University of Chicago Law School

Chicago Unbound

Journal Articles

Faculty Scholarship

2009

A New Systematic Explanation of the Types and Mitigating Effects of Exculpatory Defenses

Jim Staihar

Follow this and additional works at: https://chicagounbound.uchicago.edu/journal_articles

Part of the [Law Commons](#)

Recommended Citation

Jim Staihar, "A New Systematic Explanation of the Types and Mitigating Effects of Exculpatory Defenses," 12 New Criminal Law Review 205 (2009).

This Article is brought to you for free and open access by the Faculty Scholarship at Chicago Unbound. It has been accepted for inclusion in Journal Articles by an authorized administrator of Chicago Unbound. For more information, please contact unbound@law.uchicago.edu.

A NEW SYSTEMATIC EXPLANATION OF THE TYPES AND MITIGATING EFFECTS OF EXCULPATORY DEFENSES

Jim Staihar*

When someone performs a criminal act, there is a rebuttable presumption that she is particularly blameworthy and liable to a particularly severe punishment for the act. The presumption is rebutted when the criminal actor has an exculpatory defense. Such defenses mitigate how much a criminal actor is blameworthy and liable to be punished for her act. In this paper, I begin by spelling out a taxonomy of the main types of exculpatory defenses. Then I argue that a restorative signaling theory of punitive desert best explains why such defenses have their mitigating effects. According to the theory, how much someone is blameworthy and deserves to be punished for performing a criminal act corresponds to the severity of the burdens she is obligated to undertake to restore the conditions of trust she undermined by performing the act. The theory explains the mitigating effects of exculpatory defenses by explaining why they mitigate the severity of burdens that a criminal actor must undertake to fulfill the obligation of restoration she incurs from performing her act.

I. INTRODUCTION

A crime is a criminal act performed with a criminal state of mind. A crime's *actus reus* requirements define a criminal act, and its *mens rea* requirements define a criminal state of mind. A criminal act is a type of act

*Law and Philosophy Fellow, University of Chicago Law School. For helpful comments or discussion on these issues, I very much thank Elizabeth Anderson, Stephen Darwall, Anthony Dillof, Allan Gibbard, Thomas Green, Tom Beauchamp, Aaron Bronfman, Sarah Buss, Wayne Davis, Ed Curley, David Dick, Lee Fennell, Suzanne Goraj, Steven Kuhn, Saul Levmore, Eric Lormand, James Mattingly, Richard McAdams, Martha Nussbaum,

New Criminal Law Review, Vol. 12, Number 2, pps 205–286. ISSN 1933-4192, electronic ISSN 1933-4206. © 2009 by the Regents of the University of California. All rights reserved. Please direct all requests for permission to photocopy or reproduce article content through the University of California Press's Rights and Permissions website, <http://www.ucpressjournals.com/reprintInfo.asp>. DOI: 10.1525/nclr.2009.12.2.205.

that standardly poses a high risk of violating the rights of others. A criminal state of mind consists in disrespecting the rights of others.¹ Such disrespect consists in flouting the moral reasons against violating the rights of others, and flouting such reasons involves responding inappropriately to them.² Although a crime is defined as a criminal act performed with a criminal state of mind, the mere fact that someone performs a criminal act generates a rebuttable presumption that she³ is particularly blameworthy and liable to a particularly severe punishment for the act. Assuming she really is so blameworthy, the state would be warranted in blaming her to that degree. And if she really is liable to such a severe punishment, then the state has most reason to punish her so severely against her will.⁴ This presumption is rebutted, though, when the criminal actor has an exculpatory defense.⁵ Defenses mitigate how much a criminal actor is blameworthy and

Matt Pugsley, Peter Railton, Don Regan, Scott Shapiro, Matty Silverstein, and two anonymous reviewers for the New Criminal Law Review. For financial support while completing the paper, I thank the University of Michigan Philosophy Department and Rackham Graduate School, the John M. Olin Center for Law and Economics at the University of Michigan Law School, and the University of Chicago Law School.

1. See Stephen L. Darwall, Two Kinds of Respect, 88 Ethics 36, 40–41 (1977) (describing the concept of “moral recognition respect” whose denial captures the sense of disrespect at issue); Benjamin B. Sendor, Mistakes of Fact: A Study in the Structure of Criminal Conduct, 25 Wake Forest L. Rev. 707, 720–36 (1990) (suggesting that disrespect for the rights of others is standardly constitutive of a crime’s *mens rea* requirements). Throughout the paper, I focus only on crimes that are *mala in se*. For discussion on the types of defenses the state should recognize for crimes that are *mala prohibita* or strict liability offenses, see Jeremy Horder, Excusing Crime 237–76 (2004) (endorsing a due diligence defense and a reasonable ignorance of law excuse to such offenses).

2. At one extreme, such an inappropriate response might consist in someone’s intending to do something under conditions in which she believes that doing it would have a high risk of violating the rights of others, and none of what she believes about the consequences of doing it provides her with any reason to do it.

3. I intend to use personal pronouns, such as “he” or “she,” in a gender-neutral sense throughout the paper.

4. My sense of liability is a demanding one. On this sense, a criminal actor is liable to a punishment if and only if the state’s reasons to impose it on her against her will outweigh the state’s reasons not to. On a less demanding sense, a criminal actor is liable to a punishment if and only if the state’s imposing it on her against her will would not violate her rights. Cf. Jeff McMahan, The Basis of Moral Liability to Defensive Killing, 15 Phil. Issues 386, 386 (2005) (describing a similarly less demanding sense of liability to harmful treatment).

5. As I use the term, a criminal actor is not necessarily a criminal. A criminal actor is defined only as someone who performs a criminal act. A criminal is someone who performs a criminal act with a criminal state of mind.

liable to be punished.⁶ If a criminal actor has a full defense for her act, she is not at all blameworthy and is not liable to any punishment for it. If she has a partial defense for the act, she might be somewhat blameworthy and liable to some punishment for it, but not as much as presumed.

Although defenses have these mitigating effects, it is not obvious what qualifies as a defense or why defenses have these effects. In this paper, I have three objectives. First, I spell out a taxonomy of the main types of defenses. Second, I explain the mitigating effects of defenses on a criminal actor's liability to punishment. Third, I explain their mitigating effects on a criminal actor's blameworthiness. To explain their mitigating effects on how much a criminal actor is liable to be punished, I initially consider explanations grounded in the value requirement of my general theory of the justification of punishment. I argue none succeeds. Next I consider explanations grounded in the desert requirement of my general theory. I argue that my restorative signaling theory of punitive desert, RS, best explains them. I then argue that RS also plausibly explains the mitigating effects of defenses on how much a criminal actor is blameworthy. Given its explanatory power, I argue that RS has important implications not only for what should qualify as a defense but also for the nature and strength of the state's reasons to recognize defenses as mitigating factors in its punitive policies.

II. PRESUPPOSITIONS

In spelling out the main types of defenses and explaining their mitigating effects, I make several presuppositions that I assume obtain at the time a criminal actor is assessed for blameworthiness and liability to punishment. First, a criminal actor is unavoidably a member of a large community of persons over which the state governs as their agent. Second, any justifiable means of reducing the degree of interaction between a criminal actor and others would unavoidably leave a significant degree of interaction between them and so would unavoidably leave others significantly vulnerable to

6. Throughout the paper, I use the term "defenses" to refer only to exculpatory defenses. To be concise, I leave the exculpatory qualification implicit. I will not explain why the state should recognize non-exculpatory defenses, such as diplomatic immunity. For discussion of non-exculpatory defenses, see Paul H. Robinson, *Criminal Law Defenses: A Systematic Analysis*, 82 Colum. L. Rev. 199, 229–32 (1982).

her. Third, there are no extraordinary means of obtaining epistemic access to a criminal actor's dispositions. So I assume there are no extraordinary means of obtaining epistemic access to whether she has a particular disposition to commit crimes.

I make presuppositions 1 through 3 because I seek a practical explanation of the mitigating effects of defenses. I seek to explain their mitigating effects in the actual world given the natural facts that generally characterize the unavoidable conditions under which people actually live. Presuppositions 1 through 3 are such facts. I leave it an open question whether the considerations that constitute defenses in the actual world would also be defenses with the same mitigating effects in other worlds in which those presuppositions do not obtain.⁷

Fourth, everyone knows what acts the criminal actor has performed and not performed, and they know the beliefs, intentions, and motives with which she performed them. Fifth, everyone knows all the facts about her physical and psychological capacities. Thus, I assume everyone knows whether she has the capacity to respond appropriately to the moral reasons against violating the rights of others. Sixth, everyone forms justified beliefs about her dispositions on the basis of his knowledge specified in presuppositions 4 and 5. So if everyone's knowledge of the relevant facts justifies him in believing that she has a particularly bad disposition to commit crimes, I assume he justifiably believes that she has such a disposition. Seventh, everyone responds rationally to his justified beliefs about her dispositions. So if others are justified in believing that she has a particularly bad disposition to commit crimes, and they are rationally required to incur certain costs in response, I assume they incur such costs.

I make presuppositions 4 through 6 because I seek to explain the mitigating effects of defenses on the assumption that others fulfill their epistemic duties to each other and to the criminal actor before they punish her. Before they do so, they have a duty to discover that no facts about her acts or capacities entail that she does not deserve the punishment. They also have a duty to form justified beliefs about her dispositions to commit crimes. And when others discover facts that make someone deserving of punishment, they have a duty to promulgate such facts to those with an

7. Cf. John Rawls, *A Theory of Justice* 157–61 (1971) (developing a theory of distributive justice with a similarly practical aim and its own presuppositions about the natural facts under which it applies).

interest in them but are unaware of them. Presuppositions 4 through 6 entail that others fulfill such duties.

Also I make presuppositions 4 through 7 because I seek to explain the mitigating effects of defenses on how much a criminal actor is blameworthy and deserves to be punished. Both concepts are plausibly defined conditionally on the assumption that presuppositions 4 through 7 obtain at the time of assessment. Regarding the concept of blameworthiness, someone is particularly blameworthy for an act if and only if she did it, and others would be warranted in blaming her to this degree for doing it if presuppositions 4 through 7 were to obtain. Regarding the concept of punitive desert, someone deserves a punishment for an act if and only if she did it, and the state would not violate her rights by imposing the punishment on her against her will for doing it if presuppositions 4 through 7 were to obtain.⁸ In either case, how much someone is blameworthy or deserves to be punished for doing something does not seem to depend essentially on the moral status of blaming or punishing her under less idealized conditions. For example, whether someone is blameworthy for doing something does not seem to depend essentially on whether others would be warranted in blaming her for it under conditions in which (a) they do not know what she has done or is capable of doing; (b) they have unjustified beliefs about her dispositions; or (c) they are disposed to respond irrationally to their beliefs about her dispositions.⁹ Similarly, these less idealized conditions also seem irrelevant to the issue of whether

8. My non-retributive sense of punitive desert is an instance of J.L.A. Garcia's more general conception of "negative desert" in *Two Concepts of Desert*, 5 *Law & Phil.* 219 (1986). According to Garcia, the point of claiming that someone deserves something harmful to her is not to claim that there would be anything good about imposing it on her. There might not be. According to Garcia, the point is merely to claim that imposing it on her against her will would not violate her rights. On a retributive sense of punitive desert, though, the point of claiming that someone deserves to be punished is precisely to claim that punishing her would be intrinsically good. See Michael S. Moore, *Justifying Retributivism*, 27 *Isr. L. Rev.* 15, 19–20 (1993); cf. G.E. Moore, *Principia Ethica* 82, 262–65 (Thomas Baldwin ed., rev. ed. 1993) (claiming that the organic whole of a criminal's suffering some punishment is better than the organic whole of her suffering no punishment). Because I eschew the retributive sense, I am not committed to retributivism by claiming that criminals deserve to be punished.

9. Cf. Justin D'Arms & Daniel Jacobson, *Sentiment and Value*, 110 *Ethics* 722, 745 (2000) (noting that whether someone is warranted in feeling a particular emotion, such as blame, toward something depends on what she has evidence for believing about it).

someone deserves to be punished for doing something. So to explain the mitigating effects of defenses on how much a criminal actor is blameworthy and deserves to be punished, we should assume presuppositions 4 through 7 obtain at the time of assessment.

III. A TAXONOMY OF DEFENSES

To explain the mitigating effects of defenses, we should understand the main types of them. We can usefully begin by describing three general types along with justifications.

A. Type 1 Defenses

When someone performs a criminal act, there is a presumption that in performing it, she manifested a particularly bad degree of disrespect toward the rights of others. The presumed degree of disrespect corresponds to how badly she presumably flouted the moral reasons against violating their rights. How badly she presumably flouted such reasons corresponds to how inappropriately she presumably responded to them. In general, how inappropriately she presumably responded to such reasons depends on the beliefs, intentions, and motives with which she presumably performed the act. More precisely, how inappropriately she presumably responded to them depends on (a) what she presumably intended in performing the act, (b) what presumably motivated her to intend to perform the act, (c) how strongly she was presumably motivated to intend to perform the act, and (d) the relative strength of the reasons that she presumably had for and against intending to perform the act. What reasons she presumably had for and against intending to perform the act depends on what she presumably believed when she performed it.

Thus, a reason is meant here in a subjective sense that is relative to what the criminal actor believed when she performed the act. In this sense, the reasons she had for and against intending to perform the act were provided by what she believed when she performed it.¹⁰ So in this sense, the

10. In an objective sense, someone's reasons for and against intending to perform an act are not relative to what she believes. They are provided by facts that are true independently of what she believes. See Derek Parfit, *Rationality and Reasons*, in *Exploring Practical*

fact that someone believes that her performing a particular act would directly cause another to die against his will provides her with a strong moral reason against intending to do it. This fact provides her with such a reason even if her belief is false, or she believes that it does not provide her with such a reason. Hence, if she were to intend to perform the act without believing it would have any consequences that would count in favor of it, she would flout very badly the moral reasons against violating the putative victim's right not to be killed. Therefore, she would disrespect very badly his right not to be killed.

Given the degree of disrespect that someone presumably manifested in performing a criminal act, type 1 defenses are considerations that mitigate how badly she actually disrespected the rights of others in performing it.¹¹ In other words, they mitigate how badly she actually flouted the moral reasons against violating the rights of others in performing the act. They mitigate how inappropriately she actually responded to such reasons in performing it. If a criminal actor has a full type 1 defense, she did not disrespect the rights of others at all in performing her act.¹² If she has a partial type 1 defense, she might have disrespected the rights of others to some degree, but she did not disrespect them as badly as presumed.

Philosophy: From Action to Values 17, 17 (Dan Egonsson, Jonas Josefsson, Bjorn Petersson, & Toni Rønnow-Rasmussen eds., 2001) (noting the objective sense of reasons). Although they are conceptually distinct, subjective and objective reasons are related. A fact provides someone with an objective reason if and only if her believing that it obtains would provide her with a subjective reason.

To keep matters simple, I set aside for further analysis the issue of whether someone's beliefs must be justified for them to provide her with a subjective reason for or against intending to perform an act. Similarly, I set aside for further analysis the issue of whether the subjective reasons that someone had for and against intending to perform an act were provided not by what she believed when she performed the act, but more precisely, by the considerations that provided her with evidence to believe or not to believe certain propositions when she performed the act.

11. See T.M. Scanlon, *What We Owe to Each Other* 279–80 (1998) (noting this type of defense); P.F. Strawson, *Freedom and Resentment*, in *Freedom and Resentment and Other Essays* 7–8 (1974) (same); Gary Watson, *Responsibility and the Limits of Evil: Variations on a Strawsonian Theme*, in *Perspectives on Moral Responsibility* 119, 123 (John Martin Fischer & Mark Ravizza eds., 1993) (same); cf. John Gardner, *The Gist of Excuses*, 1 *Buff. Crim. L. Rev.* 575 (1998) (endorsing excuses which entail that in performing her act, the criminal actor satisfied the standards of character governing her role at the time of the act).

12. More precisely, if a criminal actor has a full type 1 defense, she did not disrespect the rights of others to any significant degree in performing her act.

Many different considerations can provide a criminal actor with a type 1 defense. For example, she can have a type 1 defense of compulsion because she was caused to perform the act against her will by mere force.¹³ The compulsion might have been external or internal. To illustrate a case of external compulsion, suppose a group seizes my arm and although I try to resist, they use my arm against my will to strike another person. To illustrate a case of internal compulsion, suppose I suffer a seizure that also causes my arm to strike the victim against my will. In both cases, I have a type 1 defense of compulsion for striking the victim. In neither case did I intend to do anything that I believed had a high risk of striking him or harming him in any way. In the former, I intended not to strike him or harm him in any way, and in the latter, not only did I lack the capacity to form any intentions at all, I also had no beliefs about what I was doing. So in striking the victim, I did not respond inappropriately to the moral reasons against violating his rights, specifically his right not to be assaulted. Thus, I did not disrespect his rights in striking him.

For another example, a criminal actor can have a type 1 defense of ignorance or mistake of fact because she did not believe her act would have the type of consequences that make it a criminal act.¹⁴ In other words, she did not believe her act would have the type of consequences that standardly pose a high risk of violating the rights of others. To illustrate, suppose I am target shooting with a friend. I shoot at the target believing there is no chance the bullet will harm anyone in any way. However, the bullet ricochets off the target and into my friend, causing him to die. In this case, I have a type 1 defense of ignorance or mistake of fact for killing him. Although I intended to do something that subsequently caused him to die, I did not believe that doing it would cause him to die or harm him

13. See Aristotle, *Nicomachean Ethics* 1110a1–5 (Terence Irwin trans., 2d ed. 1999) (describing defenses of compulsion). Strictly speaking, an act is intended, under some description. See Donald Davidson, *Agency*, in *Essays on Actions and Events* 43, 46 (1980); Michael E. Bratman, *Intention, Plans, and Practical Reason* 119–22 (1987). However, when someone is caused to perform a criminal act against her will by mere force, she does not intend to do it, under any description. So as I use the term, criminal acts are, strictly speaking, merely putative acts.

14. See Aristotle, *supra* note 13, at 1110b19–1111a21 (describing defenses of ignorance or mistake of fact). To keep matters simple, I set aside for further analysis the issue of whether the mistake or ignorance must be justified or reasonable to constitute a defense. See *supra* note 10. For discussion on this issue, see Victor Tadros, *Criminal Responsibility* 237–64 (2005).

in any way. So in killing my friend, I did not respond inappropriately to the moral reasons against violating his rights, specifically his right not to be killed. Hence, I did not disrespect his rights in killing him.

For another example, a criminal actor can have a type 1 defense of countervailing considerations because, although she knew the act was a criminal one, she was appropriately motivated to perform it by her belief that particular considerations obtained which (a) entailed that her performing it would not violate the rights of others or (b) provided her with most reason to perform it.¹⁵ Such an example might involve consent, self-defense, defense of others, duress, or necessity. To illustrate a case of defense of others, suppose I know that an assailant will wrongfully kill a group of innocent people unless I kill him first. As a consequence, he lacks a right against my killing him. In response, I kill him for the purpose of preventing him from killing the others. In this case, I have a type 1 defense of defense of others for killing him. Although I intended to kill him, I was appropriately motivated to do so by my belief that some consideration obtained which entailed that he lacked a right against my killing him. So in killing him, I did not disrespect his rights.

To illustrate a case of necessity, suppose I know that I will starve to death unless I steal some food from the surplus of another. As a consequence, I have most reason to steal the food even though I know that by doing so I would violate the other's property right to the food.¹⁶ In response, I steal it for the purpose saving my life. In this case, I have a type 1 defense of necessity for stealing the food. Although I intended to steal it, I was appropriately motivated to do so by my belief that some consideration obtained which provided me with most reason to steal it. So in stealing the food, I did not respond inappropriately to the moral reasons against violating the other's property right to it, even though I knew I violated this right by stealing it. For in this case, those reasons were outweighed by the reasons in favor of violating his property right to the

15. See Aristotle, *supra* note 13, at 1110a5–1110a22 (describing defenses of countervailing considerations). Again, to keep matters simple, I set aside for further analysis the issue of whether the belief must be justified for it to provide the criminal actor with a defense. See *supra* notes 10, 14.

16. Although I assume stealing the food in this context would violate the other's property right to it, there is controversy over whether his property right to the food would extend to such a case. I assume it would, in part, because by taking the food I presumably incur an obligation to compensate him for the loss.

food, namely the reasons for saving my life. Therefore, I did not disrespect his property right to the food in stealing it.

To clarify, type 1 defenses of countervailing considerations have a motivation requirement: countervailing considerations constitute a type 1 defense only if the criminal actor was appropriately motivated by them in performing her act.¹⁷ Thus, the mere fact that a criminal actor believed such considerations obtained when performing her act does not provide her with a type 1 defense for the act. Such considerations provide her with a type 1 defense only if she not only believed they obtained, but also was appropriately motivated by this belief in performing the act, such that if she had not believed they obtained, then she would not have been motivated as strongly to perform it.¹⁸

To illustrate the motivation requirement, consider an apparent case of defense of others. Suppose a misanthrope embarks on a killing spree merely because she enjoys killing others and has no regard for their rights. On her spree, she sees someone aim a gun at five innocent persons. She knows he will wrongfully kill the five unless she kills him first with her own gun. She then kills him not because she believes that doing so is necessary to prevent him from killing the five, but merely because she would enjoy killing him and has no regard for his rights. Her belief that killing him was necessary to prevent him from killing the five did not play any role in motivating her to kill him. She would have been motivated to kill him just as strongly even if she did not have this belief. As evidence that she was not motivated by this belief, we might suppose that she tried to kill the other five immediately after killing the one. In this case, she does not have a type 1 defense of defense of others for killing him. Although she

17. See John Gardner, *Justifications and Reasons*, in *Harm and Culpability* 103, 105 (A.P. Simester & A.T.H. Smith eds., 1996) (endorsing a motivation requirement on defenses of countervailing considerations); Tadros, *supra* note 14, at 273–80 (same); Marcia Baron, *Justifications and Excuses*, 2 *Ohio St. J. Crim. L.* 387, 392–93 (2005) (same).

