A Restorative Signaling Theory of Punitive Desert

by

Jim C. Staihar

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Philosophy)
in The University of Michigan
2008

Doctoral Committee:

Professor Elizabeth S. Anderson, Chair
Professor Stephen Leicester Darwall
Professor Allan F. Gibbard
Professor Thomas A. Green

Acknowledgements

For all their comments and support, I very much thank my dissertation committee: Elizabeth Anderson (chair), Stephen Darwall, Allan Gibbard, and Thomas Green. For insightful discussion and advice on writing the dissertation, I very much thank Tom Beauchamp, Aaron Bronfman, Sarah Buss, Wayne Davis, Ed Curley, David Dick, Lee Fennell, Steven Kuhn, Saul Levmore, Eric Lormand, James Mattingly, Richard McAdams, Martha Nussbaum, Matt Pugsley, Peter Railton, Don Regan, Scott Shapiro, and Matty Silverstein. For financial support while completing the dissertation, 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. For their special friendship and support, I very much thank Bob Darden, Chrysta Lienczewski, and my family.

Table of Contents

Abstract

A Restorative Signaling Theory of Punitive Desert

by

Jim C. Staihar

Chair: Elizabeth S. Anderson

I explain why and how much criminals deserve to be punished on the basis of a novel theory of punitive desert. Criminals deserve to be punished in the negative sense that the state would not violate their rights by punishing them against their will. Explaining why this is so is challenging because punishing someone involves intentionally harming her, and people have a prima facie right not to be intentionally harmed against their will. My restorative signaling theory avoids the shortcomings but incorporates the insights of the main competing theories in the literature, including expressive, consensual, unfair advantage, annulment, moral education, and deterrence theories of punitive desert.

My theory draws on the fact that when someone commits a crime without an exculpatory defense, she undermines conditions of trust, which are the conditions necessary for others' being justified in believing that she is not disposed to commit crimes. The criminal is obligated to restore such conditions because unless she does so, she will cause others to incur certain costs of insecurity. To restore the conditions of trust, the criminal must demonstrate to others that she has a good will, and to do so, she must sacrifice some of her sufficiently important personal interests for a sufficiently long time for the sake of benefiting others. According to the theory's main principle, 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.

The theory explains why punishments should ideally take the form of labor intensive community service performed under reasonable conditions of incapacitation. I

argue that the theory best explains why the state would not be justified in punishing all criminals extremely severely even though doing so could be the most efficient means of achieving deterrence. I also argue that it best explains the mitigating effects of exculpatory defenses on how much criminal actors not only deserve to be punished, but also are blameworthy. Finally, I argue that the theory illuminates what it means and when it is warranted to forgive criminals.

Chapter 1

A Restorative Signaling Theory of Punitive Desert

I. Introduction

When someone commits a crime, we standardly assume she[1] deserves to be punished for it in the sense that the state would not violate her rights by punishing her for it against her will.[2] If the punishment were not too severe, the criminal would not have an individual complaint in justice against suffering it, and the state would not owe her any compensation for the resulting harm. Our assumption, though, stands in acute need of justification. Punishing someone involves intentionally harming her, and people have a prima facie right not to be intentionally harmed against their will.[3] So punishing even a criminal against her will is prima facie to violate her rights.

In this chapter, I seek to explain why and how much criminals deserve to be punished. More specifically, I seek to justify a proportionality principle of punitive desert, which I call 'PP'. According to the principle, when someone commits a crime, she

---

1. Throughout the dissertation, I intend to use personal pronouns, such as 'he' or 'she', in a gender neutral sense.

2. 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 L. & 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 ISRAEL 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.

3. In other words, people have a prima facie claim or entitlement against such treatment. *See* 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).

deserves some punishment for it; however, the seriousness of her crime entails an upper limit on the severity of the punishment she deserves. A criminal deserves a particular severity of punishment for her crime if and only if the severity does not exceed the relevant upper limit. Three aspects of PP need clarification.

First, PP concerns only first time criminals who have no exculpatory defenses. Unless I note otherwise, I focus only on such criminals in this chapter. Second, the seriousness of a crime, in the sense relevant to punitive desert, corresponds to how badly the criminal disrespected the rights of others in committing it.[4] How badly she disrespected their rights corresponds to how badly she flouted the moral reasons against violating them. And how badly she flouted such reasons corresponds to how inappropriately she responded to them. In general, how inappropriately she responded to them depends on the beliefs, intentions, and motives with which she committed the crime. More precisely, it depends on a) what she intended in committing the crime, b) what motivated her to intend to commit it, c) how strongly she was motivated to intend to commit it, and d) the relative strength of the reasons she had for and against intending to commit it. What reasons she had depends on what she believed when she committed it.

