are doing something profoundly unfair in some cases. And, importantly, this is not about exculpating the accused or showing what he would have done in another world so that we cannot blame in this one. It is simply a claim about how far the police are permitted to stray from what the average citizen may expect to confront in his daily life.

Even if we can make sense of the possible worlds insight, the question is what follows from it. There is certainly going to be a possible world in which the respondent does not act and one in which he does (this one) and worlds in between. It is also the case that what the police officers do, and what the provocateur does, is both (1) to play God and (2) to reveal true facts about the respondent.

Now with police, our ability to draw the line on (1) can only be accomplished by looking at the precise relationship between the citizen and the state. We think that there are just ways the police should react vis-à-vis its citizens. It should not seek to too profoundly alter their moral luck.

What though, to say about the provocateur? Notice that what the provocateur does is to ultimately reveal that under certain conditions the respondent will behave wrongly. May the provocateur play God so as to justifiably rid us of a potential wrongdoer?

Let us take a strong case. Elise strongly suspects that Fred is a rapist. She puts herself in a situation where she is alone at night with him and brings her gun so that she can shoot him. Maybe, she also knows that talking about feminist literature angers Fred and will make him more likely to act on his misogynistic tendencies. And she also knows that his mother is a particularly sore subject with him (the root of all his hatred vis-à-vis women) and so she brings up Fred's relationship with his mother.

Before adjudicating Elise's case, let's be clear about where we are in the analysis. Above I argued that when the provocateur is reckless as to causing another to use harm against him, the provocateur forfeits his defensive right. I also noted that the justifiability of the risk is crucial in these cases because one may foresee risks that one is permitted to create. A bully cannot tell you that if you ever leave your house, he will kill you, such that your liberty is hostage to the scope of his threats – no matter how foreseeable it is that the bully will act on those threats. Because the provocateur may impose justifiable risks of incitement, the question is whether in determining whether the risk is impermissible, the actor is entitled to take into account that her later behavior could be viewed as justified. The argument seems almost, but is not quite, circular. If one may kill unjustified respondents, then one simply does not forfeit one's defensive rights at t_1. If the goal is to cause the conditions of your own defense, but you believe that the defense is ultimately justified even at t_1 – that is, the world is better off, or no worse off by your engaging in the conduct at t_1 – then there is no forfeiture of defensive rights at t_2.

The problem with Elise's conduct, and with Bronson's as well, is that they are vigilantes. Although they attempt to flush out the evil tendencies of others, the question is why they should be entitled to play that role. What they aim to do is to give people their just deserts under the guise of preventive force. They aim to replace the state. The problem about making oneself judge, jury, and executioner is that one is, in fact, taking over a role that has been delegated to the state. In the absence of a well-functioning legal order, we might revisit whether Bronson or Elise may act as they do. We could still worry about the differences in proportionality between the harm employed as a matter of prevention

and the harm deserved as a matter of retributive justice. But then our complaint would be more narrow. However, within our legal order, private citizens are not permitted to play both God and government and rid the world of wrongdoers. The unjustifiability of the conduct is thus the extent to which this behavior violates the rule of law values that the criminal justice system serves. This leaves open that sometimes the events at t_2 will justify the risk at t_1 – indeed, this must be true for us to lead our lives when we predict the future choices we may face.

Conclusion

Criminal law scholarship's failure to be sufficiently attentive to the way that the normative landscape may be changed by the conduct of the actors renders scholars ill-equipped to articulate why a provocateur cannot fight in her defense. This is not a matter of understanding defenses. Or justifications. It is a matter of understanding how sometimes, you get exactly what you ask for. And at those times, you simply have to take it.