18. The motivation requirement on type 1 defenses of countervailing considerations is analogous to a requirement on justified or well-founded beliefs in epistemology: someone's belief is justified or well-founded only if she not only has evidence that supports her belief, but also holds the belief on the basis of her evidence that supports it. See Earl Conee & Richard Feldman, *Evidentialism: Essays in Epistemology* 92–93 (2004); cf. Hilary Kornblith, *Beyond Foundationalism and the Coherence Theory*, 77 *J. Phil.* 597, 601–02 (1980); Alvin I. Goldman, *What Is Justified Belief?*, in *Justification and Knowledge* 1, 8–9 (George S. Pappas ed., 1979).

believed that killing him was necessary to prevent him from killing the five, she was not motivated by this belief in killing him; she was motivated only by her desire for pleasure. So she disrespected his right not to be killed merely to promote her own pleasure.

Although a criminal actor must have been appropriately motivated by countervailing considerations in order for them to provide her with a type 1 defense, her motivation need not have been perfectly virtuous. The motivation requirement requires only an appropriate regard for the rights of others. To illustrate, consider again the misanthrope, except suppose she is appropriately motivated by the rights of others. She still enjoys killing others; however, she is motivated to do so only under conditions in which doing so would not violate their rights. So suppose again that the misanthrope kills the one under conditions in which she knows that he would kill the other five unless she kills him first. In this case, though, she kills him not merely because she enjoys doing so, but also because she believes that doing so would not violate his rights since killing him is necessary to prevent him from killing the others. In this case, she has a full type 1 defense of defense of others for killing him because she was appropriately motivated by her belief that killing him was necessary to prevent him from killing the others, such that if she had not held this belief, then she would not have been motivated to kill him. She has such a defense for killing him even though her motivation was not perfectly virtuous given that she was motivated in part to kill him to promote her own pleasure. However, she did not kill him merely to promote her own pleasure: her motivation was constrained by an appropriate regard for his rights. Because she appropriately regarded his rights in killing him, she did not disrespect his rights in doing so.

Critics might argue that type 1 defenses of countervailing considerations do not have a motivation requirement.¹⁹ They might contend that a criminal actor's having any regard for the rights of others is not a necessary condition of her having such a defense. According to them, the mere fact that the criminal actor believed certain countervailing considerations obtained when she performed her act can be sufficient to provide her with a defense for the act no matter what her motivations for performing it. For in such a case, her act might have been optimific in both a subjective and

19. I refer in general to "critics" as a means of anticipating possible objections to my claims. I do not assume actual writers have raised the objections I discuss.

objective sense no matter what her motivations for performing it. It might have been subjectively optimific in the sense that its expected value might have been at least as high as the expected value of any other act then available to her. It might have been objectively optimific in the sense that its actual consequences might have been at least as good as the consequences of any other acts then available to her.²⁰ Assuming her act was optimific in both senses, she has a defense for the act no matter what her motivations for performing it. For if others were to express any blame toward her for the act, they would deter her and others from performing similar optimific acts in the future. Hence, expressing any blame toward her for the act would not itself be optimific. So she would not be blameworthy for the act.

To illustrate, consider again the first case of the misanthrope who has no regard for the rights of others. Although her trying to kill the innocent five was not optimific, her killing the one was optimific because she knew that killing him was necessary to prevent him from killing the others. Because her killing him was optimific, she has a type 1 defense of defense of others for doing so even though she did so with no regard for his rights, as she killed him merely to promote her own pleasure. If others were to express any blame toward her for killing him, they would deter her and others from performing similar optimific acts in the future. For example, they would deter misanthropes generally from killing aggressors under conditions in which doing so is necessary to prevent the aggressors from killing many others. Hence, expressing any blame toward the misanthrope for killing the victim would not itself be optimific. So she is not blameworthy for killing him.

In response, the critics have not proven that type 1 defenses of countervailing considerations do not have a motivation requirement. Their argument is unpersuasive for at least three reasons. First, the critics assume that a criminal act might be optimific even if it is performed with no regard for the rights of others. But in the standard case, a criminal act would not be optimific if it were performed with no such regard. Whether an act is optimific depends on the value or expected value of its consequences. The value or expected value of an act's consequences depends on not only the value of its further causal consequences, but also the value of the act itself and the

20. Unless I note otherwise, I use the term "optimific" to connote both its subjective and objective senses.

motives with which it is performed.²¹ If a criminal actor performs her act with no regard for the rights of others, then she performs it with a morally deficient motivation that is worse than the motivation with which she would have performed it if she had performed it with the appropriate regard for the rights of others. So in the standard case, a criminal act performed with no regard for the rights of others would not be optimific because the value or expected value of its consequences would have been higher if it had been performed with the appropriate regard for the rights of others.

Second, if a criminal actor believes that certain countervailing considerations obtain when performing her act, the critics assume that expressing blame toward her for the act would not be optimific even if she performed it with no regard for the rights of others. They assume that expressing such blame would deter her and others from performing other acts in the future that would be optimific at least considered independently of the motives with which they would be performed. But even if expressing such blame would deter others from performing some optimific acts in the future, the expression might still be optimific. For the expression might have some sufficiently valuable consequences that would not obtain in its absence. For example, the expression might deter the performance of not only some optimific criminal acts, but also some non-optimific criminal acts. For another example, the expression might encourage some to develop an appropriate regard for the rights of others. As a consequence, the expression might result in not only fewer optimific criminal acts, but also fewer non-optimific criminal acts. Assuming the non-optimific acts prevented are sufficiently bad and numerous relative to the optimific acts deterred, the expression might be optimific.

Third, the critics assume that if expressing blame toward a criminal actor for her act would not be optimific, then she is not blameworthy for the act. But whether the expression would be optimific is not relevant to whether she is blameworthy. She is blameworthy for the act if and only if others would be warranted in blaming her for the act under certain idealized conditions.²² And others' blaming her for the act would consist in

21. See G.E. Moore, *supra* note 8, at 198–99; William Shaw, *The Consequentialist Perspective*, in *Contemporary Debates in Moral Theory* 5, 6–7 (James Dreier ed., 2006); Tadros, *supra* note 14, at 278–79.

22. The idealized conditions plausibly include the satisfaction of my presuppositions 4 through 7. See *supra* text accompanying notes 7–9.

their feeling an attitude of moral blame toward her for the act. Such an attitude might consist in resentment or indignation.²³ So she is blameworthy for the act if and only if others would be warranted in feeling an attitude of moral blame toward her for the act under certain idealized conditions. But others' feeling an attitude of moral blame toward her does not necessarily involve their expressing the attitude. They might feel such an attitude without expressing it. Thus, the issue of whether they would be warranted in feeling such an attitude is distinct from the issue of whether they would be warranted in expressing it. They might be warranted in feeling such an attitude even if they are not warranted in expressing it.²⁴ As a consequence, even if others would not be warranted in expressing an attitude of moral blame toward a criminal actor for her act because the expression would not be optimific, the criminal actor might still be blameworthy for the act.

The critics might concede that someone is blameworthy if and only if others would be warranted in feeling an attitude of moral blame toward her. But they might contend that others are warranted in feeling an attitude of moral blame toward someone if and only if feeling such an attitude toward her would be optimific. And they might contend that others would feel an attitude of moral blame toward someone only if they would also express it toward her. For they might contend that it is psychologically impossible to feel an attitude of moral blame toward someone without expressing it toward her.²⁵ As a consequence, feeling an attitude of moral blame toward someone would be optimific only if expressing the attitude toward her would be optimific. So feeling an attitude of moral blame toward someone would be warranted only if expressing the attitude toward her would be optimific. Thus, assuming it would not be optimific to express an attitude of moral blame toward someone, others would not be warranted in feeling such an attitude toward her, and so she would not be blameworthy.

The critics' response is unpersuasive for at least two reasons. First, we can plausibly deny that it is psychologically impossible to feel an attitude

23. See, e.g., Strawson, *supra* note 11, at 13–15.

24. See Scanlon, *supra* note 11, at 269.

25. Cf. Justin D'Arms & Daniel Jacobson, *The Moralistic Fallacy: On the "Appropriateness" of Emotions*, 61 *Phil. & Phenomenological Res.* 65, 77 (2000) (stating that "[i]n normal human psychology . . . the relationship between feeling an emotion and expressing it . . . is exceedingly tight").

of moral blame without expressing it. We can concede that others generally have a motivational tendency to express their attitudes of moral blame. So we can concede that it might be difficult for others to restrain themselves from expressing such attitudes. But this does not entail that it is psychologically impossible to do so: it entails only that a concerted effort is required to do so. Assuming others can feel an attitude of moral blame without expressing the attitude, it might be optimific for them to feel the attitude even if it would not be optimific for them to express the attitude.

Second, suppose for the sake of argument that an attitude of moral blame is not optimific. More generally, suppose for the sake of argument that it would not be optimific to feel an attitude of moral blame toward a criminal actor who performed her act with no regard for the rights of others but nevertheless believed certain countervailing considerations obtained when she performed it. Contrary to the critics, the attitude might still be warranted, and so the criminal actor might still be blameworthy for the act. For many of the considerations that bear on whether an attitude of moral blame is optimific are irrelevant to whether the attitude is warranted. Whether the attitude is optimific depends on the value of all the consequences of feeling the attitude. But whether the attitude is warranted depends on a distinct and narrower set of considerations.²⁶ In general, whether an attitude of moral blame toward someone for performing an act is warranted depends only on three considerations. First, it depends on what the assessor is justified in believing about the external and internal facts about the act itself, including the external and psychological causes of the act and its consequences. Second, it depends on what the assessor is justified in believing about the actor's capacities. Third, it depends on the content of the attitude itself, and what the assessor is justified in believing about her relation to the actor.²⁷ Thus, the consequences of having or not

26. For discussion on the problem of identifying the considerations that are relevant to whether an attitude of moral blame is warranted or, in other words, fitting or rational, see, e.g., D'Arms & Jacobson, *supra* note 9; D'Arms & Jacobson, *supra* note 25; Stephen Darwall, *The Second-Person Standpoint: Morality, Respect, and Accountability* 15–17 (2006); Allan Gibbard, *Wise Choices, Apt Feelings* 36–40 (1990); Parfit, *supra* note 10; Wlodek Rabinowicz & Toni Rønnow-Rasmussen, *The Strike of the Demon: On Fitting Pro-Attitudes and Value*, 114 *Ethics* 397 (2004).

27. For example, suppose a criminal actor performs an act in which she disrespects the rights of a particular victim. Other things being equal, only the criminal actor is warranted in feeling guilty toward herself for the act, given that guilt is a self-reactive attitude. Only

having the attitude are irrelevant to whether it is warranted. Because the value of such consequences does bear on whether the attitude is optimific, the attitude might be warranted even if it is not optimific, and vice versa.

To illustrate the first possibility, suppose a criminal actor assaults me without any defense. She then threatens to assault me again if I feel any resentment toward her for assaulting me. We might suppose that if I were to resent her, I would evince some sign of the resentment such that she would know I were feeling it, and she would assault me as a consequence. Given the threat, we might suppose that my resenting her for assaulting me would not be optimific. Being assaulted again would make the consequences of my feeling the resentment all things considered worse than the consequences of my not feeling it. So assuming the resentment would not be optimific, I might be warranted in desiring not to feel it and in undertaking a process of getting myself not to feel it.²⁸ But even if the resentment would not be optimific, it might still be warranted. For the consequences of my resenting her for assaulting me are irrelevant to whether the resentment would be warranted. Because she disrespected my rights in assaulting me, I am presumably warranted in resenting her for assaulting me. So she might be blameworthy for assaulting me even if blaming her for it would not be optimific.

To illustrate the second possibility, suppose someone performs a supererogatory act of donating to charity. A criminal actor then threatens to assault me and the donor if I do not feel an attitude of indignation toward the donor for giving to charity. Given the threat, we might suppose that my indignation toward the donor would be optimific. My and the donor's being assaulted would make the consequences of my not feeling the indignation all things considered worse than the consequences of my feeling it. So assuming the indignation would be optimific, I might be warranted in desiring to feel the indignation and in undertaking a process of getting myself to feel it. But even if the indignation would be optimific,

the victim is warranted in feeling resentment toward the criminal actor for the act, given that resentment is a personal reactive attitude. And anyone is warranted in feeling indignation toward the criminal actor for the act, given that indignation is an impersonal reactive attitude. See Strawson, *supra* note 11, 13–15 (discussing the types of reactive attitudes and their relations to each other).

28. See Gibbard, *supra* note 26, at 37 (distinguishing the issue of whether an attitude is rational from the issue of whether desiring the attitude is rational, and noting that a person can rationally desire not to have a rational attitude); Parfit, *supra* note 10, at 27 (same).

it would not be warranted. For the consequences of my not feeling the indignation are irrelevant to whether it would be warranted. Because the donor did not disrespect anyone's rights in giving to charity, I would not be warranted in feeling indignation toward her for giving to charity. So the donor would not be blameworthy for giving to charity even if blaming her for it would be optimific.

In summary, the critics have not proven that type 1 defenses of countervailing considerations do not have a motivation requirement. On the contrary, they do. If a criminal actor merely believed that countervailing considerations obtained when she performed her act but was not appropriately motivated by them, then she disrespected the rights of others in performing the act. Other things being equal, someone's disrespecting the rights of others in performing an act makes her blameworthy for the act even if blaming her for it would not be optimific.

Although type 1 defenses of countervailing considerations have a motivation requirement, a caveat is in order. In practice, it is standardly very difficult to determine whether the required motivation obtained. To avoid wrongly denying a criminal actor a defense, the state should recognize a strong presumption that a criminal actor had the required motivation if she believed countervailing considerations obtained at the time of her act. So the state should recognize a strong presumption that the criminal actor did not disrespect the rights of others if she believed countervailing considerations obtained at the time of her act. The presumption should be overridden only when the state meets a high burden of proof in demonstrating that the presumption is not satisfied. As a consequence, the motivation requirement might not have significant bite in practice.

B. Justifications

Justifications are a subset of type 1 defenses. They are distinct from other type 1 defenses in that they apply only to criminal acts that were intended under some description.²⁹ More precisely, justifications apply only to the specific intention that a criminal actor had to perform her act. To claim that a criminal actor has a justification for her act is elliptical for claiming

29. See *supra* note 13.

that she has a justification for the intention she had to perform it.³⁰ Like the presumptions generated by a criminal act itself, the intention of a criminal actor generates similar presumptions regarding how much she is blameworthy and liable to punishment for the intention. In particular, the intention of a criminal actor generates similar presumptions regarding how badly she disrespected the rights of others in having the intention. As type 1 defenses, justifications are considerations that mitigate how much a criminal actor is blameworthy and liable to punishment for intending to perform her act by mitigating how badly she actually disrespected the rights of others in intending to perform the act.

Justifications can be subjective or objective. A criminal actor has a subjective justification for her intention if and only if she has a type 1 defense for it. A criminal actor has an objective justification for her intention if and only if she has an infeasible type 1 defense for it. She has an infeasible type 1 defense for the intention if and only if she actually has this type 1 defense for the intention, and she would have had this defense for the intention even if she had the intention under conditions in which (a) she knew all the facts and (b) other things were equal. Objective justifications are a subset of subjective ones. Objective justifications are distinct from other subjective ones in that objective justifications are partly defined in terms of what the criminal actor would have believed if she knew all the facts, whereas subjective justifications are defined only in terms of what the criminal actor actually believed when intending to perform her act.

Both subjective and objective justifications can be full or partial. A criminal actor has a full (or partial) subjective justification for her intention if and only if she has a full (or partial) type 1 defense for it. A criminal actor has a full (or partial) objective justification for her intention if and only if she has a full (or partial) infeasible type 1 defense for it. And she has a full (or partial) infeasible type 1 defense for the intention if and only if she actually has a full (or partial) type 1 defense for the intention, and she would have had this full (or partial) type 1 defense for the intention even if she had the intention under conditions in which (a) she knew all the facts and (b) other things were equal.

30. I do not assume the intention a criminal actor had to perform her act was an intention to perform the act under its description as a criminal act. I merely assume it was an intention to do something whose actual consequences made it a criminal act.

To illustrate the distinction between justifications and other type 1 defenses, consider again the cases of external and internal compulsion. In both cases, I strike another person. In the one, I do so merely because a group seizes my arm and uses it against my will to strike the victim. In the other, I do so merely because I suffer a seizure that causes my arm to strike him. In both cases, I have a type 1 defense of compulsion for striking the victim because I did not disrespect the rights of others in doing so. However, in neither case do I have a justification. For in neither case do I perform the act intentionally. So in neither case do I intend to do anything for which I could have a justification. Thus, justifications do not include type 1 defenses of compulsion although they do include type 1 defenses of ignorance or mistake of fact and countervailing considerations.

To illustrate the distinction between objective and mere subjective justifications, consider two cases of self-defense. In the first, suppose I see two persons load guns and aim them at me as if they are going to shoot me to death. In response, I believe they will kill me unless I prevent them from shooting. I also believe I would prevent them from shooting me by shooting them to death first with my own gun. There are no other ways I believe I could prevent their killing me. To prevent their killing me, I intend to kill them, which I then do. Assuming my beliefs were true, and there was no less harmful way that I could have prevented their killing me, I have both a subjective and an objective justification of self-defense for intending to kill them. For given my actual beliefs and motive, I actually have a type 1 defense of self-defense for the intention. And this defense is indefeasible because I would have had this defense for intending to kill them even if I had intended to kill them under conditions in which (a) I knew all the facts and (b) all other things were equal.

The second case is the same as the first except, unknown to me, the two persons load their guns only with blanks and so are not an objective threat to me. In this case, I still have a subjective justification of self-defense for intending to kill them. For given my actual beliefs and motive, I still actually have a type 1 defense of self-defense for the intention. But I no longer have an objective justification of self-defense for the intention because my defense of self-defense is not indefeasible: I would not have had this defense for intending to kill them if I had intended to kill them under conditions in which (a) I knew all the facts and (b) all other things were equal. For if I had known all the facts, I would have known they

posed no threat to me. So by intending to kill them, I would have disrespected their right not to be killed.

Given my conception of justifications, we can see where it stands on seven potentially controversial issues about how we should understand them. First, the mere fact that a criminal actor has both no false beliefs and a type 1 defense for intending to perform her act does not provide her with an objective justification for her intention. To illustrate, consider a case of self-defense that is the same as the first except, unknown to me, I could have prevented the two from shooting me by merely uttering to them a religious slogan that would have identified me as a member of their religious group. It is not that I had the false belief that uttering the slogan would not have prevented them from shooting me. It is that I had no beliefs about this at all. Here I do have a type 1 defense of self-defense for intending to kill them, and I had no false beliefs. So I have a subjective justification for the intention. But I do not have an objective justification for the intention because my type 1 defense of self-defense for it is not indefeasible. If I had known all the facts, and other things were equal, then I would have known that I could have prevented their killing me merely by uttering the slogan. So I would have known that killing them was not the least harmful means to preventing them from killing me. Hence, I would not have had a type 1 defense of self-defense for intending to kill them.

Second, the mere fact that a criminal actor intends to perform her act on the basis of some false beliefs does not entail that she lacks an objective justification for her intention. To illustrate, consider a case of self-defense that is the same as the first except, unknown to me, if the two had shot, a gust of wind would have blown their bullets into a tree off of which the bullets would have ricocheted back into me, thereby killing me. In this case, I intend to kill the two on the basis of my false belief that if they had shot, their bullets would have flown directly into me. Other things being equal, if I had believed that a gust of wind would have blown their bullets into a tree, I would not have believed they were a threat to me and so would not have intended to kill them. Nevertheless, I still have an objective justification for intending to kill them. For given my actual beliefs and motive, I have a type 1 defense of self-defense for the intention. And this defense is indefeasible. If I had known all the facts, and other things were equal, then I would have known that my killing them was necessary to prevent their killing me. So I still would have had the type 1 defense of self-defense for intending to kill them.

Third, the mere fact that a criminal actor has an objective justification for intending to perform her act does not entail that (a) no one has an objective justification for intending to resist her act or (b) third parties could have an objective justification for intending to perform the act on her behalf.³¹ To illustrate, suppose I and another person consent to box each other but only each other on the condition that each of us boxes without the assistance of third parties. As a consequence, I have an objective justification for intending to strike him because I have an infeasible type 1 defense of consent for this intention. But he also has an objective justification for intending to resist my striking him because he also has an infeasible type 1 defense of consent for this intention. Furthermore, no third parties could have an objective justification for intending to strike my opponent on my behalf because no third parties could have an infeasible type 1 defense for this intention. In particular, no third parties could have an infeasible type 1 defense of consent for intending to strike him on my behalf given the restricted scope of his consent.