Thus, a reason is meant here in a subjective sense that is relative to one's beliefs. In this sense, the reasons a criminal had for and against intending to commit her crime were provided by what she believed when she committed it.[5] In this sense, the fact that

4. *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). Throughout the dissertation, I focus only on crimes that are mala in se. I leave it an open question whether and how much people deserve to be punished for crimes that are mala prohibita.

5. 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 PHILOSOPHY: FROM ACTION TO VALUES 17, 17 (Dan Egonsson, Jonas Josefsson, Bjorn Petersson & Toni Ronnow-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

someone believes that her performing a particular act would directly cause another to die provides her with a strong moral reason against intending to do it. And 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. So if she were 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. Hence, the act would constitute a very serious crime because she would disrespect very badly the rights of others in performing it.

Third, the upper limits of PP have both an ordinal and a cardinal ranking. According to the ordinal ranking, a criminal's upper limit is relatively higher for committing relatively more serious crimes. In other words, a criminal deserves to be punished more severely for committing more serious crimes. For any crime, the cardinal ranking specifies the absolute magnitude of the upper limit on the severity of punishment that someone deserves for committing it. In other words, for any crime, the cardinal ranking specifies the absolute severity of the most severe punishment that someone deserves for committing it.

In seeking a justification of PP, I begin by considering five theories of punitive desert. I show that a consideration of their shortcomings and insights motivates a restorative signaling theory of punitive desert, which I call 'RS'. After providing an initial defense of RS, I respond to several objections that critics might raise against it. I conclude that RS plausibly justifies PP and so plays an important role in constraining the conditions under which the state may punish a criminal against her will. For this reason, RS should play an important role in any general theory of the justification of punishment.

---

be justified for them to provide her with a subjective reason. Similarly, I set aside for further analysis the issue of whether someone's subjective reasons are provided not by her beliefs, but more precisely, by the considerations that provide her with evidence to believe or not to believe certain propositions.

II. A General Theory of the Justification of Punishment

Before attempting to justify PP, let us identify its role in my 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.[6] 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 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.

6. 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).

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.

Within my general theory, PP determines whether punishing a criminal would satisfy the desert requirement. It does not determine whether the punishment would satisfy the general rights requirement or the value requirement. The mere fact that a criminal deserves a punishment does not entail that it satisfies the other two requirements.[7] To determine whether a deserved punishment satisfies them, we need, among other things, to know more about the actual consequences of the punishment, and we need a more general theory of rights and a theory of the valuable aims of punishment, which are the important values that the state might promote by imposing a deserved punishment on someone.[8] Thus, PP plays an important but also limited role in determining whether the state would be justified in punishing a criminal against her will. PP only determines whether the punishment would violate the criminal's rights.

III. Presuppositions

In spelling out a theory of punitive desert, I make several presuppositions that I assume obtain at the time a criminal is assessed for punitive desert. First, a criminal 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

---

7. So the state might not be justified in imposing a deserved punishment on a criminal. We might inquire into the practical difference between a) the state's imposing a deserved punishment on a criminal when it is not justified in doing so and b) its imposing an undeserved punishment on a criminal when it is not justified in doing so. In the latter but not the former case, the state would violate the criminal's rights by imposing the punishment on her. Hence, in the latter but not the former case, the criminal would have an individual complaint in justice against suffering the punishment, and the state would owe her compensation for the resulting harm.

8. My use of the term 'valuable aims of punishment' is similar to H.L.A. Hart's use of the term 'general justifying aims of punishment.' However, I use my term to refer to the valuable aims of punishing a particular person, whereas Hart uses his term to refer to the valuable aims of a general system of punishing persons. *See* H.L.A. HART, *Prolegomenon to the Principles of Punishment*, *in* PUNISHMENT AND RESPONSIBILITY: ESSAYS IN THE PHILOSOPHY OF LAW 1, 8-11 (1968).