Fourth, the mere fact that a criminal actor has a full subjective justification for intending to perform her act does entail that her intention is subjectively right in a weak sense. Someone's intention is subjectively right in a weak sense if and only if she does not disrespect the rights of others by having the intention given her actual beliefs and motives. Similarly, the mere fact that a criminal actor has a full objective justification for intending to perform her act does entail that her intention is objectively right in a weak sense. Someone's intention is objectively right in a weak sense if and only if she would not disrespect the rights of others by having the intention and some motive for it under conditions in which (a) she knew all the facts and (b) other things were equal.

31. Compare George P. Fletcher, *Rethinking Criminal Law* 759–62 (1978) (endorsing such an entailment), with Mitchell N. Berman, *Justifications and Excuses, Law and Morality*, 53 *Duke L.J.* 1, 62–64 (2003) (rejecting such an entailment), Michael Corrado, *Notes on the Structure of a Theory of Excuses*, 82 *J. Crim. L. & Criminology* 465, 466–67, 491–93 (1992) (same), Joshua Dressler, *New Thoughts About the Concept of Justifications in the Criminal Law: A Critique of Fletcher's Thinking and Rethinking*, 32 *UCLA L. Rev.* 61, 87–98 (1984) (same), R.A. Duff, *Rethinking Justifications*, 39 *Tulsa L. Rev.* 829, 830 (2004) (same), Kent Greenawalt, *The Perplexing Borders of Justification and Excuse*, 84 *Colum. L. Rev.* 1897, 1918–27 (1984) (same), and Douglas N. Husak, *Conflicts of Justifications*, 18 *Law & Phil.* 18 (1999) (same).

However, the mere fact that a criminal actor has a full subjective justification for intending to perform her act does not entail that her intention is subjectively right in a strong sense. Someone's intention is subjectively right in a strong sense if and only if she has most reason to have the intention given her actual beliefs.³² Similarly, the mere fact that a criminal actor has a full objective justification for intending to perform her act does not entail that the intention is objectively right in a strong sense. Someone's intention is objectively right in a strong sense if and only if she would have most reason to have the intention under conditions in which (a) she knew all the facts and (b) other things were equal.³³

To illustrate, consider the previous boxing case. In it, I have a full subjective and objective justification for intending to strike my opponent. Because I know he consented to box me, I do not disrespect his rights by intending to strike him, and I would not do so even if I knew all the facts, and other things were equal. So my intention is both subjectively and objectively right in a weak sense. But the intention might be neither subjectively nor objectively right in a strong sense. For by striking him, I know I would cause him severe pain, and such pain might provide me with most reason not to intend to strike him even though I know striking him would not violate his rights.

Fifth, the mere fact that a criminal actor's intention is objectively right in a weak or strong sense does not entail that she has a subjective or objective justification for the intention.³⁴ To illustrate, suppose I killed someone. In doing so, I intended to kill him merely because I did not like him. However, unknown to me, if I had not killed him, he would have wrongfully killed

32. There is a stronger sense of a subjective justification under which a justified intention is necessarily subjectively right in the strong sense. However, I am concerned here only with the sense of a justification that mitigates the degree to which a criminal actor disrespects the rights of others in intending to perform her act. For I assume that a criminal actor is not blameworthy or liable to any punishment for her act if she did not disrespect the rights of others in intending to perform the act. A criminal actor's intention need not be subjectively right in the strong sense for her to avoid disrespecting the rights of others in having the intention.

33. There is a stronger sense of an objective justification under which a justified intention is necessarily objectively right in the strong sense. It is not my concern here. See *supra* note 32.

34. See George P. Fletcher, *The Right Deed for the Wrong Reason: A Reply to Mr. Robinson*, 23 UCLA L. Rev. 293 (1975) (rejecting such an entailment); Corrado, *supra* note 31, at 489 (same).

several others. Other things being equal, this makes my intending to kill him objectively right in both a weak and strong sense. But I do not have a subjective or objective justification for the intention because I do not actually have a type 1 defense for it. By intending to kill him under the conditions of my ignorance, I disrespected his right not to be killed merely because he was unlikely, and I disrespected it just as badly as I would have if my killing him were not in fact necessary to prevent him from wrongfully killing the others.

Sixth, the fact that a criminal act is necessary to obtain some information could standardly provide someone with only a subjective but not an objective justification for intending to perform it. Although she could have a type 1 defense for intending to perform the act as a means to obtaining the information, the type 1 defense would standardly not be indefeasible. For if she knew all the facts, and other things were equal, she would know the information sought independently of her performing the act. To illustrate, suppose I intend to torture someone as a necessary means to extracting information from him that I need to prevent a disaster he would otherwise cause. But if I knew the information sought independently of torturing him, I would know how to prevent the disaster without torturing him. Given my actual beliefs and motive, I have a subjective justification for the intention because I have a type 1 defense of defense of others for it. But I do not have an objective justification for the intention because my type 1 defense for it is not indefeasible. If I knew all the facts, and other things were equal, then I would know the information sought independently of torturing him, and so I would know how to prevent the disaster without torturing him. So by intending to torture him, I would disrespect his right not to be tortured.

Seventh, the concept of a type 1 defense is conceptually prior to the concepts of both a subjective and an objective justification. Both concepts of a justification are defined in terms of a type 1 defense, whereas the latter cannot be defined in terms of the former. So the fact that a criminal actor has a subjective or objective justification entails that she has a type 1 defense, but the fact that she has a type 1 defense does not entail that she has a subjective or objective justification.³⁵ Similarly, the concept of a subjective justification is conceptually prior to the concept of an objective

35. To illustrate, consider again cases of compulsion.

justification. Subjective justifications cannot be defined in terms of objective justifications, whereas the latter can be defined in terms of the former. On such a definition, a criminal actor has an objective justification for intending to perform her act if and only if (a) she has a subjective justification for the intention and (b) the type 1 defense that constitutes her subjective justification for it is infeasible. Thus, the fact that a criminal actor has an objective justification entails that she has a subjective one, but the fact that she has a subjective justification does not entail that she has an objective one.³⁶

C. Type 2 Defenses

When someone performs a criminal act, there is not only a presumption that in performing it, she disrespected the rights of others to a particularly bad degree. Given this presumption, there is also a presumption that she has a particularly bad disposition to commit crimes at the time of assessment. More precisely, there is a presumption that her act is particularly strong evidence that she has the presumed disposition to commit crimes at the time of assessment. In other words, there is a presumption that her act justifies others in believing with a particularly high credence that she has the presumed disposition to commit crimes at the time of assessment. The badness of someone's disposition to commit crimes is a function of at least four factors. First, it is primarily a function of the seriousness of the crimes she is willing to commit, where the seriousness of a crime corresponds to how badly someone disrespects the rights of others in committing the crime. Second, it is a function of the range of people against whom she is willing to commit crimes. Third, it is a function of the range of situations in which she is willing to commit crimes. Fourth, it is a function of the frequency with which she is willing to commit crimes.

A type 2 defense is a consideration that blocks the otherwise justified inference from the fact that someone performed a criminal act to her having the presumed disposition to commit crimes at the time of assessment.³⁷ A type 2 defense blocks this inference primarily by blocking the

36. To illustrate, consider again cases involving ignorance or mistake of fact.

37. Proponents of type 2 defenses generally endorse a "character theory of excuses." According to a character theory, a consideration constitutes an excuse if it blocks the otherwise justified inference from the fact that someone performed a criminal act to her having

otherwise justified inference from the fact that someone performed a criminal act to her being willing to commit crimes of the presumed degree of seriousness. If a criminal actor has a full type 2 defense, her act is not strong evidence that she has any disposition to commit crimes. In other words, her act does not justify others in believing with any credence that she has any disposition to commit crimes.³⁸ So it does not justify others in believing with any credence that she is willing to commit any crimes. If a criminal actor has a partial type 2 defense, her act might be strong evidence that she has some disposition to commit crimes. But her act is not as strong of evidence that she has as bad a disposition to commit crimes as was presumed. In other words, her act might justify others in believing with some credence that she has some disposition to commit crimes. But it does not justify them in believing with a credence as high as the presumed one that she has a disposition to commit crimes as bad as the presumed one. So her act might justify others in believing with some credence that she is willing to commit some crimes. But it does not justify them in believing with a credence as high as the presumed one that she is willing to commit crimes as serious as the presumed ones.

a motivational defect in her character. See, e.g., Michael D. Bayles, *Character, Purpose, and Criminal Responsibility*, 1 *Law & Phil.* 5 (1982); Michael D. Bayles, *Hume on Blame and Excuse*, 2 *Hume Stud.* 17 (1976); R.B. Brandt, *A Motivational Theory of Excuses in the Criminal Law*, in *Criminal Justice: Nomos XXVII* 165 (J. Roland Pennock & John W. Chapman eds., 1985); Richard B. Brandt, *A Utilitarian Theory of Excuses*, 78 *Phil. Rev.* 337, 353–58 (1969); Richard B. Brandt, *Blameworthiness and Obligation*, in *Essays in Moral Philosophy* 3 (A.I. Melden ed., 1958); David Hume, *A Treatise of Human Nature* 412, 477, 575 (L.A. Selby-Bigge & P.H. Nidditch eds., 2d ed. 1978); Peter Arenella, *Character, Choice and Moral Agency*, 7 *Soc. Phil. & Pol'y* 59 (1990); Tadros, *supra* note 14, at 293–321; George Vuoso, *Background, Responsibility, and Excuse*, 96 *Yale L.J.* 1661 (1987); cf. Scanlon, *supra* note 11, at 277–79 (suggesting that some considerations constitute excuses because they “sever the connection between the action or attitude and the agent’s judgments and character”); Strawson, *supra* note 11, at 8 (stating that “[w]e shall not feel resentment against the man he is for the action done by the man he is not; or at least we shall feel less”); Watson, *supra* note 11, at 123 (describing excuses that, according to Strawson, “present the other . . . as acting uncharacteristically due to extraordinary circumstances”).

38. More precisely, if a criminal actor has a full type 2 defense for her act, the act does not justify others in believing with any significant additional credence that she has any disposition to commit crimes. Even if a person has not committed any crimes, others are still justified in believing with some positive baseline credence that she is disposed to commit some crimes. No one is justified in being certain that anyone is not so disposed.

Many different considerations can provide a criminal actor with a type 2 defense. For example, all type 1 defenses are a subset of type 2 defenses because if a criminal actor did not disrespect the rights of others in performing her act, the act is not strong evidence that she is disposed to commit crimes.³⁹ In addition, though, several considerations are type 2 defenses that are not type 1 defenses. So a criminal actor can have a type 2 defense even though in performing her act, she manifested the presumed degree of disrespect toward the rights of others. For example, a criminal actor can have a type 2 defense because she performed her act as a result of an enduring mental illness that undermined her capacity to respond appropriately to the moral reasons against violating the rights of others,⁴⁰ and by the time of assessment, her illness or its symptoms have been eliminated by therapy or medication.

For another example, a criminal actor can have a type 2 defense because she performed her act under conditions that would standardly cause a temporary radical distortion in an agent's system of normative self-governance,⁴¹ which consists in all the norms she accepts regarding what to intend in particular situations.⁴² Such a distortion is radical if the norms the agent accepted before the distortion are radically different from and inconsistent with the ones she accepted under the distortion. Such a distortion is temporary if its cause is temporary, and after its cause subsides, she returns to accepting her *ex ante* system of normative self-governance.

If someone committed a crime under such distorting conditions, her crime does justify others in believing that under such conditions, she was disposed to commit crimes. However, she has a type 2 defense for her crime

39. Although type 1 defenses are a subset of type 2 ones, their salience warrants a separate exposition.

40. For the epistemic or volitional nature of such an incapacity, see *infra* text accompanying notes 49, 50.

41. See Fletcher, *supra* note 31, at 802 (stating that an "excuse represent[s] a limited, temporal distortion of the actor's character").

42. I assume that if a person accepts a norm regarding what to intend, the norm plays a distinctive role in her practical reasoning. Most generally, I assume that if she accepts a norm regarding what to intend in particular situations, then she adopts the norm as her policy for determining what to intend in those situations. More specifically, I assume that if she accepts a norm against intending to do something in particular situations, and she believes those situations obtain, then she will not intend to do it in those situations. In this sense, the acceptance of such a norm is an intention controlling attitude.

because her committing it does not justify others in believing that after those conditions subside, she is still disposed to commit crimes. For the fact that under distorting conditions she was disposed to commit crimes is not strong evidence that she is so disposed *ex post*. In general, what she was disposed to do under the distorting conditions is not strong evidence that she is similarly disposed *ex post*. Although under the distorting conditions she accepted a norm permitting her committing crimes, *ex ante* she might very well have accepted norms forbidding her committing crimes, and *ex post* she might very well return to accepting such norms again.

Examples of distorting conditions include intoxication, hypnosis, somnambulism, provocation, and other conditions that would temporarily impair someone's capacity to respond appropriately to the moral reasons against violating the rights of others.⁴³ To illustrate, consider a case of provocation. Suppose someone returns home and without any warning of an affair finds her husband intimately involved with his paramour. Given the extreme emotional disturbance she immediately experiences, she develops a strong motivational tendency to assault both and loses her capacity to respond appropriately to the moral reasons against violating their right not to be assaulted. As a consequence, she immediately strikes both without the benefit of a cooling down period. In this case, the woman does not have a type 1 defense for striking her husband and his paramour. She manifested the presumed degree of disrespect toward their right not to be assaulted. However, she does have a type 2 defense of provocation for striking them. The extreme emotional disturbance she naturally experienced in this situation would standardly cause a temporary radical change in an agent's system of normative self-governance. So her striking the victims does justify others in believing that in her emotionally disturbed state, she was disposed to commit crimes. But it does not justify others in believing that she is still disposed to commit crimes after cooling off. Although while in her emotional state, she accepted a norm permitting her committing crimes of assault, *ex ante* she might very well have accepted norms forbidding her committing any crimes, and *ex post* she might very well accept such norms again.

To illustrate another distorting condition, suppose unknown to me, my doctor injects into me a drug that makes me intoxicated. While intoxicated,

43. Cf. Horder, *supra* note 1, at 139–90 (endorsing an excuse of diminished capacity).

she then hypnotizes me, specifically implanting in me a hypnotic suggestion to kill someone wrongfully. In my intoxicated hypnotic state, I intentionally kill him. In this case, I do not have a type 1 defense for killing the victim. In killing him, I manifested the presumed degree of disrespect toward his right not to be killed. However, I do have a type 2 defense of intoxication and hypnosis for killing him. Being intoxicated and hypnotized are distorting conditions that would standardly cause a temporary radical change in an agent's system of normative self-governance. So my killing the victim does justify others in believing that in my intoxicated hypnotic state, I was disposed to commit crimes. But it does not justify others in believing that I am still disposed to commit crimes when I am no longer intoxicated or hypnotized. Although while intoxicated and hypnotized, I accepted a norm permitting my committing crimes of murder, *ex ante* I might very well have accepted norms forbidding my committing any crimes, and *ex post* I might very well accept such norms again.

Given my description of type 2 defenses, critics might argue that someone's having a full type 2 defense for a crime is generally inconsistent with her having no type 1 defense for it. In a standard case of blameworthiness and liability to punishment, a person's present self is blameworthy and liable to be punished for the crime of a person's past self. Similarly, in a standard case of a defense, a person's present self has a defense for the crime of a person's past self. In the cases at issue, the present self apparently has a full type 2 defense but no type 1 defense for the past self's crime. And in the cases at issue, the present self's system of normative self-governance is radically different from the system of the past self. According to the critics, their accepting radically different systems of normative self-governance generates the following dilemma for my claim that the present self has a full type 2 defense but no type 1 defense for the past self's crime.

On the first horn of the dilemma, suppose the present self and the past self are not the same person in the sense that they are not parts of the same person's life. Although they have the same body, they are not the same person because they have radically different psychological traits in virtue of their accepting radically different systems of normative self-governance. In this case, the present self does have a full type 2 defense for the past self's crime. Although the past self's crime does justify others in believing that she was disposed to commit crimes, it does not justify others in believing that the present self is similarly disposed. For the present self would

not persist as the same person under the conditions in which the past self would have committed crimes. In addition, though, to the full type 2 defense, the present self also has a type 1 defense for the past self's crime. Because they are different persons, neither the present self nor any of her own past selves actually disrespected the rights of others in virtue of the past self's doing so in committing the crime. So if they are different persons, the present self has not only a full type 2 defense but also a type 1 defense for the past self's crime.

On the second horn of the dilemma, suppose the present self and the past self are the same person in the sense that they are parts of the same person's life. Although they have radically different psychological traits, they are the same person because they have the same body and are psychologically connected or continuous in other relevant ways. In this case, the present self does lack a type 1 defense for the past self's crime. For the past self did disrespect the rights of others in committing the crime, and the past self is a past self of the present self. In addition, though, to lacking a type 1 defense, the present self also lacks a full type 2 defense for the past self's crime. For the past self's crime justifies others in believing that she was disposed to commit crimes under certain conditions. And because the present self and the past self are the same person, her crime also justifies others in believing that the present self is disposed to commit crimes under those same conditions. So if they are the same person, the present self has neither a full type 2 defense nor a type 1 defense for the past self's crime. As a consequence, whether or not the present self and the past self are the same person, the former could not have a full type 2 defense but no type 1 defense for the latter's crime. So in the cases at issue, someone could not have a full type 2 defense but no type 1 defense for committing a crime.

To illustrate the first horn of the dilemma, reconsider the case in which I, my present self, apparently have a full type 2 defense but no type 1 defense for my apparent past self's killing someone while intoxicated and hypnotized. Suppose, though, that my apparent past self is not in fact my past self: we are not the same person in the sense that we are not both parts of the same person's life. Although we have the same body, we are not the same person because we have radically different psychological traits. In particular, I accept norms forbidding my committing any crimes, whereas he accepted norms permitting his committing crimes of murder. In this case, I have a full type 2 defense for his killing. Although his killing justifies others in believing that he was disposed to commit crimes, it does not

justify others in believing that I am similarly disposed. For I would not persist as the same person under the conditions of intoxication and hypnotism in which he would have committed crimes. In addition, though, to having a full type 2 defense, I also have a type 1 defense for his killing. Because we are different persons, neither I nor any of my past selves disrespected the rights of others in virtue of his doing so in killing the victim. So if we are different persons, I have not only a full type 2 defense, but also a type 1 defense for his killing.

To illustrate the second horn, reconsider the same case. Suppose, though, that my apparent past self is in fact my past self: we are the same person in the sense that we are both parts of the same person's life. Although we have radically different psychological traits, we are the same person because we have the same body and are psychologically connected or continuous in other relevant ways. In this case, I do lack a type 1 defense for my past self's killing because he is my past self, and he did disrespect the rights of others in killing the victim. In addition, though, to lacking a type 1 defense, I also lack a type 2 defense for his killing. For his killing justifies others in believing that I am now disposed to commit crimes when in a similarly intoxicated hypnotic state. So if I and my apparent past self are the same person, I have neither a full type 2 defense nor a type 1 defense for his killing. As a consequence, whether or not we are the same person, I could not have a full type 2 defense but no type 1 defense for his killing.

In response, there is no dilemma for the cases at issue. Someone's having a full type 2 defense for a crime is not generally inconsistent with her having no type 1 defense for it. The first horn of the purported dilemma is implausible because it rests on an implausible theory of personal identity. In the cases at issue, the present self and the past self do have radically different psychological traits in virtue of their accepting radically different systems of normative self-governance. So they do have radically different personalities or characters. But this fact by itself does not entail that they are different persons in the sense that they are not parts of the same person's life. For example, almost every old adult has many past selves who were very young children with radically different psychological traits from her present self. In spite of their radical psychological differences, they are the same person.⁴⁴

44. See Thomas Reid, *Of Mr. Locke's Account of Our Personal Identity*, in *Personal Identity* 113 (John Perry ed., 1975).

In the cases at issue, the selves might also be the same person even though they have radically different psychological traits. Three facts could support such a possibility. First, the selves have not only the same brain, but also the same body.⁴⁵ Second, they might be psychologically connected to each other in the sense that they might have some significant psychological traits in common, and the traits of the past self might have caused the traits of the present self in the normal way that preserves the relation of personal identity between selves. For example, the present self might remember some experiences of the past self. Third, even if they are not psychologically connected to each other, they might be psychologically continuous with each other in the sense that the present self might be connected to the past self by a chain of selves such that each self in the chain is psychologically connected to its immediately preceding self in the chain.⁴⁶ Assuming they have the same brain and body, and are psychologically connected to or continuous with each other, and there is no branching between them, then they are the same person on any plausible theory of personal identity even though they have radically different psychological traits.

The second horn of the purported dilemma is implausible because its sense of a disposition is too broad. According to the broad sense, someone is presently disposed to do something if she would do it under any possible conditions. But the sense of a disposition at issue in type 2 defenses is much narrower. It is the sense of a disposition that is constitutive of a character trait.⁴⁷ According to the character sense, someone is presently disposed to do something only if she would do it under a relevant range of conditions. And the relevant range consists only in those conditions in which her present system of normative self-governance is held fixed. In other words, the relevant range consists only in those conditions in which she would accept the same system of normative self-governance that she actually accepts at present. So the relevant range does not include conditions in which she

45. See Bernard Williams, *The Self and the Future*, in *Problems of the Self* 46 (1973) (defending the possibility that the relation of personal identity consists in a relation of bodily continuity).

46. See Derek Parfit, *Personal Identity*, 80 *Phil. Rev.* 3 (1971) (defending a psychological continuity theory of personal identity); Derek Parfit, *Reasons and Persons* 204–09 (1984) (same); Sydney Shoemaker, *Persons and Their Pasts*, 7 *Am. Phil. Q.* 269 (1970) (same).

47. Cf. Richard B. Brandt, *Traits of Character: A Conceptual Analysis*, 7 *Am. Phil. Q.* 23 (1970) (describing character traits in terms of dispositions).

would accept a system of normative self-governance that is radically different from the one she actually accepts at present. What she would do under such conditions is not relevant to what she is presently disposed to do in the character sense of a disposition at issue in type 2 defenses.

To illustrate the distinction between these senses of a disposition, consider someone who accepts a norm of extreme honesty that forbids her from lying under any conditions. In the character sense of a disposition, she is disposed to be honest and not disposed to be dishonest. In the relevant range of conditions, she would never lie because she would accept a norm of extreme honesty in those conditions. Nevertheless, she might still have a broad disposition to be dishonest. For example, it might be the case that if she were to suffer certain types of brain damage, she would lie because, under those conditions, she would accept a norm permitting her to lie. Such a broad disposition, though, does not bear on what she is presently disposed to do in the character sense of a disposition.

Now in the cases at issue, a person has a full type 2 defense but no type 1 defense for committing a crime. Given that she has no type 1 defense for the crime, it justifies others in believing that she has a broad disposition to commit crimes because it justifies others in believing that she would commit crimes under the type of conditions that actually elicited her crime. So the crime justifies others in believing that under the eliciting conditions, she would accept norms permitting her to commit crimes. But given that she has a full type 2 defense for the crime, it does not justify others in believing that she is presently disposed to commit crimes in the character sense of a disposition. Her crime does not justify others in believing that at present, she actually accepts norms permitting her to commit crimes. For given that she has a full type 2 defense for the crime, the fact that she accepted such norms under the eliciting conditions is not strong evidence that she accepts such norms at present after those conditions have subsided.

To illustrate, reconsider the case in which I have a full type 2 defense but no type 1 defense for killing someone while I am intoxicated and hypnotized. Given that I have no type 1 defense for the crime, it justifies others in believing that I have a broad disposition to commit crimes because it justifies others in believing that I would commit crimes under conditions in which I were similarly intoxicated and hypnotized. So my crime justifies others in believing that under such distorting conditions, I would accept norms permitting my committing crimes. But given that I have a full type 2 defense for the crime, it does not justify others in believing that

I am presently disposed to commit crimes in the character sense of a disposition. The crime does not justify others in believing that at present, I actually accept norms permitting my committing crimes. For given that I have a full type 2 defense for the crime, the fact that I accepted such norms while intoxicated and hypnotized is not strong evidence that I accept such norms at present after those distorting conditions have subsided.

D. Type 3 Defenses

When someone performs a criminal act, there is a presumption that in performing it, she flouted certain moral reasons against violating the rights of others. As a consequence, there is a presumption that at the time of assessment, she is disposed to flout a relevant class of moral reasons against violating the rights of others, where the relevant class is a function of the moral reasons she is presumed to have flouted in performing the act. Type 3 defenses are considerations that undermine a criminal actor's capacity to respond appropriately to the relevant class of moral reasons at the time of assessment.⁴⁸ If a criminal actor has a full type 3 defense, she lacks the capacity to respond appropriately to any of the relevant moral reasons. If she has a partial type 3 defense, she might have the capacity to respond appropriately to some of the relevant moral reasons, but not to all of them and particularly not to the strongest of them.

A criminal actor might lack the capacity in two respects. First, she might lack the epistemic capacity to recognize the relevant reasons or to appreciate them in the sense of recognizing them as reasons. The scope of her epistemic incapacity might vary. At one extreme, she might have a broad incapacity to recognize or appreciate any reasons whether moral or nonmoral. At another extreme, she might have only a narrow incapacity to recognize or appreciate the relevant moral reasons. Second, even if she has the epistemic capacity, she might lack the volitional capacity to control what she intends to do in response to her judgments about what the relevant moral reasons count for or against her intending.⁴⁹ So she might

48. Cf. Scanlon, *supra* note 11, at 280 (describing defenses that consist in an incapacity to respond appropriately to reasons); Strawson, *supra* note 11, at 8–10 (same); Watson, *supra* note 11, at 123 (same); Model Penal Code art. 4 (1985) (same).

49. Assuming she suffers from such an incapacity, she might judge that she has most reason not to intend to do something but intend to do it anyway, and vice versa.

lack the capacity to control what norms she accepts governing intentions in response to her judgments about what norms the relevant moral reasons count for or against her accepting.⁵⁰ The scope of her volitional capacity might also vary. At one extreme, she might have a broad incapacity to control what she intends to do in response to her judgments about what any reasons count for or against her intending. At another extreme, she might have only a narrow incapacity to control what she intends to do in response to her judgments about what the relevant moral reasons count for or against her intending. Examples of such epistemic or volitional incapacities include insanity, more specific forms of mental illness such as kleptomania, and possibly other limiting conditions more prevalent among the normal population.

Some considerations that provide type 3 defenses also provide type 1 and type 2 ones. To illustrate, suppose I suffer from an incurable mental illness that makes me permanently incapable of understanding the concept of a person. So I am permanently incapable of recognizing others as thinking, feeling, and demanding subjects. Now suppose I kill someone to acquire a resource, such as food, that he would otherwise consume or take away. In this case, my mental illness provides me with a type 1, type 2, and type 3 defense for killing him. It provides me with a type 1 defense because I could not have disrespected the victim's rights without recognizing him as a person. It provides me with a type 2 defense because it blocks the otherwise justified inference from the fact that I killed him to my being disposed to commit crimes. For it blocks the otherwise justified inference from the fact that I killed him to my being disposed to disrespect the rights of others.⁵¹ And in virtue of being permanent, my mental illness provides me with a type 3 defense because my incapacity to understand the concept of a person makes me incapable, at the time of assessment, of responding appropriately to any of the moral reasons against violating the

50. Assuming she suffers from such an incapacity, she might judge that she has most reason to accept a particular norm governing intentions but reject it anyway, and vice versa.

51. In this case, I am disposed to perform criminal acts. That is, I am disposed to perform acts that satisfy the *actus reus* requirements of a crime. However, I am not disposed to commit crimes because I am incapable of satisfying their *mens rea* requirements. For I am incapable of disrespecting the rights of others in performing criminal acts in virtue of the fact that I am incapable of recognizing others as persons. For example, we might suppose that if I were ever to strangle someone to death, I would believe I were merely squeezing a lemon.

rights of others. The illness makes me incapable of recognizing or appreciating such reasons.

Although there is some overlap, not all considerations that provide type 3 defenses also provide type 1 or type 2 ones. Unlike type 1 defenses, type 3 ones are consistent with a criminal actor's disrespecting the rights of others to the presumed degree in performing her act. Type 3 defenses focus only on considerations that obtain at the time of assessing the criminal actor for blameworthiness and liability to punishment. Specifically, type 3 defenses focus only on her capacities at the time of assessment rather than at the time she performed her act.⁵² So if the criminal actor lacks the relevant capacity at the time of assessment, she has a type 3 defense for her act even if she had the capacity when she performed it.⁵³ Type 1 defenses, though, focus only on considerations that obtained at the time the criminal actor performed her act. Specifically, type 1 defenses focus only on how badly she disrespected the rights of others in performing the act. So if she did so to the presumed degree, she lacks a type 1 defense for the act even if she has a type 3 defense for it because she lacks the relevant capacity at the time of assessment.

To illustrate, suppose I intentionally kill someone for no sound reason under conditions in which I have the capacity to respond appropriately to the moral reasons against doing so. After killing him, though, I develop the mental illness that makes me incapable of understanding the concept of a person. In this case, I lack a type 1 defense for killing the victim because in doing so, I disrespected his rights to the presumed degree. But I have a type 3 defense for killing him because at the time of assessment, I lack the capacity to respond appropriately to any of the moral reasons against violating the rights of others.

Unlike type 2 defenses, type 3 ones are consistent with a criminal act's providing others with knowledge that the actor has the presumed disposition to commit crimes. Type 3 defenses are concerned with whether the criminal actor has the capacity to respond appropriately to the relevant moral reasons. But they are not concerned with whether she actually responds appropriately to such reasons. She might not actually respond

52. If a criminal actor performed her act due to a mere temporary incapacity at the time of the act, the temporary incapacity would constitute a type 2 but not a type 3 defense.

53. See R.A. Duff, *Trials and Punishments* 14–38 (1986) (emphasizing this possibility).

appropriately to them even if she has the capacity to.⁵⁴ Similarly, type 3 defenses are concerned with whether the criminal actor has the capacity to be disposed to respond appropriately to such reasons. But they are not concerned with whether she is actually so disposed. She might not be so disposed even if she has the capacity to be.⁵⁵ Thus, type 3 defenses are concerned with whether the criminal actor has the capacity to be disposed to choose not to commit crimes on the basis of the moral reasons against committing them. But type 3 defenses are not concerned with whether she is actually so disposed. Again, she might not be so disposed even if she has the capacity to be. Type 2 defenses, though, are concerned with whether the criminal actor is actually disposed to commit crimes insofar as they are concerned with whether her act justifies others in believing that she is actually so disposed. Thus, if her act provides others with knowledge that she has the presumed disposition to commit crimes, then she lacks a type 2 defense for the act even if she has a type 3 defense for it because she lacks the capacity to respond appropriately to the relevant moral reasons at the time of assessment.

To illustrate, suppose again that I intentionally kill someone for no sound reason. Given that I disrespected his rights in killing him, I lack a type 1 defense for the act. Suppose also that I lack a type 2 defense for the act because it provides others with knowledge that I am disposed to commit crimes at the time of assessment. Nevertheless, I might still have a type 3 defense for the act because at the time of assessment, I might lack the capacity to respond appropriately to the relevant moral reasons against violating the rights of others. In this case, I do have the epistemic capacity to

54. I assume someone might not do something that she has the capacity to do. And I assume someone might do something that she has the capacity not to do. I set aside for further analysis the potential problems that the truth of determinism might raise for these assumptions. I leave it an open question whether determinism is true. And I leave it an open question whether these assumptions are consistent with the truth of determinism. Cf. David Copp, "Ought" Implies "Can," Blameworthiness, and the Principle of Alternate Possibilities, in *Moral Responsibility and Alternative Possibilities* 265, 291–95 (David Widerker & Michael McKenna eds., 2003) (arguing that determinism is consistent with an actor's ability to do otherwise).

55. Type 3 defenses are concerned with whether a criminal actor has the capacity to accept certain norms against violating the rights of others. But they are not concerned with whether she actually accepts such norms. She might not accept them even if she has the capacity to.

recognize the relevant moral reasons, but I might suffer from an extreme form of psychopathy that makes me incapable of recognizing them as reasons. For example, in killing the victim, I did recognize that I was killing a person, but as a result of the psychopathy, I might have lacked then and now the capacity to recognize this fact as a moral reason against killing him.⁵⁶ Alternatively, even if I have all the epistemic capacities, I might suffer from a mental illness that makes me incapable of controlling what I intend to do in response to my judgments about what the relevant moral reasons count for or against intending. For example, I might have judged that I had most reason not to intend to kill the victim, but as a result of the mental illness, I might have intended to do so anyway. Assuming I continue to suffer from such capacities at the time of assessment, they provide me with a type 3 defense but not a type 2 defense for killing the victim.

E. Additional Types of Defenses

At this point, I have spelled out three general types of defenses that account for a very wide range of more particular defenses. In two ways, though, these three types do not exhaust every type of defense. First, there might be other general types that pick out particular defenses left out from the three covered so far. Later in the paper, I will argue that the theory of punitive desert that best explains the mitigating effects of these three types does entail a more general fourth type of defense. Expounding the fourth type must await our exposition of my restorative signaling theory of punitive desert. Second, there are many finer-grained distinctions to be drawn among the more particular defenses picked out by the more general types. In other words, within the general types of defenses, we could demarcate many subtypes.

We have already seen that justifications are a subset of type 1 defenses whose salience warrants a separate exposition. Other salient subtypes include exemptions, failure of proof defenses, and excuses. An exemption is an enduring incapacity to respond appropriately to any moral reasons.⁵⁷ Under an exemption, a person would not be blameworthy or liable to punishment for any acts she might perform, including criminal acts. Exemptions are a subset of type 3 defenses.

56. For an actual case of a criminal who might have suffered from such an incapacity, see the story of Robert Harris discussed in Watson, *supra* note 11, at 131–37.

57. See Tadros, *supra* note 14, at 124–29.

Consider now failure of proof defenses.⁵⁸ In general, the state is justified in punishing someone for committing a crime only if a statute explicitly prohibits the crime.⁵⁹ The statute spells out the crime's essential elements. Some of the elements consist in *actus reus* requirements. They describe the type of act that someone must perform to commit the crime. Other elements consist in *mens rea* requirements. They describe the state of mind with which someone must perform the act to commit the crime. If someone does not satisfy all the elements of the crime, then she has a failure of proof defense to the charge that she committed it. This defense consists in her not satisfying one of the crime's elements. Failure of proof defenses are a subset of type 1 defenses.

Failure of proof defenses are especially significant because they have distinctive implications for the state's burden of proving at trial that a defendant not only committed a crime, but also is liable to punishment for committing the crime.⁶⁰ At trial, the state has the initial burden of proving beyond a reasonable doubt that the defendant satisfied all the elements of the crime charged.⁶¹ So the state has the initial burden of proving beyond a reasonable doubt that the defendant lacks a failure of proof defense to the charge that she committed the crime. However, the state might not have the initial burden of proving that the defendant lacks other types of defenses, such as type 3 ones.⁶² The defendant herself might have the initial burden of proving that she has these other defenses before the state incurs a burden of proving otherwise.

The concept of an excuse is open to broader and narrower conceptions.⁶³ On a broader conception, the concept of an excuse picks out any defense that a criminal actor might have for her act. On a narrower conception, the concept of an excuse refers only to a class of defenses not picked out by some range of other salient types of defenses, such as justifications, exemptions, and failure of proof defenses.

58. See Joshua Dressler, *Understanding Criminal Law* 181–82 (1995); Robinson, *supra* note 6, at 204–08.

59. See Model Penal Code § 1.05(1) (1985).

60. See Dressler, *supra* note 58, at 51–61; Robinson, *supra* note 6, at 250–64.

61. See Model Penal Code § 1.12(1) (1985).

62. See *id.* § 1.12(2)–(4).

63. See Marcia Baron, *Excuses, Excuses*, 1 *Crim. L. & Phil.* 21, 37 (2007) (crediting Antony Duff with pointing out broader and narrower senses of “excuse”).

Whether a defense falls under the extension of one of these more particular subtypes might have significant implications. But for the purposes of explaining the mitigating effects of defenses in general, we can abstract away from these more fine-grained distinctions between defenses and focus only on the general types. For in virtue of explaining the mitigating effects of the more general types, we would thereby explain the mitigating effects of all the more particular subtypes of defenses, such as justifications, exemptions, failure of proof defenses, and excuses.

IV. A GENERAL THEORY OF THE JUSTIFICATION OF PUNISHMENT

My general theory specifies the general conditions under which the state would be justified in punishing someone against her will. On the strong sense of justification at issue here, it specifies the general conditions under which the state would not be open to any warranted attitude of moral disapproval for punishing someone against her will. So the theory specifies the general conditions under which the state has most reason to punish someone against her will. According to the theory, the state is justified in imposing a punishment on someone against her will if and only if five requirements are satisfied. First, according to the desert requirement, the person must deserve the punishment. The desert requirement is warranted because if the state were to impose an undeserved punishment on someone against her will, it would violate her rights. The fact that the state would violate someone's rights by doing something provides the state with an overriding reason not to do it.⁶⁴ Second, according to the general rights requirement, imposing the punishment on the person must not violate the rights of anyone else. Third, according to the value requirement, the expected value of the consequences of imposing the punishment on the person must be at least as high as the expected value of the consequences of

64. To keep matters simple, I set aside the fact that rights standardly have thresholds such that the state might have most reason to violate someone's rights when the consequences of not violating them would be sufficiently bad. So I set aside the possibility that in some emergency situations, the state might have most reason to impose an undeserved punishment on someone against her will because doing so would be necessary to "avoid catastrophic moral horror." Robert Nozick, *Anarchy, State, and Utopia* 30 (1974).

any other available act that would not violate anyone's rights. The value requirement is warranted because if the expected value of such an alternative act were higher than the expected value of imposing the punishment on her, the state would have more reason to perform the alternative act instead. Fourth, according to the epistemic requirement, the state must know that the first three requirements are satisfied in imposing the punishment on her. Fifth, according to the motivation requirement, the state must be appropriately motivated by its knowledge that the first three requirements are satisfied in imposing the punishment on her.

Assuming the state would punish criminal actors against their will, both the desert and the value requirements are *prima facie* plausible grounds on which to explain the mitigating effects of defenses on a criminal actor's liability to punishment. Regarding the desert requirement, criminal actors with full defenses might not deserve any punishment. Those with partial defenses might deserve to be punished only to a mitigated degree. Regarding the value requirement, the expected value of punishing criminal actors with full defenses might be lower than the expected value of not punishing them at all. The expected value of punishing those with partial defenses to an unmitigated degree might be lower than the expected value of punishing them to a mitigated degree. We will consider each requirement in turn.

V. THE VALUE REQUIREMENT

Suppose the benefits of a punishment consist in its good consequences, and its costs consist in its bad consequences. Suppose the net benefit of a punishment consists in the difference between its benefits and costs. To explain on the basis of the value requirement the mitigating effects of defenses on a criminal actor's liability to punishment, we must prove two claims. First, we must show that the expected costs would outweigh the expected benefits of punishing criminal actors with full defenses. Second, we must show that the expected net benefit of punishing criminal actors with partial defenses to an unmitigated degree would be lower than the expected net benefit of punishing them to a mitigated degree. We will focus primarily on trying to prove the first claim regarding full defenses. If we find such a proof, we can consider whether it generalizes to prove the second claim regarding partial defenses. But if the first is unprovable, the

second is too. In the following analysis, whenever I refer to a defense, I refer to a full defense unless I note otherwise. To be concise, I leave the expected qualification to costs and benefits implicit.

A. First Argument

According to one argument, the consequences of a punishment include the punishment itself, and the punishment itself is intrinsically bad.⁶⁵ So a punishment is itself a cost, which I call its “punitive cost.” To satisfy the value requirement, a punishment must have some benefit that outweighs its punitive cost. One benefit of a punishment is its deterrence benefit. The state deters people from performing criminal acts by following a policy of threatening to punish anyone who performs criminal acts and punishing those who it discovers have performed them. Such a policy achieves deterrence by providing people with a strong prudential incentive not to perform criminal acts for fear of being punished for performing them. The deterrence benefit of a punishment is the benefit of preventing all the criminal acts that it deters people from performing. Assuming the main benefit of a punishment is its deterrence benefit, some might contend that the punitive cost would outweigh the benefits of any punishment that has no deterrence benefit. They might argue that punishing criminal actors with defenses would have no deterrence benefit. They might contend that criminal actors with defenses are undeterrable. In other words, when someone performs a criminal act with a defense, she performs the act under conditions in which she could not be deterred from performing it by any threat to punish her for performing the act under its description as a criminal act. Because criminal actors with defenses are undeterrable, punishing them would deter no one from performing any criminal acts.⁶⁶ So punishing them would have no deterrence benefit. Therefore, the punitive costs would outweigh the benefits of punishing them.

65. To keep matters simple, I assume every punishment is intrinsically bad. I also assume that the degree to which a punishment is intrinsically bad is proportional to its severity. The more severe a punishment, the worse it is.

66. Cf. Jeremy Bentham, *An Introduction to the Principles of Morals and Legislation* 158–62 (J.H. Burns & H.L.A. Hart eds., 1970) (arguing that punishing criminal actors with some defenses would be inefficacious as it would not prevent any mischief).

To illustrate, consider a defense of ignorance or mistake of fact. If a criminal actor has such a defense, she did not believe her act had the properties that made it a criminal act. So she could not have been deterred from performing her act by any threat to punish her for performing such an act under its description as a criminal act. Because criminal actors with a defense of ignorance or mistake of fact are undeterrable, punishing them would deter no one from performing any criminal acts, and so punishing them would have no deterrence benefits.

In response, the first argument is unpersuasive for two reasons among others. According to the first objection, suppose criminal actors with some defenses are undeterrable. So punishing them would not deter others from performing criminal acts under those same defense eliciting conditions. Nevertheless, punishing them might deter people from performing criminal acts under other conditions, and so punishing them might have deterrence benefits that would outweigh the punitive costs of punishing them.⁶⁷ For punishing people for performing a particular type of criminal act under a particular type of condition can generally deter others from performing different types of criminal acts under different types of conditions.

Punishing criminal actors with defenses can have such a general deterrence effect in two ways. First, punishing them can make others more vividly aware of how bad it would be to suffer a punishment in general. By seeing, hearing about, or talking to someone punished, others might better appreciate the severity of the suffering involved in being punished. Second, punishing them can make others discount to a lesser degree their probability of being punished for performing criminal acts in general. For punishing criminal actors with defenses indicates that the state is actively trying to detect and punish criminal actors in general. And the state's punishing them reduces the range of defenses available to people charged with committing crimes. If the state were not to punish them, some might commit crimes without those defenses because they believe they can avoid being punished for their crimes by deceiving the state into believing that they performed their acts with those defenses.⁶⁸ But if the state were to

67. See H.L.A. Hart, *Prolegomenon to the Principles of Punishment*, in *Punishment and Responsibility: Essays in the Philosophy of Law* 1, 19 (1968).

68. See *id.* Some might worry there is tension between my presuppositions 4 and 5 and the possibility of someone's committing a crime because she believes she will deceive the state into believing falsely that she performed her criminal act with a defense. For to deceive

punish criminal actors with defenses, others would not discount their probability of punishment in response to the possibility of such deception.

According to the second objection, not all criminal actors with defenses are undeterrable. For example, consider type 3 defenses in which a criminal actor lacks the epistemic capacity to recognize and appreciate as such the moral reasons against violating the rights of others. Such an actor might still retain the capacity to recognize and appreciate merely prudential reasons against performing criminal acts. Assuming she retains the capacity to respond appropriately to prudential reasons, she might be deterrable. For another example, consider duress. In a standard case of duress, someone performs a criminal act in response to a threat to harm her unless she performs it. Assuming the threatened harm was very severe and imminent, it would have been very difficult for the criminal actor to resist performing the act. So she was not easily deterrable. But she still might have been deterrable by the state's threat of an extraordinarily severe punishment for not resisting under duress.⁶⁹ Such a punishment would harm her much more severely than the harm she faced for resisting under duress.

B. Second Argument

A second argument might concede that criminal actors with some defenses are deterrable, and punishing criminal actors with any defense can have some deterrence benefits. But with respect to particular defenses,

the state in this way, she must deceive the state about her capacities or the beliefs, intentions, and motives with which she performed her act. In fact, though, there is no tension. According to presuppositions 4 and 5, when a criminal actor is assessed for blameworthiness and liability to punishment, everyone knows all the facts about her capacities and the beliefs, intentions, and motives with which she performed her act. However, there is no presupposition that everyone believes that presuppositions 4 and 5 will be satisfied at the time of assessment. In particular, there is no presupposition that a criminal actor will believe at the time of her act that everyone will know all the relevant facts at the time she is assessed for blameworthiness or liability to punishment. Presuppositions 4 and 5 are consistent with a criminal actor's believing at the time of her act that the state will not know all the relevant facts at the time of assessment.

69. For this reason, considerations of deterrence arguably favor punishing criminal actors with a duress defense extraordinarily severely. See Anthony Kenny, *Duress Per Minas* as a Defence to Crime: II, in *Law, Morality and Rights* 345, 352 (M.A. Stewart ed., 1983); J.L. Mackie, *Duress and Necessity as Defences to Crime: A Postscript*, in *Law, Morality and Rights* 365, 367 (M.A. Stewart ed., 1983); 2 Sir James Fitzjames Stephen, *A History of the Criminal Law of England* 107 (London, MacMillan 1883).

there are other reasons to believe that the punitive costs would outweigh the benefits of punishing criminal actors with them. Consider again duress. Assuming it is extraordinarily difficult to deter people from complying with a threat under duress, people under duress are deterrable only by punishing them extraordinarily severely for not resisting the threat. An extraordinarily severe punishment, though, would be very bad in itself. Moreover, such a punishment would be worse than the harm someone would cause by performing the criminal act demanded under duress. Hence, the punitive cost of punishing a criminal actor with a duress defense extraordinarily severely not only would be very high, but also would outweigh its deterrence benefit.⁷⁰ So the punitive costs would outweigh the benefits of punishing criminal actors with a duress defense.

In response, we might concede that an extraordinarily severe punishment would be worse than the harm someone would cause by performing a single criminal act under duress. But the second argument is still unpersuasive because, among other reasons, punishing a criminal actor with a duress defense extraordinarily severely might deter multiple people from performing multiple criminal acts under duress. The extraordinarily severe punishment might be better than the aggregate harm that others would cause by performing multiple criminal acts under duress. So the deterrence benefit of punishing a criminal actor with a duress defense extraordinarily severely not only might be very high, but also might outweigh its punitive cost.

C. Third Argument

A third argument might concede that punishing a criminal actor with a duress defense would deter multiple people from performing multiple criminal acts under duress. But in addition to having a deterrence benefit, the punishment would also have a deterrence cost, which is the cost of preventing all the acts it would deter people from performing. Some might contend that the deterrence cost of the punishment would outweigh its deterrence benefit because criminal acts performed under duress are optimistic. They are optimistic because in a case of duress, the threatened harm

70. Cf. Bentham, *supra* note 66, at 159, 163–64 (arguing that punishing criminal actors with some defenses would be unprofitable or too expensive because “the mischief it would produce would be greater than what it prevented”).

would be worse for the person under duress than the alternative harm she would directly cause by performing the criminal act demanded.⁷¹ Hence, assuming that punishing criminal actors with a duress defense would deter only optimific acts, the deterrence costs would outweigh the deterrence benefits of punishing them. So the overall costs would outweigh the overall benefits of punishing them.

To illustrate, suppose someone threatens to assault me unless I commit a theft. My committing the theft would be optimific because the harm of my being assaulted would be worse than the alternative harm I would directly cause by committing the theft. Thus, assuming that punishing me for committing the theft would deter only optimific acts, the deterrence cost would outweigh the deterrence benefit of the punishment. So its overall costs would outweigh its overall benefits.

In response, the third argument is unpersuasive for three reasons among others. First, suppose for the sake of argument that in a case of duress, the threatened harm would be worse for the person under duress than the alternative harm she would directly cause by performing the criminal act demanded. Nevertheless, performing the criminal act might not be optimific. For performing it is not to resist the threat. And not resisting the threat might result in the indirect harm of increasing the incidence of duress and hence the number of victims of duress. Conversely, resisting the threat might result in the indirect benefit of decreasing the incidence of duress and the number of victims of duress.

On the one hand, by resisting the threat and hence not performing the criminal act demanded, the person under duress would signal that she is threat resistant in the sense that she is disposed to resist unauthoritative threats under duress. In so doing, she would also set a precedent of threat resistance for others to follow when they are under duress. As a result, others might also resist threats under duress, and thereby signal that they too are threat resistant. As a consequence, potentially threatening agents might infer that she and others too are threat resistant. Thus, they might refrain from putting others under duress as a means of fulfilling their criminal objectives. For if someone wants others to perform a criminal act, and he knows they are likely to be threat resistant, then it is not rational

71. Cf. Peter Westen & James Mangiafico, *The Criminal Defense of Duress: A Justification, Not an Excuse—and Why It Matters*, 6 *Buff. Crim. L. Rev.* 833 (2003).

for him to place them under duress as a means of getting them to perform the act. So by increasing the perceived incidence of threat resistance, resisting threats under duress might reduce the incidence of duress and the number victims of duress.⁷²

On the other hand, by not resisting the threat and hence performing the criminal act demanded, the person under duress would signal that she is not threat resistant. In so doing, she would also set a precedent of threat non-resistance for others to follow when they are under duress. As a result, others might also not resist threats under duress, and thereby signal that they too are not threat resistant. As a consequence, potentially threatening agents might infer that she and others too are not threat resistant. Thus, they might engage in putting others under duress as a means of fulfilling their criminal objectives. For if someone wants others to perform a criminal act, and he knows they are likely to be not threat resistant, then it might be prudentially rational for him to place them under duress as a means of getting them to perform the act. So by decreasing the perceived incidence of threat resistance, not resisting threats under duress might increase the incidence of duress and the number of victims of duress. Assuming the indirect benefits of resistance and the indirect harms of non-resistance would be sufficiently large, performing the criminal act demanded under duress would not be optimific.

Second, suppose for the sake of argument that the criminal acts of those with a duress defense are optimific. Nevertheless, the deterrence benefits might still outweigh the deterrence costs of punishing criminal actors with a duress defense. In addition to deterring optimific criminal acts performed under duress, punishing them might also deter non-optimific criminal acts performed under other conditions. In response to punishing them, potential criminal actors in general might become more vividly aware of how bad being punished would be for them, and they might discount to a lesser degree their probability of punishment. For example, punishing criminal actors with a full defense of duress might deter others from performing non-optimific criminal acts under conditions in which they would really have only a partial defense of duress but mistakenly believe they have a full defense of duress. The deterrence benefits might outweigh

72. Cf. Parfit, *Reasons and Persons*, *supra* note 46, at 20–23, 457–61 (explaining why it would probably be rational for a person to be transparently disposed to ignore threats).

the deterrence costs of punishing even optimific criminal acts performed under duress if the non-optimific acts deterred are sufficiently numerous and harmful relative to the optimific ones deterred.

Third, suppose for the sake of argument that punishing criminal actors with a duress defense would deter only optimific criminal acts. So the direct deterrence costs would outweigh the direct deterrence benefits of punishing them. Nevertheless, the overall deterrence benefits might still outweigh the overall deterrence costs of punishing them because punishing them might have the indirect deterrence benefit of decreasing the incidence of duress and hence the number of victims of duress. For punishing them might increase the number of people who are threat resistant by increasing the number of people who are deterred from performing criminal acts demanded under duress. Hence, the deterrence effect of punishing criminal actors with a duress defense might increase the incidence of threat resistance. Assuming potentially threatening agents would be aware of this higher incidence, they might refrain from putting others under duress as a means of fulfilling their criminal objectives. Thus, by increasing the perceived incidence of threat resistance, the deterrence effect of punishing criminal actors with a duress defense might decrease the incidence of duress and the number of victims of duress.⁷³ Assuming this indirect deterrence benefit would be sufficiently large, the overall deterrence benefits might outweigh the overall deterrence costs of punishing them. So the overall benefits might outweigh the overall costs of punishing them.⁷⁴

D. Fourth Argument

A fourth argument might concede that punishing criminal actors with a duress defense would have indirect deterrence benefits. But according to the argument, the overall costs would still outweigh the overall benefits of

73. By parity of reasoning, not punishing criminal actors with a duress defense might make people less threat resistant and, consequently, might increase the incidence of duress and the victims of duress. See Kenny, *supra* note 69, at 348, 352–53; *Abbott v. The Queen*, [1977] A.C. 755 (Lord Salmon); cf. Parfit, *Reasons and Persons*, *supra* note 46, at 20–23, 457–61.

74. As an apparently paradoxical implication of this analysis, a punishment might be optimific even if it deters the performance of only optimific criminal acts. The appearance of paradox, though, dissipates in light of the fact that deterring the performance of even optimific criminal acts can have indirect benefits that would outweigh the costs of deterring such acts.

punishing criminal actors with a duress defense or a range of other defenses. The relevant range consists in those defenses that preclude a criminal actor from having a fair opportunity to avoid performing her act. If the state were to punish criminal actors with these defenses, it would deny people a fair opportunity to avoid being punished for performing criminal acts with these defenses.⁷⁵ Hence, by punishing them, the state would restrict people's ability to plan on making choices that would safeguard them from being punished.⁷⁶ As a consequence of punishing them, people would incur at least three additional costs of insecurity in response to the possibility of being punished for performing criminal acts with these defenses. First, they would invest in services to help them avoid performing criminal acts with these defenses. Second, they would avoid engaging in valuable activities that would place them at an unduly high risk of performing criminal acts with these defenses. Third, they would fear the possibility of performing criminal acts with these defenses. Because punishing criminal actors with these defenses would result in additional costs of insecurity, the overall costs would outweigh the overall benefits of punishing them.

To illustrate, consider again duress. Criminal actors with a duress defense did not have a fair opportunity to avoid performing their acts because they knew they would have suffered serious harm if they had avoided performing them. If the state were to punish them, it would deny people a fair opportunity to avoid being punished for performing criminal acts under duress. Hence, by punishing them, the state would restrict people's ability to plan on making choices that would safeguard them from being punished. As a consequence of punishing them, people would incur at least three additional costs of insecurity in response to the possibility of being punished for performing criminal acts under duress. First, they would invest more in services to reduce their risk of being put under duress. For example, they might invest more in protective services to help them defend against potential threats, and they might invest more in surveillance schemes to help them avoid interacting with those who are likely to put

75. H.L.A. Hart emphasizes this point in *supra* note 67, at 22–24 and in *Legal Responsibility and Excuses*, in *Punishment and Responsibility: Essays in the Philosophy of Law* 28 (1968). Both Hart and T.M. Scanlon emphasize the importance of providing people with a fair opportunity to avoid being punished. See Scanlon, *supra* note 11, at 263–67.

76. Both Hart and Scanlon emphasize the value of this ability. See their works cited in *supra* note 75.

them under duress. Second, they would avoid more valuable activities that would leave them too vulnerable to duress. Third, they would fear more the possibility of being put under duress. Because punishing criminal actors with a duress defense would result in additional costs of insecurity, the overall costs would outweigh the overall benefits of punishing them.

In response, the fourth argument is unpersuasive for at least two reasons. First, we might concede that punishing criminal actors with the relevant defenses would result in some additional costs of insecurity. However, punishing them might also result in the reduction of other costs of insecurity. As we have noted, punishing criminal actors with defenses might have general deterrence effects. Thus, punishing them might result in the performance of fewer criminal acts and in fewer people being victims of criminal acts. So if the state were to punish criminal actors with the relevant defenses, it might provide more people with a fair opportunity to avoid being victims of criminal acts.⁷⁷ Hence, by punishing them, the state might increase the confidence with which people can plan on not being victims of criminal acts. As a consequence of punishing them, people might reduce the costs of insecurity they incur in response to their lower probability of being victims of criminal acts. First, they might invest less in services aimed at protecting them from being victims of criminal acts. Second, they might engage in valuable activities that would otherwise place them at an unduly high risk of being victims of criminal acts. Third, they might fear less the possibility of being victims of criminal acts. Assuming punishing criminal actors with the relevant defenses would result in a sufficiently large reduction in some costs of insecurity, punishing them might have the benefit of reducing costs of insecurity overall. So the overall benefits might outweigh the overall costs of punishing them.

To illustrate, consider again duress. As we have noted, punishing criminal actors with a duress defense might deter people from performing criminal acts in general and might reduce the incidence of duress as a result of making others more threat resistant. So punishing them might result in fewer people performing criminal acts and fewer people being victims of criminal acts. More specifically, punishing them might result in fewer victims of duress. Thus, if the state were to punish them, it might provide more people with a fair opportunity to avoid being victims of criminal acts

77. See Sanford H. Kadish, *Excusing Crime*, 75 Cal. L. Rev. 257, 263–64 (1987).

in general and duress in particular. Hence, by punishing them, the state might increase the confidence with which people can plan on not being victims of criminal acts. As a consequence of punishing them, people might reduce the costs of insecurity they incur in response to their lower probability of being victims of criminal acts. First, for example, they might invest less in services aimed at reducing their risk of being put under duress or being victims of criminal acts performed under duress. Second, for example, they might engage in more valuable activities that would otherwise leave them too vulnerable to such criminal acts. Third, for example, they might fear less the possibility of being victims of such criminal acts. Assuming that punishing criminal actors with a duress defense would result in a sufficiently large reduction in some costs of insecurity, punishing them might have the benefit of reducing costs of insecurity overall. So the overall benefits might outweigh the overall costs of punishing them.

Second, suppose for the sake of argument that punishing criminal actors with the relevant defenses would increase costs of insecurity overall. Nevertheless, punishing them might still have significant deterrence effects and so significant deterrence benefits. Assuming the deterrence benefits would be sufficiently large, they might outweigh the additional costs of insecurity of punishing them. So the overall benefits might still outweigh the overall costs of punishing them.

E. Fifth Argument

A more general fifth argument might concede that punishing criminal actors with any defense could reduce costs of insecurity overall and have significant deterrence benefits. But according to the argument, the overall costs would still outweigh the overall benefits of punishing them because the punitive costs of punishing them would be extremely high. They would be extremely high because punishing them would be extremely bad. Punishing them would be extremely bad not so much because each would suffer extremely severely in being punished. After all, the punishments need not be extremely severe. Rather punishing them would be extremely bad because they do not deserve to be punished. So punishing them would violate their rights, and violating their rights would be extremely bad considered individually and especially in the aggregate.

In response, the fifth argument is problematic for at least four reasons. First, it merely assumes that criminal actors with defenses deserve no

punishment. However, this assumption stands in need of justification. Given that criminal actors without defenses deserve some punishment, it is not clear why those with defenses deserve no punishment. Because the argument provides no justification for its central assumption, it is unpersuasive.

Second, suppose the argument were to justify its central assumption. Then the argument would be superfluous. If criminal actors with defenses deserve no punishment, then this fact by itself explains why they are not liable to any punishment independently of whether the costs would outweigh the benefits of punishing them. For if they do not deserve any punishment, then punishing them would violate their rights. And the state has a decisive reason not to violate someone's rights regardless of whether the costs would outweigh the benefits of violating her rights. So the state would not be justified in imposing an undeserved punishment on someone regardless of whether the costs would outweigh the benefits of doing so.

Third, suppose again that criminal actors with defenses deserve no punishment. The argument is still problematic because it is not clear that punishing a criminal actor with a defense would be extremely bad, as the argument assumes. So it is not clear that the punitive costs of punishing her would be extremely high, as the argument assumes. The state's punishing her would violate her rights. Hence, given the state's decisive reason not to violate anyone's rights, it has a decisive reason not to punish her. But this fact does not obviously entail that the state's punishing her would be extremely bad.

A person's rights provide others with a decisive agent-relative reason not to violate them. An agent-relative reason is a reason whose description contains an essentially indexical reference to the subject of the reason, whereas an agent-neutral reason is one whose description does not contain such a reference.⁷⁸ Specifically, people's rights provide someone with a decisive agent-relative reason not to violate them herself. But people's rights do not provide someone with a decisive agent-neutral reason to minimize the violations of their rights in general. For example, a person does not have most reason to violate the rights of one in order to prevent another from violating the rights of several others.⁷⁹

78. See Thomas Nagel, *The View from Nowhere* 152–53 (1986) (distinguishing agent-relative from agent-neutral reasons).

79. See Nozick, *supra* note 64, at 30–33 (describing rights as side constraints).

So the state's decisive reason not to punish a criminal actor with a defense is an agent-relative one. As a consequence, the state is warranted for agent-relative reasons in desiring to an extremely strong degree that it not punish her. But again this fact does not obviously entail that the state's punishing her would be extremely bad. For the narrow sense of value at issue is an impartial one.⁸⁰ On this sense, something is good only if everyone is warranted for agent-neutral reasons in desiring that it obtain.⁸¹ Conversely, something is bad only if everyone is warranted for agent-neutral reasons in desiring that it not obtain. More specifically, something is extremely bad only if everyone is warranted for agent-neutral reasons in desiring to an extremely strong degree that it not obtain. The fact, though, that the state is warranted for agent-relative reasons in desiring to an extremely strong degree that it not punish a criminal actor with a defense does not obviously entail that everyone is warranted for agent-neutral reasons in desiring to an extremely strong degree that the state not punish her. Whether there is such an entailment stands in need of a justification that the argument does not provide.

Fourth, suppose for the sake of argument that punishing a criminal actor with a defense would be extremely bad because punishing her would violate her rights. So suppose the punitive costs of punishing criminal actors with defenses would be extremely high considered individually and especially considered in the aggregate. Nevertheless, the deterrence benefits might still outweigh the punitive costs of punishing them. For punishing them might deter people generally from performing criminal acts that would violate the rights of others. As a consequence, punishing them might prevent the violation of the rights of others. If violating someone's rights would be extremely bad, then preventing the violation of someone's rights would be extremely good. Hence, punishing criminal actors with defenses might have the extremely good deterrence effect of preventing people from violating the rights of others. So the deterrence benefits of

80. This sense of value is employed in the standard formulation of act-consequentialism. See, e.g., Samuel Scheffler, *The Rejection of Consequentialism* 1–2 (rev. ed. 1994).

81. I presuppose here that everyone is vividly aware of the things whose value is being assessed. On a more precise restatement of this necessary condition on goodness, something is good only if everyone would be warranted for agent-neutral reasons in desiring that it obtain if she were vividly aware of what it is like. Parallel restatements are available for the following necessary conditions on badness.

punishing them might be extremely high. Thus, the deterrence benefits might outweigh the punitive costs of punishing them even if the latter would be extremely high.

In the final analysis, punishing criminal actors with full defenses would have significant costs, which might even be extremely high. In particular, punishing them would result in significant punitive costs and costs of insecurity. However, punishing them might also have significant benefits, which also might be extremely high. In particular, punishing them might result in significant deterrence benefits. Thus, it is epistemically indeterminate whether the overall costs would outweigh the overall benefits of punishing them. There is no robust reason to believe one way or the other. So the value requirement does not explain the mitigating effects of full defenses on a criminal actor's liability to punishment.

Parallel claims hold with respect to partial defenses. Punishing criminal actors with partial defenses to an unmitigated degree would also have significant additional costs, such as punitive costs and costs of insecurity. However, punishing them to an unmitigated degree might also have significant additional benefits, such as deterrence benefits. Thus, it is epistemically indeterminate whether the net benefit of punishing them to an unmitigated degree would be lower than the net benefit of punishing them to a mitigated degree. So the value requirement also does not explain the mitigating effects of partial defenses on a criminal actor's liability to punishment.

VI. THE DESERT REQUIREMENT

Compared to the value requirement, the desert requirement seems a much more promising basis on which to explain the mitigating effects of defenses on a criminal actor's liability to punishment. Intuitively, a criminal actor with a full defense deserves no punishment, and more generally, a criminal actor with a partial defense deserves to be punished only to a mitigated degree. This intuition, though, stands in need of justification. As we have noted, it is not clear why criminal actors with full defenses deserve no punishment given that criminal actors without them deserve some. More generally, it is not clear why criminal actors with partial defenses deserve to be punished only to a mitigated degree given that criminal actors without them deserve to be punished to an unmitigated degree. To explain

these mitigating effects, I will consider theories of punitive desert that might provide a *prima facie* plausible basis on which to explain them. I will focus primarily on whether the theories can explain the mitigating effects of full defenses. If a theory does not explain them, then it does not explain the mitigating effects of partial defenses either. If a theory does explain them, we will consider whether it also provides a more general explanation of the mitigating effects of partial defenses. Ultimately, I argue that my restorative signaling theory of punitive desert best explains the mitigating effects of both full and partial defenses.

A. A Fairness Theory

A fairness theory might state the following necessary condition of punitive desert: a criminal actor deserves to be punished for her act only if she had a fair opportunity to avoid performing it. If a criminal actor has a defense for her act, then she lacked a fair opportunity to avoid performing it. So criminal actors with defenses do not deserve to be punished.

To illustrate, consider a defense of ignorance or mistake of fact. If a criminal actor has such a defense, she did not believe her act had the properties that made it a criminal act. So she could not have been motivated to avoid performing the act by those properties. And she could not have been deterred from performing the act by any threat to punish her for performing such an act under its description as a criminal act. As a consequence, she lacked a fair opportunity to avoid performing it. Thus, she does not deserve to be punished for it.

For another illustration, consider duress. In a standard case, a criminal actor with a duress defense knew she would have suffered serious harm unless she performed her act. Given the serious harm she would have suffered if she had avoided performing her act, she lacked a fair opportunity to avoid performing it. So she does not deserve to be punished for it.

In response, the fairness theory is open to at least two objections. First, the theory does not identify a necessary condition of defenses. A criminal actor might have a defense even if she had a fair opportunity to avoid performing her act. To illustrate, consider a criminal actor with a defense of consent or defense of others and who knew the state did not recognize such considerations as defenses. She could have known that her act was a criminal one punishable by the state, and she might have known that she could have avoided the act without suffering serious harm as a result. So

she could have had a fair opportunity to avoid performing the act even though she performed it with a defense of consent or defense of others. Thus, the theory is limited in its explanatory power. At best, it could only explain the mitigating effects of a subset of defenses.

Second, the theory does not identify a necessary condition of punitive desert. A criminal actor might deserve to be punished for her act even if she lacked a fair opportunity to avoid performing it. To illustrate, consider a case involving a “counterfactual intervenor.”⁸² Suppose I perform a criminal act with no defenses. Suppose also that throughout the process of my performing it, a bystander secretly observed me ready to intervene whenever I might have expressed any reluctance to perform the act. If I had expressed such reluctance, the bystander would have threatened to harm me seriously unless I performed the act. In this case, I lacked a fair opportunity to avoid performing it. For if I had avoided performing it, I would have suffered serious harm. Nevertheless, I still deserve to be punished for the act.

B. A Second Fairness Theory

A second fairness theory might concede the objections to the first. But it might state a revised necessary condition of punitive desert, and try to explain the mitigating effects of only a subset of defenses. According to the revised condition, a criminal actor deserves to be punished for her act only if (a) she disrespected the rights of others in performing it and (b) she had a fair opportunity to avoid disrespecting the rights of others in

82. I borrow this term from John Martin Fischer & Mark Ravizza, *Responsibility and Control: A Theory of Moral Responsibility* 29–30 (1998) and Copp, *supra* note 54, at 268. As they use the term, a counterfactual intervenor is the potentially intervening mechanism in any Frankfurt-style counterexample to the principle of alternate possibilities. According to the principle, someone is blameworthy for performing an act only if she could have done otherwise. In other words, someone is not blameworthy for performing an act if she literally could not have done otherwise. Although I am not directly engaging this principle here, my counterexamples to the fairness theories of punitive desert are similar in form to a Frankfurt-style counterexample to this principle. I set aside for further analysis, though, the issue of whether this principle is sound and the issue of whether Frankfurt-style counterexamples undermine it. For more discussion on these issues, see, e.g., Harry Frankfurt, *Alternate Possibilities and Moral Responsibility*, 66 *J. Phil.* 829 (1969); Peter van Inwagen, *Ability and Responsibility*, 87 *Phil. Rev.* 201 (1978); and *Moral Responsibility and Alternative Possibilities: Essays on the Importance of Alternative Possibilities* (David Widerker & Michael McKenna eds., 2003).

performing it. If a criminal actor has some range of defenses, then she does not satisfy at least one part of the revised condition. Thus, criminal actors with the relevant defenses do not deserve to be punished. The revised condition of the second theory reflects the idea that ultimately criminal actors deserve to be punished not for performing their acts *per se*, but rather for the way they governed themselves in performing them.⁸³ In particular, they deserve to be punished for disrespecting the rights of others in performing their acts.

The revised condition is not open to the counterexample to the first theory. In that case, I lacked a fair opportunity to avoid performing the criminal act. But I also disrespected the rights of others in performing it, and I had a fair opportunity not to do so because I was free to perform the act with only appropriate motives. For suppose I had expressed reluctance to perform the act, and the bystander had intervened, threatening to harm me seriously unless I performed it. Then I would have been free to perform the act solely in order to avoid the threatened harm; this motivation would have been consistent with my performing the act with an appropriate regard for the rights of others and so with no disrespect for the rights of others. So the revised condition does not unacceptably entail that I do not deserve to be punished for performing the act.

In response, the second theory does not identify a necessary condition of punitive desert. A criminal actor might deserve to be punished for her act even if she lacked a fair opportunity to avoid disrespecting the rights of others in performing it. To illustrate, consider another case involving a counterfactual intervenor. Suppose again that I perform a criminal act with no defenses. So I disrespected the rights of others in performing it. Suppose also that throughout the process of my performing the act, a bystander secretly observed me ready to intervene whenever I might have expressed any reluctance to perform it. If I had expressed such reluctance, the bystander would have injected into me a mind altering drug that would have diminished my capacity to respond appropriately to the moral reasons against violating the rights of others. As a consequence of the drug, I would have performed the act, and I would have disrespected the rights of others in performing it. In this case, I lacked a fair opportunity to avoid not only performing the act, but also disrespecting the rights of

83. See Scanlon, *supra* note 11, at 268–69.

others in performing it. For if I had tried to avoid performing the act, I would have been injected with a drug that would have caused me not only to perform it, but also to disrespect the rights of others in performing it. Nevertheless, I still deserve to be punished for the act.

C. A Third Fairness Theory

A third fairness theory might concede the objections to the first two. But it might state another necessary condition of punitive desert and again try to explain the mitigating effects of only a subset of defenses. According to the condition, a criminal actor deserves to be punished for her act only if she had a fair opportunity to avoid being punished for it. If a criminal actor has some range of defenses, and the state were to punish her for the act, then she would not have had a fair opportunity to avoid being punished for it. Thus, criminal actors with the relevant defenses do not deserve to be punished.

The revised condition might not be open to the two previous counterexamples in which I perform a criminal act with no defenses. In the first case, suppose the state punishes me for performing the criminal act under conditions in which the state recognizes a defense of duress as a mitigating factor. Then I had a fair opportunity to avoid being punished for the act. For if I had tried not to perform the act, the bystander would have put me under duress. And if I had performed the act under duress, I would have had a duress defense for it, and the state would not have punished me for it. In the second case, suppose the state punishes me for performing the criminal act under conditions in which the state recognizes a defense of intoxication as a mitigating factor. Then I had a fair opportunity to avoid being punished for the act. For if I had tried not to perform the act, the bystander would have caused me to become intoxicated by injecting the drug into me. And if I had performed the act while intoxicated, I would have had an intoxication defense for it, and the state would not have punished me for it. So in neither case would the revised condition unacceptably entail that I do not deserve to be punished for the act.

In response, there might be significant costs to the state's adopting or not adopting the revised condition as a constraint on the conditions under which it punishes criminal actors. On the one hand, if the state were not to adopt the constraint, it would restrict people's ability to plan on making choices that would safeguard them from being punished. So as we

have discussed, people might incur additional costs of insecurity in response to the possibility of being punished for performing criminal acts with the relevant defenses. On the other hand, if the state were to adopt the constraint, it would restrict people's ability to plan on making choices that would safeguard them from being victims of criminal acts. Punishing criminal actors with the relevant defenses would have a general deterrence effect that would be absent under the constraint. If the state were not to punish criminal actors with the relevant defenses, it would provide fewer people with a fair opportunity to avoid being victims of criminal acts. So as we have discussed, people might incur additional costs of insecurity in response to their higher probability of being victims of criminal acts.

But whether or not the benefits would outweigh the costs of adopting the revised condition as a constraint, it is not a necessary condition of punitive desert. A criminal actor might deserve to be punished for her act even if she lacked a fair opportunity to avoid being punished for it. To illustrate, consider variants of the previous cases. In the first, suppose the state punishes me for performing the criminal act under conditions in which the state does not recognize a defense of duress as a mitigating factor. Then I lacked a fair opportunity to avoid being punished for it. For if I had tried not to perform the act, the bystander would have put me under duress. And if I had performed the act under duress, the state would have punished me for it anyway. In the second case, suppose the state punishes me for performing the criminal act under conditions in which the state does not recognize a defense of intoxication as a mitigating factor. Then I also lacked a fair opportunity to avoid being punished for the act. For if I had tried not to perform it, the bystander would have caused me to become intoxicated by injecting the drug into me. And if I had performed the act while intoxicated, the state would have punished me for it anyway. So in both cases, the revised condition unacceptably entails that I do not deserve to be punished for the act.

In clarification, there is something objectionable about the state in each case. But what is objectionable is not the state's punishing me for my criminal acts or imposing on me an undeserved punishment. Given that I performed the acts with no defenses, I deserved to be punished for them even though I lacked a fair opportunity to avoid being punished for them because the state does not recognize the relevant defenses. Rather what is objectionable is precisely the state's policy of not recognizing these defenses as mitigating factors. The policy is objectionable because criminal

actors with these defenses do not deserve to be punished. The claim, though, that criminal actors with defenses do not deserve to be punished is precisely the claim that stands in need of justification.

D. A Fourth Fairness Theory

A fourth fairness theory might state a related principle of punitive desert drawing on the content of the previous fairness theories. According to the fourth principle, a criminal actor deserves to be punished for her act only if she believed that she had a fair opportunity to avoid (a) performing the act, (b) disrespecting the rights of others in performing the act, or (c) being punished for the act. If a criminal actor has some range of defenses, and the state punishes her for the act, then she believed that she lacked a fair opportunity to avoid the principle's three conditions. Therefore, criminal actors with the relevant defenses do not deserve to be punished.

The fourth principle is not open to any of the previous counterexamples. In those cases, I performed the criminal act unaware of the counterfactual intervenor. I did not believe there was a bystander who would have put me under duress or injected a drug into me if I had expressed any reluctance to perform the act. Hence, in each of the previous cases, I believed that I had a fair opportunity to avoid (a) performing the criminal act, (b) disrespecting the rights of others in performing it, and (c) being punished for it. Thus, the fourth principle does not unacceptably entail that I do not deserve to be punished for the criminal act.

In response, the fourth principle is also not a sound necessary condition of punitive desert. A criminal actor might deserve to be punished even if she believed that she lacked a fair opportunity to avoid all the conditions specified in the fourth principle. To illustrate, consider another variant of our previous cases involving a counterfactual intervenor. Suppose I am a misanthrope, and I go on a crime spree, committing ten crimes in succession, one immediately after the other. I commit each crime merely to cause others harm. Before I commit the fifth crime, though, I notice a bystander. I know that if I show any reluctance to perform the fifth criminal act, the bystander will inject a drug into me that will cause me not only to perform the criminal act, but also to disrespect the rights of others in performing it. I also know that the state does not recognize an intoxication defense, and the state will inevitably punish me for performing the fifth criminal act. I choose to perform the fifth criminal

act, however, merely to cause others harm and with no regard for the rights of others. My beliefs about the bystander were motivationally inert, as evidenced by the fact that I committed several similar crimes immediately before and after the fifth one outside the gaze of the bystander. Even if I had been unaware of the bystander's presence, I would have been motivated to perform the fifth criminal act just as strongly as I was actually motivated to perform it. As a consequence, I have no defense for the fifth criminal act. I deserve to be punished for the act even though I believed that I lacked a fair opportunity to avoid (a) performing it, (b) disrespecting the rights of others in performing it, and (c) being punished for it.

In the final analysis, the fairness theories and their potential variants are problematic for at least two reasons. First, it is doubtful that any fair opportunity of avoidance principle is a necessary condition of punitive desert. Such principles seem open to counterexamples involving counterfactual intervenors or nonstandard punitive policies involving strict liability. So it is doubtful that such principles could explain why criminal actors with any full defenses deserve no punishment. Second, even if such a principle is a necessary condition of punitive desert, its explanatory power would likely be limited. As we have seen, it is doubtful that such principles could explain why criminal actors with some full defenses do not deserve to be punished. For criminal actors with some full defenses, such as consent and defense of others, seem capable of having a fair opportunity to avoid any relevant consequence. It is also doubtful that such principles could explain why criminal actors with mere partial defenses deserve to be punished only to a mitigated degree. Criminal actors with some partial defenses might have only a partially unfair opportunity to avoid some relevant consequence. But it is not clear how such degrees of unfairness could explain why they deserve to be punished only to a mitigated degree.

E. An Expressive Theory

An expressive theory might point out that punishing someone expresses an attitude of moral blame toward her.⁸⁴ Such an attitude might be resentment or indignation. Given the expressive aspect of punishment, an expressive

84. See Joel Feinberg, *The Expressive Function of Punishment*, in *Doing and Deserving: Essays in the Theory of Responsibility* 98 (1970).

theory might state the following necessary condition of punitive desert: a criminal actor deserves a punishment only if the punishment would not express an unwarranted attitude of moral blame toward her. So a criminal actor deserves a punishment only if the punishment would not express too much blame toward her. Punishing criminal actors with full defenses would express too much blame toward them because punishing them would express some blame toward them, and they are not at all blameworthy. Thus, they do not deserve to be punished. More generally, punishing criminal actors with partial defenses to an unmitigated degree would express too much blame toward them because punishing them to an unmitigated degree would express an unmitigated degree of blame toward them, and they are blameworthy only to a mitigated degree. Hence, they do not deserve to be punished to an unmitigated degree.

In response, the theory is open to at least two objections. First, it merely assumes that criminal actors with full defenses are not at all blameworthy, and those with partial defenses are blameworthy only to a mitigated degree. But these claims stand in need of justification. Given that criminal actors without full defenses are blameworthy to some degree, it is not clear why those with full defenses are not also blameworthy to some degree. And given that criminal actors without partial defenses are blameworthy to an unmitigated degree, it is not clear why those without partial defenses are not also blameworthy to an unmitigated degree. Because the theory provides no justification for these assumptions, it is unpersuasive.

Second, the theory is circular. On a standard view, attitudes of moral blame consist at least partly in certain demands.⁸⁵ To feel an attitude of moral blame toward someone is to make or feel certain demands on her. So to express an attitude of moral blame toward someone is to express certain demands on her. One demand that is constitutive of an attitude of moral blame is the demand to undertake a punishment or, more precisely, the burdens constitutive of a punishment.⁸⁶ Punishing someone expresses

85. See Darwall, *supra* note 26, at 17; Strawson, *supra* note 11, at 14–15, 21–22; Watson, *supra* note 11, at 121, 126–28.

86. Attitudes of moral blame are sentiments of justice. See J.S. Mill, *Utilitarianism* 87–107 (Oxford 1998). As such, they are a subset of all attitudes of moral disapproval. I do not claim that a demand to undertake a punishment is constitutive of all the latter. I only claim that such a demand is constitutive of moral blame. See *id.* at 93, 95 (claiming that attitudes of moral blame as sentiments of justice involve both a desire to punish their objects

an attitude of moral blame toward her by expressing its constitutive demand on her to undertake the punishment. So to claim that a punishment expresses an unwarranted attitude of moral blame toward someone is to claim that the punishment expresses an unwarranted demand on her to undertake the punishment. And to claim that a demand on someone to undertake a punishment is unwarranted presupposes that she does not deserve the punishment.⁸⁷ As a consequence, the expressive theory is circular in virtue of the fact that it presupposes what it aims to explain.⁸⁸

To clarify, consider criminal actors with full defenses. The theory tries to explain why they do not deserve any punishment on the assumption that any punishment would express too much blame toward them. But to claim that any punishment would express too much blame toward them is to claim that any punishment would express an unwarranted demand on them to undertake the punishment. And to claim that a demand on them to undertake any punishment would be unwarranted presupposes that they do not deserve any punishment. The claim, though, that they do not deserve any punishment is precisely the claim to be explained. Consider now criminal actors with partial defenses. The theory tries to explain why they do not deserve an unmitigated punishment on the assumption that such a punishment would express too much blame toward them. But to claim that an unmitigated punishment would express too much blame toward them is to claim that such a punishment would express an unwarranted demand on them to undertake the punishment.

and the judgment that their objects can be properly punished); cf. Thomas Baldwin, *Punishment, Communication, and Resentment*, in *Punishment and Political Theory* 124, 128, 130, 132 (Matt Matravers ed., 1999) (claiming that resenting someone involves the judgment that she should be punished); Joseph Butler, *Sermon VIII: Upon Resentment*, in *The Works of Joseph Butler* 141 (W.E. Gladstone ed., 1896) (suggesting that attitudes of resentment and indignation toward someone involve a desire that she be punished).

87. The considerations that determine whether others are warranted in demanding someone to undertake a punishment are narrower than the considerations that determine whether others have most reason to desire to demand her to undertake the punishment. The relevant considerations are the ones that determine whether she is obligated to others to undertake the punishment. Cf. Copp, *supra* note 54, at 271–75; Darwall, *supra* note 26, at 96–99. For related discussion on the distinction between the considerations that make an attitude warranted, fitting, or rational and the considerations that make an attitude desirable, see the works cited in *supra* note 26.

88. Cf. Baldwin, *supra* note 86, at 130, 132 (expounding a similar circularity objection to attempts to justify punishing someone on the grounds that punishing her would express a warranted attitude of resentment toward her).

And to claim that a demand on them to undertake an unmitigated punishment would be unwarranted presupposes that they do not deserve such a punishment. The claim, though, that they do not deserve an unmitigated punishment is precisely the claim to be explained. Because the theory presupposes the claims it aims to explain, it is circular.

F. A Restorative Signaling Theory

My restorative signaling theory, RS, assumes that someone deserves a punishment for an act if and only if her undertaking the punishment is necessary to fulfill an obligation she incurs from performing the act.⁸⁹ If the latter condition obtains, she deserves the punishment for three related reasons. First, she is obligated to undertake the punishment. For in general, if someone has an obligation to do something, she has a derivative obligation to undertake the means necessary to her doing it.⁹⁰ Second, she is obligated to consent to undertake the punishment.⁹¹ In general, if someone

89. For a full defense of RS, see my *A Restorative Signaling Theory of Punitive Desert* (2008) (unpublished Ph.D. dissertation, University of Michigan) (on file with the University of Michigan Philosophy Department).

90. This is a transmission principle concerning obligations. For similar transmission principles concerning "oughts," see Mark Schroeder, *Means-end Coherence, Stringency, and Subjective Reasons*, *Phil. Stud.* (forthcoming).

91. Since RS is grounded in a principle of consent, it could be considered a consent based theory of punitive desert. However, RS is not grounded in a principle of actual consent. It does not assume that criminals deserve to be punished because they actually consent to undertake a punishment. Compare C.S. Nino, *A Consensual Theory of Punishment*, 12 *Phil. & Pub. Aff.* 289, 297–300 (1983) (arguing that criminals actually consent to forfeit their moral rights against being punished for their crimes because in committing them, they knowingly forfeit their legal rights against being punished for them), with Larry Alexander, *Consent, Punishment, and Proportionality*, 15 *Phil. & Pub. Aff.* 178 (1986) (criticizing Nino's consent based theory of punitive desert) and T.M. Scanlon, *Punishment and the Rule of Law*, in *The Difficulty of Tolerance: Essays in Political Philosophy* 219 (2003) (same). Even if criminals do not actually consent to undertake a punishment, RS assumes they deserve to be punished because they are obligated to consent to undertake a punishment. On one interpretation, the fact that criminals are obligated to consent to undertake a punishment makes punishing them even against their will consistent with Kant's categorical imperative to treat others as ends in themselves and never as mere means to an end. See Immanuel Kant, *Groundwork of the Metaphysics of Morals* 4:433 (Mary Gregor ed., 1997); cf. Derek Parfit, *Climbing the Mountain* chs. 4–5 (forthcoming) (discussing interpretations of this formulation of Kant's categorical imperative and its relation to various principles of consent).

is obligated to undertake something or must undertake something to fulfill an obligation, then she has an obligation to consent to undertake it. Third, she would be unjustly enriched by not undertaking the punishment.⁹² For by not undertaking the punishment, she would not only violate her obligation to others, but also obtain a benefit from the violation consisting in her freedom from the burdens necessary to fulfill the obligation. In general, if someone obtains a benefit from violating an obligation to others, the benefit constitutes an unjust enrichment. Each of these related claims entails that she has no right to be free from the punishment.⁹³ So if she refuses to undertake the punishment voluntarily, the state would not violate her rights by imposing the punishment on her against her will.

Given its basic principle, RS explains why criminals without defenses deserve to be punished by explaining why they must undertake a punishment to fulfill an obligation they incur from committing their crimes.

92. Insofar as an unjust enrichment constitutes an unfair advantage, RS could be considered a type of unfair advantage theory of punitive desert. For examples of such theories, see Herbert Morris, *Persons and Punishment*, in *Theories of Punishment* 76, 79 (Stanley E. Grupp ed., 1971); Jeffrie Murphy, *Marxism and Retribution*, 2 *Phil. & Pub. Aff.* 217, 228–29 (1973); George Sher, *Desert* 69–90 (1987); Richard Dagger, *Playing Fair with Punishment*, 103 *Ethics* 473 (1993). Unlike these particular theories, though, RS does not assume that criminals obtain a benefit from committing their crimes. They might not. Rather RS assumes that criminals stand to obtain a benefit from violating an obligation they incur from committing their crimes. Relatedly, RS does not assume that the wrong for which criminals deserve to be punished consists in their obtaining an unfair advantage or an unjust enrichment. Rather RS assumes that the wrong consists in their disrespecting the rights of others. Also unlike these particular theories, RS can account for the proportionality of punitive desert, as I argue below. See Richard Burgh, *Do the Guilty Deserve Punishment?*, 79 *J. Phil.* 193, 209–10 (1982) (criticizing other unfair advantage theories of punitive desert on the grounds that they cannot account for its proportionality); David Dolinko, *Some Thoughts about Retributivism*, 101 *Ethics* 537, 545–49 (1991) (same); Matt Matravers, *Justice and Punishment: The Rationale of Coercion* 45–72 (2000) (same).

93. I take this point to be basic for the purposes of this paper. In general, if someone is obligated to undertake a burden or is obligated to consent to undertake it or would be unjustly enriched by not undertaking it, she has no right to be free from it. In other words, under any one of these conditions, she has no claim or entitlement to be free from the burden. Cf. Mill, *supra* note 86, at 93 (stating that “Duty is a thing which may be *exacted* from a person, as one exacts a debt. . . . Reasons of prudence, or the interest of other people, may militate against actually exacting it; but the person himself . . . would not be entitled to complain.”); Joel Feinberg, *Duties, Rights, and Claims*, in *Rights, Justice, and the Bounds of Liberty: Essays in Social Philosophy* 130 (1980) (analyzing rights in terms of claims).

According to RS, a person has an obligation of trust not to undermine the conditions that are necessary for others' being justified in believing that she is not disposed to commit crimes. Call such conditions "conditions of trust." A person's obligation of trust is grounded in the fact that she would standardly cause others to incur at least three costs of insecurity by undermining conditions of trust.⁹⁴ First, if someone undermines them, others rationally must invest in costly precautionary measures to protect themselves from her. For example, others might need to engage in costly surveillance schemes to reduce their interactions with her and to employ costly protective services when interacting with her is unavoidable. Second, if someone undermines conditions of trust, others rationally must forgo pursuing some personally and socially valuable activities that would invariably leave them too vulnerable to her. Third, if someone undermines conditions of trust, others will rationally experience fear in response to their higher subjective probability of her committing crimes against them. To avoid imposing such costs on others, a person has not only the obligation of trust, but also an obligation of restoration. According to the obligation of restoration, if someone undermines conditions of trust, she is obligated to restore those conditions as quickly as is reasonably possible by demonstrating to others that she is no longer disposed to commit crimes. She must restore them as quickly as is reasonably possible because the longer she takes to restore them the longer others must incur the costs of insecurity.⁹⁵

When someone commits a crime with no defenses, she undermines conditions of trust.⁹⁶ Her crime is strong evidence that she is disposed to commit crimes; therefore, others are not justified in believing otherwise.⁹⁷ So she is

94. Cf. Thomas Hobbes, *Leviathan* ch. 13 (Richard Tuck ed., Cambridge 1996) (noting the costs of insecurity someone rationally must incur in response to being justified in believing that others are disposed to engage in acts of aggression).

95. To be concise, I leave the quickness condition implicit in my following discussion of the obligation of restoration.

96. David Hoekema and Susan Dimock also point out that criminals undermine conditions of trust by committing their crimes. See Susan Dimock, *Retributivism and Trust*, 16 *Law & Phil.* 37 (1997); David A. Hoekema, *Trust and Obedience: Toward a New Theory of Punishment*, 25 *Isr. L. Rev.* 332 (1991).

97. Some crimes under some descriptions are literally unrepeatable, like matricide or patricide. But even these undermine conditions of trust because even these are strong evidence that their perpetrator is disposed to commit crimes. For when someone commits a crime without a defense, her specific crime is strong evidence that she is disposed to commit

obligated to restore the conditions of trust she undermined by committing her crime. Given her obligation of restoration, she is obligated to undertake any means that are necessary to fulfill it. Hence, she is obligated to undertake any burdens that are necessary to fulfill it. So assuming she must undertake certain burdens to fulfill her obligation of restoration, the state may impose them on her as a punishment against her will without violating her rights. According to the main principle of RS, a criminal deserves a punishment for her crime that is no more severe than the burdens she must undertake to fulfill the obligation of restoration she incurs from committing her crime. In other words, a criminal deserves to be punished for her crime no more severely than the burdens she is obligated to undertake to restore the conditions of trust she undermined by committing her crime.⁹⁸

A criminal with no defenses must in fact undertake some burdens to fulfill her obligation of restoration. By committing her crime, she undermined conditions of trust by providing others with strong evidence that she is disposed to commit crimes. To fulfill her obligation of restoration,

a broader range of crimes comparable in seriousness. Two findings, among others, support the broader inference.

First, there is a high rate of recidivism among criminals: a high percentage of criminals commit multiple crimes at different times. See, e.g., Patrick A. Langan & David J. Levin, U.S. Dep't Just., *Recidivism of Prisoners Released in 1994*, available at <http://www.ojp.usdoj.gov/bjs/pub/pdf/rpr94.pdf> (among 300,000 prisoners released in 15 U.S. states, 67.5% were rearrested for a new offense, and 46.9% were reconvicted for a new crime within 3 years of their release); Michael R. Gottfredson & Travis Hirschi, *A General Theory of Crime* 107–08, 177, 230–31, 253 (1990) (noting the high stability of criminals' dispositions to commit crimes, and citing numerous research studies in support).

Second, there is a high rate of versatility or, in other words, a low rate of specialization among criminals: a high percentage of criminals commit multiple types of crimes. See, e.g., Gottfredson & Hirschi, *supra*, at 91–94, 256, 266 (discussing the high rate of versatility among criminals, and noting numerous research studies in support); Chester L. Britt, *Versatility*, in *The Generality of Deviance* 173 (Travis Hirschi & Michael R. Gottfredson eds., 1994) (same); Leonore M.J. Simon, *Do Criminal Offenders Specialize in Crime Types?*, 6 *Applied & Preventive Psychol.* 35 (1997) (same, noting the high rate of versatility even among white collar criminals, sex offenders, and those who commit crimes of domestic violence).

98. As a corollary, a criminal does not deserve a punishment for her crime that is more severe than the burdens she must undertake to fulfill the obligation of restoration she incurs from committing her crime. In other words, a criminal does not deserve to be punished for her crime more severely than the burdens she is obligated to undertake to restore the conditions of trust she undermined by committing her crime.

she must annul such evidence by providing others with strong countervailing evidence that she is no longer disposed to commit crimes. To provide others with such countervailing evidence, she must demonstrate that she has a good will.⁹⁹ Assuming she has a good will, she has a stable disposition not to commit crimes even under conditions in which she could commit them without fear of being detected and punished.

To demonstrate that she has a good will, a criminal must demonstrate that she has a stable disposition to be appropriately motivated by the moral reasons against committing crimes. Hence, she must demonstrate that she has a stable disposition to be appropriately motivated by the moral reasons against violating the rights of others. To demonstrate this, she must demonstrate that she has a stable disposition to care highly about the interests of others.¹⁰⁰ Such a stable disposition to care is strong evidence of a stable disposition to respect the rights of others. For the moral reasons against violating the rights of others are grounded in the interests of others. To demonstrate that she has such a disposition to care, she must demonstrate that she has acted with a sufficiently high degree of benevolence for a sufficiently long time after committing her crime.¹⁰¹ For the fact that someone acts with a sufficiently high degree of benevolence for

99. As Annette Baier states, when we trust others, we are confident that they have a good will toward us; therefore, "reasonable trust will require grounds for such confidence in another's good will. . . ." Annette Baier, *Trust and Antitrust*, 96 *Ethics* 231, 235 (1986).

100. Two points of clarification are in order. First, the attitude of care at issue is one of respectful care. It is an attitude of caring about others only in ways that respect them as persons. See Darwall, *supra* note 26, at 126–30 (distinguishing an attitude of respect from an attitude of mere care). Second, I mean only a thin sense of care, which might consist in an austere attitude of other regard. In this sense, for someone to care about the interests of others just is for her to place weight on their interests in deliberating about what to do. In this sense, for someone to care highly about the interests of others just is for her to weight their interests to a sufficiently high degree relative to her own in deliberating about what to do. This thin sense of care does not have some of the implications of a thicker sense. For example, in the thin sense, someone might care highly about the interests of others without being disposed to grieve when they are harmed or rejoice when they are benefited. Cf. Lawrence A. Blum, *Friendship, Altruism and Morality* 12–15, 146–49 (1980) (discussing a thicker sense of care and other altruistic emotions in which these attitudes do involve dispositions to respond toward their objects with particular feelings).

101. I refer here only to a thin sense of benevolence that consists in caring about the interests of others in the same thin sense I spell out in *supra* note 100. In this sense, a high degree of benevolence involves weighting the interests of others to a sufficiently high degree relative to one's own in deliberating about what to do.

a sufficiently long time is strong evidence that she has a stable disposition to care highly about the interests of others.¹⁰² To demonstrate that she has acted with such benevolence, she must sacrifice some of her sufficiently important personal interests for a sufficiently long time for the sake of benefiting others. To make such a sacrifice for others, the criminal must standardly engage in labor-intensive community service, and she must standardly do so while incapacitated in some form to mitigate the costs of insecurity that others rationally must incur during the interim.¹⁰³ Thus, to fulfill her obligation of restoration, a criminal must undertake some burdens.¹⁰⁴

At this point, we have shown that RS explains why criminals with no defenses deserve some punishment for their crimes. RS also explains how much they deserve to be punished. According to RS, the absolute severity of the most severe punishment that a criminal deserves for her crime corresponds to the absolute severity of the burdens she must undertake to fulfill her obligation of restoration.¹⁰⁵ In general, the absolute severity of such burdens does correspond to the absolute severity of the most severe punishment that a criminal with no defenses seems to deserve for her crime.

To illustrate, consider someone who commits a moderately serious crime with no defenses and, thus, disrespects the rights of others to a moderately bad degree in committing it. Her crime is strong evidence that she has a moderately bad disposition to commit crimes. Thus, her crime is strong evidence that she has a moderately bad disposition to disrespect the

102. See, e.g., Michael Bacharach & Diego Gambetta, *Trust in Signs*, in *Trust in Society* 148, 154 (Karen S. Cook ed., 2001) (noting benevolence as a trust-warranting property).

103. For this reason, punishments under RS should ideally take the form of labor-intensive community service performed under reasonable conditions of incapacitation.

104. Like other trust building or maintaining processes, the process of a criminal's fulfilling her obligation of restoration has a "multi-layered inferential structure." *Id.* at 162. In addition to undertaking the required burdens, other steps might also be necessary to restore the conditions of trust, such as apologizing for the crime and compensating any victims. The criminal also might need to undergo some form of therapy and take steps to eliminate aspects of her situation that pressure her to commit crimes, such as unemployment and corrupting social influences. Much will depend on the specifics of the case. Because these other steps are not necessarily burdensome for the criminal, they need not be part of her punishment.

105. See the Appendix for determining the precise severity of the burdens required to fulfill one's obligation of restoration.

rights of others. So her crime is strong evidence that there is a moderately bad deficiency in the degree to which she cares about others. On reflection, to annul this evidence and demonstrate that she has a stable disposition to care highly about others, she must demonstrate that she has acted with a moderately high degree of benevolence for a moderately long time after committing her crime. To demonstrate this, she must sacrifice some of her moderately important personal interests for a moderately long time for the sake of benefiting others. So she must undertake no more than a moderately severe burden to restore the conditions of trust she undermined by committing her crime. Hence, she deserves no more than a moderately severe punishment for her crime. By parity of reasoning, a mildly serious criminal with no defenses deserves no more than a mildly severe punishment for her crime, and an extremely serious criminal with no defenses deserves an extremely severe punishment for hers.

Given RS's explanation of why and how much criminals with no defenses deserve to be punished, we can see its explanation of how much criminal actors presumably deserve to be punished for their acts. According to RS, a criminal actor deserves to be punished for her act no more severely than the burdens she must undertake to fulfill the obligation of restoration she incurs from performing her act. When someone performs a criminal act, there is a presumption that she committed a particularly serious crime in performing it. Thus, she presumably disrespected the rights of others to a particularly bad degree in performing it. As a consequence, she presumably has a particularly bad disposition to commit crimes. So she is presumably disposed to flout a relevant class of the moral reasons against violating the rights of others. And so presumably there is a particularly bad deficiency in the degree to which she cares about the interests of others. Thus, there is a presumption that by performing her act, she undermined conditions of trust to a particularly bad degree. On the assumption that she has the capacity to restore those conditions of trust, she is presumably obligated to do so. Assuming she has an obligation of restoration, there is a presumption that she must undertake certain burdens to fulfill it. Hence, she presumably deserves to be punished for her act as severely as those burdens, but no more severely. Given RS's explanation of how much criminal actors presumably deserve to be punished for their acts, we can now see how it explains the mitigating effects of the main types of defenses.

1. Type 1 Defenses

Consider type 1 defenses, which mitigate how badly a criminal actor disrespected the rights of others in performing her act. Suppose a criminal actor has a full type 1 defense. Then she did not commit a crime in performing her act. Although she might have violated the rights of others, she did not disrespect their rights in performing the act. For example, she might have been caused to perform the act against her will by mere force; she might not have believed her act would have the type of consequences that make it a criminal act; or she might have been appropriately motivated to perform the act by her belief that particular considerations obtained which (a) entailed that her performing it would not violate the rights of others or (b) provided her with most reason to perform it. As a consequence, the act is not strong evidence that she is disposed to commit crimes. Thus, she did not undermine any conditions of trust by performing the act. So she does not incur an obligation of restoration from performing it. Therefore, she does not deserve to be punished for the act.

To illustrate, consider a criminal actor with a full type 1 defense of duress. Suppose she commits a theft in response to another's threat to kill her unless she commits it. In this case, her theft is not a crime. Although she might have violated the rights of the theft victim, she did not disrespect his rights because her reasons in favor of committing the theft outweighed her reasons against.¹⁰⁶ As a consequence, her theft is not strong evidence that she is disposed to commit crimes. So she did not undermine any conditions of trust by committing the theft, and so she does not incur an obligation of restoration from committing it. Therefore, she does not deserve to be punished for committing the theft.

Suppose a criminal actor has a mere partial type 1 defense. Then although she committed a crime in performing her act, the crime was not as serious as presumed. Thus, the defense mitigates the seriousness of the crime she committed in performing the act. It mitigates how badly she

106. I assume a person can have agent-relative reasons to care about her own interests out of proportion to the interests of others. See, e.g., Nagel, *supra* note 78, at 171-75; Scheffler, *supra* note 80, at 20; Bernard Williams, *A Critique of Utilitarianism*, in *Utilitarianism: For and Against* 75, 108-18 (J.J.C. Smart & Bernard Williams eds., 1973). So I assume someone might have most reason not to resist a threat under duress even though the indirect effects of not resisting the threat would make doing so non-optimific for the reasons we discussed earlier.

disrespected the rights of others in performing the act. So the defense mitigates the badness of the disposition to commit crimes that others are justified in believing she has on the basis of her act. And so the defense mitigates the deficiency in the degree to which she cares about others that they are justified in believing she has on the basis of her act. In sum, the defense mitigates how badly she undermined conditions of trust by performing the act. As a consequence, the defense mitigates the severity of burdens she must undertake to restore the conditions of trust she undermined by performing the act. So it mitigates the severity of burdens she must undertake to fulfill her obligation of restoration. Therefore, the defense mitigates how much she deserves to be punished for the act.

2. Type 2 Defenses

Consider type 2 defenses, which block the otherwise justified inference from the fact that someone performed a criminal act to her having the presumed disposition to commit crimes at the time of assessment. Suppose a criminal actor has a full type 2 defense. Then she might have committed a crime in performing her act. But given the defense, the act is not strong evidence that she is disposed to commit crimes at the time of assessment. For example, she might have performed the act under conditions that would standardly cause a temporary radical distortion in an agent's system of normative self-governance. Thus, she did not undermine any conditions of trust at the time of assessment by performing her act. So at the time of assessment, she does not incur an obligation of restoration from performing the act. Therefore, she does not deserve to be punished for the act.

To illustrate, consider a criminal actor with a full defense of somnambulism. Suppose she intentionally killed others while she was sleepwalking.¹⁰⁷ In this case, she did commit a very serious crime in killing them. She disrespected their rights very badly in doing so. Nevertheless, her killing them is not strong evidence that she is disposed to commit crimes at the time of assessment, when she is awake. For she killed the others under conditions, namely a particular state of sleep, that would standardly cause a temporary radical distortion in an agent's system of normative

107. For an actual case involving homicidal somnambulism, see *The Queen v. Parks*, [1992] 2 S.C.R. 871 (Canada).

self-governance.¹⁰⁸ Thus, she did not undermine any conditions of trust at the time of assessment by killing the others. So at the time of assessment, she does not incur an obligation of restoration from killing them. Hence, she does not deserve to be punished for the crime.

Suppose a criminal actor has a mere partial type 2 defense. Then she might have committed a crime in performing the act, and the crime might have been as serious as presumed. She might have disrespected the rights of others to the presumed degree in performing the act. Nevertheless, the defense still mitigates the badness of the disposition to commit crimes that others are justified in believing she has at the time of assessment on the basis of her act. Thus, the defense mitigates how badly she undermined conditions of trust at the time of assessment by performing her act. So it mitigates the severity of burdens she must undertake to restore the conditions of trust she undermined by performing the act. And so it mitigates the severity of burdens she must undertake to fulfill her obligation of restoration at the time of assessment. Therefore, the defense mitigates how much she deserves to be punished for the act.

3. Type 3 Defenses

Consider type 3 defenses, which are considerations that undermine a criminal actor's capacity to respond appropriately to the relevant class of moral reasons at the time of assessment. Suppose a criminal actor has a full type 3 defense. Then she might have committed a crime in performing her act, and the act might be strong evidence that she is disposed to commit crimes at the time of assessment. So the act might be strong evidence that she is disposed to flout a class of moral reasons against violating the rights of others. Thus, she might have undermined conditions of trust by performing her act.

Nevertheless, even if the criminal actor did undermine conditions of trust, she is not obligated to restore them. For a person is obligated to do something only if others can fairly demand her to do it.¹⁰⁹ And others cannot fairly demand someone to do something that she lacks the capacity to

108. Cf. Rosalind Cartwright, *Sleepwalking Violence: A Sleep Disorder, a Legal Dilemma, and a Psychological Challenge*, 161 *Am. J. Psychiatry* 1149, 1150 (2004) (describing the state of sleep in which sleepwalking violence standardly occurs).

109. Cf. Copp, *supra* note 54, at 271–75; Darwall, *supra* note 26, at 96–99.

do.¹¹⁰ So a person is obligated to do something only if she has the capacity to do it.¹¹¹ If a criminal actor has a full type 3 defense for her act, she lacks the capacity to respond appropriately to any of the relevant moral reasons that she is presumably disposed to flout. Hence, she lacks the capacity to restore any of the conditions of trust she undermined by performing her act. Assuming others know she lacks the capacity to respond appropriately to the relevant moral reasons, there is nothing she could do to justify others in believing that she is disposed to respond appropriately to them. Because she lacks the capacity to restore any of the conditions of trust she undermined by performing her act, she is not obligated to restore any of them. As a consequence, she does not incur an obligation of restoration from performing her act. Therefore, she does not deserve to be punished for the act.

To illustrate, suppose someone commits a theft with a full type 3 defense of kleptomania, where kleptomania is an incapacity to respond appropriately to the moral reasons against committing theft. In this case, her theft is a crime. She disrespected the rights of others in committing it. Given that she committed the theft as a kleptomaniac, her theft is strong evidence only that she is disposed to commit crimes of theft.¹¹² So her theft is strong evidence only that she is disposed to flout the moral reasons against committing theft. Thus, she did undermine some conditions of trust by committing her crime.

Nevertheless, she is not obligated to restore those conditions of trust to any degree because she does not have the capacity to do so. To restore them to any degree, she must have the capacity to demonstrate to others that she is disposed to respond appropriately to the moral reasons against committing theft. To have this capacity, she must have the capacity to respond appropriately to such reasons. But as a kleptomaniac, she lacks this capacity. Assuming others know she lacks the capacity to respond appropriately to

110. See Copp, *supra* note 54, at 271–75.

111. This is similar to the more general principle that someone ought to do something only if she can do it. See *id.*; Immanuel Kant, *The Metaphysics of Morals* 6:380 (Mary Gregor ed., 1996) (stating that “he must judge that he *can* do what the law tells him unconditionally that he *ought* to do”).

112. For this reason, kleptomania is also a partial type 2 defense to a crime of theft. In the standard case, a theft is strong evidence of a broader disposition to commit a broader range of crimes. See *supra* note 97 (pointing out the high degree of versatility among criminals).

the moral reasons against committing theft, there is nothing she could do to justify their believing that she is disposed to respond appropriately to them. As a consequence, she does not incur an obligation of restoration from committing her theft. Therefore, she does not deserve to be punished for the theft.¹¹³

Suppose a criminal actor has a mere partial type 3 defense. Then she might have committed a crime in performing her act, and the crime might have been as serious as presumed. The act might justify others in believing that she has the presumed disposition to commit crimes. So the act might justify others in believing that she is disposed to flout the presumed class of moral reasons against violating the rights of others. Thus, she might have undermined conditions of trust to the presumed degree by performing her act.

However, even if she did undermine conditions of trust to the presumed degree, she is not obligated to restore these conditions fully because she lacks the capacity to do so. To restore them fully, she must demonstrate that she is disposed to respond appropriately to all the relevant moral reasons she is presumably disposed to flout. But she lacks the capacity to demonstrate this because she lacks the capacity to respond appropriately to all the relevant moral reasons in virtue of having a partial type 3 defense. Assuming others know she lacks this capacity, there is nothing she could do to justify their believing that she is disposed to respond appropriately to all the relevant moral reasons.

Because she has the capacity to respond appropriately only to a subset of the relevant moral reasons, she has the capacity to demonstrate that she is disposed to respond appropriately only to a subset of them. So she is obligated to demonstrate that she is disposed to respond appropriately only to a subset of the relevant moral reasons. Hence, the defense mitigates the degree to which she is obligated to restore the conditions of trust she undermined by performing her act. Under her obligation of restoration, she

113. Although the kleptomaniac does not incur an obligation of restoration from committing her theft, she might still incur an obligation to compensate her victims for their losses, and if her disposition to commit theft is sufficiently bad, she might incur an obligation to incapacitate herself by undertaking a civil commitment. But neither requiring her to compensate her victims nor incapacitating her would constitute a punishment *per se*. For neither would express an attitude of moral blame toward her and neither would be necessarily or intentionally burdensome for her.

is obligated to restore only to a partial degree the conditions of trust she undermined by performing her act. As a consequence, the defense mitigates the severity of burdens she must undertake to fulfill the obligation of restoration she incurs from performing her act. Thus, the defense mitigates how much she deserves to be punished for the act.

4. Type 4 Defenses

Given its account of the three main types of defenses, RS generates a more general fourth type of defense. Type 4 defenses are considerations that mitigate the severity of burdens a criminal actor must undertake to fulfill the obligation of restoration she incurs from performing her act. If a criminal actor has a full type 4 defense, she need not undertake any burdens to fulfill the obligation. So she does not deserve any punishment for her act. If a criminal actor has a partial type 4 defense, she might need to undertake some burdens to fulfill the obligation, but the required burdens are not as severe as presumed. So the defense mitigates how much she deserves to be punished for her act. Although all type 1, type 2, and type 3 defenses are a subset of type 4 ones, some considerations can be type 4 defenses without falling under the extension of the other types. Consider three plausible candidates: childhood, brainwashing, and a rotten social background.

Consider childhood. In some cases, childhood provides a type 1 defense. Children are sometimes incapable of understanding the effects of their acts on others. When they perform criminal acts, they sometimes do not believe their acts have the properties that make them criminal ones.¹¹⁴ In other cases, childhood provides a type 2 but not a type 1 defense. Unlike adults, children often do not have settled dispositions. As they grow, their dispositions constantly change, often in radical ways. So the mere fact that a child was disposed to disrespect the rights of others when she performed a criminal act is sometimes not strong evidence that she is similarly disposed later at the time of assessment.¹¹⁵ In other cases, childhood provides a type 3 but not a type 2 or type 1 defense. Children sometimes lack the capacity to respond appropriately to the moral reasons against violating the rights of others.¹¹⁶ In other cases, though, childhood arguably provides a type 4 defense but not one of

114. See Scanlon, *supra* note 11, at 280.

115. See *id.* at 280–81.

116. See *id.* at 280.

the other types. Relative to adults, the dispositions of children are standardly much more malleable in the sense that they are much more responsive to the demands of authority figures. So even if children do undermine conditions of trust to the presumed degree, and even if they have the capacity to respond appropriately to the moral reasons against violating the rights of others, they can standardly go a long way toward restoring those conditions merely by undertaking a demanding but nonpunitive course in moral education. The same cannot be said for adults in the standard case. Hence, the fact that a criminal is a child mitigates the severity of burdens she must undertake to fulfill her obligation of restoration.

Consider brainwashing, which is a means of changing a person's beliefs or the norms she accepts. Brainwashing is coercive persuasion standardly carried out through means of manipulation and domination.¹¹⁷ When brainwashing is successful, it causes a radical change in the victim's system of normative self-governance. In particular, brainwashing can cause someone who accepts norms against committing crimes to accept norms requiring her to commit them. Now suppose such a person commits a crime because she has been brainwashed into accepting a norm requiring her to commit it.¹¹⁸ In such a case, she might not have a type 1 defense because she might have disrespected the rights of others to the presumed degree in committing the crime. She might not have a type 2 defense because her crime might be strong evidence that she has the presumed disposition to commit crimes at the time of assessment. More strongly, she might actually have such a disposition at the time of assessment. She might not have a type 3 defense because she might still

117. Brainwashing a victim often involves several elements, such as (1) isolating her; (2) exercising total control over her environment, especially the information she receives and what is communicated to her; (3) physically debilitating her by, for example, providing her with an inadequate diet, insufficient sleep, and poor sanitation; (4) requiring her to perform repetitive tasks, like copying written material; (5) manipulating her feelings of guilt and anxiety; (6) threatening to annihilate her unless she accepts the beliefs or norms of her apparently all-powerful captors; (7) degrading her for not accepting them; (8) subjecting her to peer pressure; and (9) requiring her to act contrary to the beliefs or norms that her captors want her to reject. See Richard Delgado, *Ascription of Criminal States of Mind: Toward a Defense Theory for the Coercively Persuaded ("Brainwashed") Defendant*, 63 Minn. L. Rev. 1, 2–3 (1978).

118. For cases that arguably involve such brainwashing, see the cases of American prisoners of war during the Korean conflict at *U.S. v. Batchelor*, 19 C.M.R. 452 (1954); *U.S. v. Fleming*, 19 C.M.R. 438 (1954); *U.S. v. Olson*, 20 C.M.R. 46 (1955). See also the case of Patricia Hearst at *U.S. v. Hearst*, 412 F. Supp. 873 (N.D. Cal. 1976).

have the capacity to respond appropriately to the moral reasons against violating the rights of others. Thus, she might have undermined conditions of trust by committing her crime, and she might incur an obligation to restore those conditions at the time of assessment. Nevertheless, she still might have a type 4 defense for her crime. When brainwashed people are freed from their captors and undergo a rigorous but non-punitive psychiatric process of de-programming or deconditioning, they are highly likely to reject the changes that were induced in them from the brainwashing.¹¹⁹ So when a criminal commits a crime as a result of brainwashing, she can go a long way toward restoring conditions of trust by merely separating from her captors and undergoing non-punitive treatment.¹²⁰ Hence, the fact that a criminal was brainwashed can mitigate the severity of burdens she must undertake to fulfill her obligation of restoration.¹²¹

Consider a rotten social background.¹²² Someone with such a background is standardly born into an environment involving, among other things, extreme poverty, an abundance of corrupting influences, a scarcity of positive influences, and a perceived lack of viable opportunities for escaping the environment. Such a background can place a high degree of pressure on someone to commit crimes, especially property crimes.¹²³ Suppose

119. See Delgado, *supra* note 117, at 9, 9 n.40, 30–32; Robert Jay Lifton, *Thought Reform and the Psychology of Totalism: A Study of "Brainwashing" in China* 86–151 (1961).

120. To obtain such separation and treatment, the brainwashed criminal will likely require the assistance of the state.

121. In addition to providing a type 4 defense, brainwashing might also provide a type 2 or type 3 defense. A criminal actor could have a type 2 defense of brainwashing if she performed her act as result of brainwashing, and she obtained treatment before the time of assessment. She could have a type 3 defense of brainwashing if her brainwashing undermined her capacity to respond appropriately to the moral reasons against violating the rights of others at the time of assessment. Although brainwashing can be a partial defense, I leave it an open question whether it can be a full defense.

122. *U.S. v. Alexander*, 471 F.2d 923, 961 (D.C. Cir. 1973) (Bazelon, J., dissenting) (using the phrase "rotten social background").

123. See, e.g., David L. Bazelon, *The Morality of the Criminal Law*, 49 S. Cal. L. Rev. 385 (1976); Richard Delgado, "Rotten Social Background": Should the Criminal Law Recognize a Defense of Severe Environmental Deprivation?, 3 *Law & Inequality* 9 (1985). Although a rotten social background can exert pressure on someone to commit crimes, it should be noted that only a minority of those with such a background actually succumb to such pressure and commit crimes. See Stephen J. Morse, *The Twilight of Welfare Criminology: A Reply to Judge Bazelon*, 49 S. Cal. L. Rev. 1247, 1259 (1976).

someone commits a crime, such as theft, in response to pressures from her rotten social background. Although she might lack a type 1, type 2, or type 3 defense for the crime, she still might have a type 4 defense for it. Even if she undermined conditions of trust by committing the crime, and has the capacity to respond appropriately to the moral reasons against violating the rights of others, she can restore the conditions to a significant degree by merely pursuing non-criminal opportunities to live under better social conditions. Such opportunities include, among others, opportunities for meaningful employment, opportunities for severing personal ties with corrupting influences, and opportunities to establish personal ties with positive influences.¹²⁴ By pursuing such opportunities, the criminal would free herself from many of the factors in her rotten social background that pressured her to commit crimes. Thus, the fact that a criminal has a rotten social background can mitigate the severity of burdens she must undertake to fulfill her obligation of restoration.¹²⁵

In the final analysis, RS explains why defenses mitigate how much criminal actors deserve to be punished for their acts by explaining why defenses mitigate the severity of burdens they must undertake to fulfill the obligation of restoration they incur from performing their acts. In some cases, criminal actors with defenses do not incur an obligation of restoration from performing their acts. They might not have undermined any conditions of trust by performing their acts, and even if they did undermine conditions of trust, they might lack the capacity to restore them. In other cases, criminal actors with defenses do incur an obligation of restoration, but their defenses nevertheless mitigate the severity of burdens they must undertake to fulfill the obligation.

124. Obtaining such opportunities would likely require the assistance of the state. Although their provision might be costly for the state, such costs are acceptable given that their deprivation arguably constitutes an injustice.

125. In addition to providing a type 4 defense, a rotten social background might also provide a type 2 or type 3 defense. A criminal actor could have a type 2 defense of a rotten social background if she performed her act against such a background, and she pursues non-criminal opportunities to live under better social conditions before the time of assessment. She could have a type 3 defense of a rotten social background if her background undermined her relevant capacities at the time of assessment. Although a rotten social background can be a partial defense, I leave it an open question whether it can be a full one.

VII. BLAMEWORTHINESS

So far we have focused on how RS explains the mitigating effects of defenses on punitive desert. In light of this explanation, we can see how RS explains their mitigating effects on blameworthiness. As we have noted, attitudes of moral blame contain certain demands, such as the demand to undertake a punishment, that are constitutive of the attitudes themselves. According to RS, an attitude of moral blame contains a demand to restore conditions of trust and a demand to undertake certain burdens to restore those conditions. When others blame a criminal actor for her act, they presuppose that she undermined certain conditions of trust by performing the act, and they demand her to restore those conditions by undertaking certain burdens. The higher the severity of burdens they demand her to undertake, the more they blame her for the act. The lower the severity of burdens, the less they blame her for the act.

According to RS, the degree to which a criminal actor is blameworthy for her act corresponds to the severity of burdens that others are warranted in demanding her to undertake to restore the conditions of trust she undermined by performing the act. The higher the severity of burdens they are warranted in demanding her to undertake, the more she is blameworthy for the act. The lower the severity of burdens, the less she is blameworthy for the act. According to RS, others are warranted in demanding a criminal actor to undertake certain burdens to restore the conditions of trust she undermined by performing her act if and only if she must undertake those burdens to fulfill the obligation of restoration she incurs from performing the act. Thus, by explaining the mitigating effects of defenses on the severity of burdens that criminal actors must undertake to fulfill their obligation of restoration, RS also explains their mitigating effects on how much criminal actors are blameworthy.

VIII. CONCLUSION

RS has important implications for the nature and strength of the state's reasons to recognize defenses as mitigating factors in its punitive policies. There are two possible reasons why the state might recognize them. On the one hand, the state might do so on the basis of considerations of value or efficiency. In particular, it might do so because the costs would outweigh the benefits of punishing criminal actors with defenses to an unmitigated

degree. On the other hand, the state might do so on the basis of considerations of justice. By showing that defenses mitigate how much criminal actors deserve to be punished, RS shows that their mitigating effects are indeed ultimately grounded in considerations of justice. So even if it would be costly or inefficient to punish criminal actors with defenses only to a mitigated degree, the state is still obligated to do so as a means to respecting their rights.

IX. APPENDIX

According to RS, a criminal deserves to be punished no more severely than the burdens she must undertake to fulfill her obligation of restoration. To fulfill it, she must send others a restorative signal, which is a directly observable property that is strong evidence of her acting with the required degree of benevolence for the required time after committing her crime.¹²⁶ By definition, a restorative signal is credible; it justifies others in believing that the criminal has acted with the required degree of benevolence for the required time. To be credible and not excessively burdensome, a restorative signal must satisfy two conditions.

According to “the incentive compatibility condition,”¹²⁷ a restorative signal cannot be too costly for criminals who are benevolent to the required degree for the required time. So the all-in costs of a restorative signal must not be greater than its all-in benefits for criminals who are benevolent to the required degree for the required time. The all-in benefits of a restorative signal for a criminal consist in all the expected consequences of her sending it that she desires to obtain; its all-in costs for her consist in all the expected consequences of her sending it that she

126. A signal is a directly observable property that is strong evidence of its bearer’s possessing another property that is not directly observable. The concept of a signal has interdisciplinary application. See, e.g., A. Michael Spence, *Market Signaling* (1974) (economics); Amotz & Avishag Zahavi, *The Handicap Principle* (Naama Zahavi-Ely & Melvin Patrick Ely trans., 1997) (biology); Erving Goffman, *The Presentation of Self in Everyday Life* (rev. ed. 1959) (sociology); Douglas G. Baird, Robert H. Gertner & Randal C. Picker, *Game Theory and the Law* 122–58 (1994) (law). For an elementary game-theoretic analysis of signaling, see Avinash Dixit & Susan Skeath, *Games of Strategy* 263–310 (2d ed. 2004). For an application of signaling theory specifically to trust, see Bacharach & Gambetta, *supra* note 102.

127. Bacharach & Gambetta, *supra* note 102, at 160.

desires not to obtain.¹²⁸ According to “the non-pooling condition,”¹²⁹ a restorative signal must be too costly for criminals who are not benevolent to the required degree for the required time. So the all-in costs of a restorative signal must be greater than its all-in benefits for criminals who are not benevolent to the required degree for the required time.¹³⁰ Thus, how burdensome a restorative signal must be depends on how benevolently a criminal must act and for how long to fulfill her obligation of restoration.

Suppose a criminal must act with a particular degree of benevolence, *B*, for a particular time, *t*, to fulfill her obligation of restoration. Suppose a person who is *B* cares equally about promoting a particular amount, *S*, of her own interests and promoting a minimal interest, *M*, of others. *B*’s benevolence factor, *f*, equals *S/M*, where $S \gg M$.¹³¹ So a person who is *B* is willing to sacrifice *S* of her own interests solely for the sake of promoting *M* interests of others.¹³² To demonstrate that she has acted with *B* for *t*, she must sacrifice *tS* of her own interests solely for the sake of promoting *tM* interests of

128. I take the concept of all-in benefits and costs from Bacharach & Gambetta, *supra* note 102, at 176–77. They distinguish between a person’s “all-in payoffs” and her “raw payoffs.” A person’s raw payoffs from an act consist only in its expected consequences that she cares about and that affect her well-being understood in the narrowest sense. Roughly speaking, her raw payoffs from the act consist only in its expected consequences that she cares about for her own sake. Her all-in payoffs from the act consist in both her raw payoffs from it and its expected consequences that she cares about but that do not contribute to her well-being understood in the narrowest sense. Roughly speaking, her all-in payoffs from the act consist in both her raw payoffs from it and its expected consequences that she cares about for the sake of something other than herself, such as other persons. Ultimately, a person’s all-in payoffs determine which acts she performs.

129. Bacharach & Gambetta, *supra* note 102, at 160.

130. A restorative signal need not satisfy the non-pooling condition for literally all criminals: some anomalous exceptions are consistent with its being sufficiently credible. Thus, a restorative signal can have a “semi-sorting equilibrium.” Bacharach & Gambetta, *supra* note 102, at 160.

131. Other things being equal, people are not obligated to act with such a high degree of benevolence. Indeed, other things being equal, acting with such a high degree of benevolence would constitute a rationally deficient form of self-abnegation. But when a person commits a crime, other things are not equal. The criminal incurs an obligation of restoration, and she must act with such a high degree of benevolence to fulfill it.

132. In clarification, two conditions determine the required absolute magnitudes of *S* and *M*. First, the absolute magnitude of the ratio *S/M* must be sufficiently high. Second, the absolute magnitude of the difference between *S* and *M* must also be sufficiently high. A criminal’s sacrificing *S* of her own interests solely for the sake of promoting *M* interests of others will demonstrate the required degree of benevolence only if *S/M* and *S–M* are both sufficiently high.

others.¹³³ Hence, a burden with a severity of $-tS$ is the least severe burden she must undertake to demonstrate that she has acted with B for t . For this is the least severe burden that would satisfy the incentive compatibility and non-pooling conditions on the credibility of a restorative signal.

First, the burden would satisfy the incentive compatibility condition because for a criminal who is B for t , sacrificing tS of her own interests solely for the sake of promoting tM interests of others would not be too costly. For her, the all-in costs of doing so would not be greater than the all-in benefits of doing so; they would be equal. The all-in benefits to her would equal the degree to which she cares about the interests she would promote in others: ftM . The all-in costs to her would equal the degree to which she cares about the interests she would sacrifice in herself: tS . Because $f=S/M$, $ftM=tS$.

Second, the burden would satisfy the non-pooling condition because for a criminal who is not B for t , sacrificing tS of her own interests solely for the sake of promoting tM interests of others would be too costly. For her, the all-in costs of doing so would be greater than the all-in benefits of doing so. To consider a simple case, suppose for t she is benevolent to a lesser degree B'' whose benevolence factor, f'' , is such that $f''<f$. The all-in benefits to her would equal $f''tM$. The all-in costs to her would equal tS . Because $f=S/M$ and $f''<f$, $tS>f''tM$.

Third, any less severe burden would violate the non-pooling condition because for a criminal who is not B for t , sacrificing any less than tS of her own interests solely for the sake of promoting tM interests of others would not be too costly. For her, the all-in costs of doing so would be less than or equal to the all-in benefits of doing so. Suppose the lesser sacrifice is $tS-E$. Suppose also that for t she is benevolent to a lesser degree B'' whose benevolence factor, f'' , is such that $(tS-E)/tM \leq f'' < f$. The all-in benefits to her would equal $f''tM$. The all-in costs to her would equal $tS-E$. Because $f'' \geq (tS-E)/tM$, $tS-E \leq f''tM$.

On a related note, the badness of the deficiency in the degree to which a criminal cares about others depends on two similar conditions. Suppose the criminal is willing to harm M interests of others to promote S of her own interests. The badness of her deficiency depends on the absolute magnitude of both S/M and $S-M$. The lower they are the worse the deficiency.

133. The units of t consist in the time required for the criminal to sacrifice S of her own interests for the sake of promoting M interests of others. So t corresponds to the number of benevolent acts she must undertake to fulfill her obligation of restoration. As a corollary, t corresponds to the number of times the criminal must sacrifice S of her own interests to promote M interests of others.