criminal and others would unavoidably leave a significant degree of interaction between them and so would unavoidably leave others vulnerable to her to a significant degree. Third, there are no extraordinary means of obtaining epistemic access to a criminal'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 1-3 because I seek a practical explanation of why and how much criminals deserve to be punished. That is, I seek to explain why and how much they deserve to be punished in the actual world given the natural facts that generally characterize the unavoidable conditions under which people actually live. Presuppositions 1-3 are such facts. I leave it an open question whether and how much criminals would deserve to be punished in other worlds in which those presuppositions do not obtain.[9]

Fourth, everyone knows all the external and internal facts about all the acts the criminal has performed, including the external and psychological causes of her acts and their consequences. Fifth, everyone knows all the facts about her physical and psychological capacities. Sixth, everyone forms justified beliefs about her dispositions on the basis of his knowledge specified in presuppositions 4-5. 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 4-6 because I seek to explain why and how much

9. *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).

criminals deserve to be punished on the assumption that others fulfill their epistemic duties to each other and to the criminal 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 interest in them but are unaware of them. Presuppositions 4-6 entail that others fulfill such duties.

Also I make presuppositions 4-7 because the concept of punitive desert is plausibly defined conditionally on the assumption that they obtain at the time of assessment. On this definition, 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 4-7 were to obtain. The definition is plausible because how much someone deserves to be punished for doing something does not seem to depend essentially on the moral status of punishing her under less idealized conditions. For example, whether someone deserves a particular punishment for doing something does not seem to depend essentially on whether the state would violate her rights by imposing it on her under conditions in which a) others 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.

IV. Five Theories of Punitive Desert

A. An Expressive Theory

Consider an expressive theory, which I call 'E.' Punishing someone expresses an attitude of moral blame toward her.[10] Such an attitude might consist in resentment or

10. *See* JOEL FEINBERG, *The Expressive Function of Punishment*, *in* DOING AND DESERVING: ESSAYS IN THE THEORY OF RESPONSIBILITY 98 (1970).

indignation.[11] According to E, a person deserves a punishment if and only if imposing it on her would be the least harmful means of expressing a warranted attitude of moral blame toward her.[12] A criminal deserves some punishment because everyone is warranted in feeling an attitude of moral blame toward her, and punishing her to some degree would be the least harmful means of expressing that attitude toward her.

E's defense of PP is problematic for at least three reasons. First, E is circular. E merely assumes that punishing a criminal would express a warranted attitude of moral blame toward her. On a standard view, an attitude of moral blame consists at least partly in certain demands.[13] Feeling an attitude of moral blame toward someone involves feeling or making certain demands on her. One demand that is constitutive of an attitude of moral blame is the demand to undertake a punishment.[14] Punishing someone expresses an attitude of moral blame toward her by expressing its constitutive demand on her to undertake the punishment. So to claim that punishing a criminal would express a warranted attitude of moral blame toward her is to claim that the punishment would

---

11. *See, e.g.*, P.F. STRAWSON, *Freedom and Resentment, in* FREEDOM AND RESENTMENT AND OTHER ESSAYS 1 (1974).

12. In addition, some theories might insist on the punishment's expressing moral blame in a way that communicates it or makes it understandable to the criminal. *See, e.g.*, R.A. DUFF, PUNISHMENT, COMMUNICATION, AND COMMUNITY (2001); Igor Primoratz, *Punishment as Language*, 64 PHIL. 187 (1989).

13. *See* STEPHEN DARWALL, THE SECOND-PERSON STANDPOINT: MORALITY, RESPECT, AND ACCOUNTABILITY 17 (2006); STRAWSON, *supra* note 11, at 14-15, 21-22; Gary Watson, *Responsibility and the Limits of Evil: Variations on a Strawsonian Theme*, *in* PERSPECTIVES ON MORAL RESPONSIBILITY 121, 126-28 (John Martin Fischer & Mark Ravizza eds., 1993).

14. I take attitudes of moral blame to be 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 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).

express a warranted demand on her to undertake it. But to claim that a demand on a criminal to undertake a punishment is warranted presupposes that she deserves the punishment. The claim that a criminal deserves a punishment is precisely the claim to be explained. Because E presupposes the claim it aims to explain, it is circular.[15]

Second, punishing a criminal is not the least harmful means of expressing a warranted attitude of moral blame toward her. Punishing a criminal involves significantly harmful treatment, which might consist in the deprivation of an important liberty or harm to some other important personal interest. It does not seem that any attitude of moral blame is ever expressible only through such harmful treatment. Blame seems expressible through behavior that does not significantly harm the criminal.[16] For example, we might express blame toward a criminal by simply a) communicating our feelings to her verbally in a sufficiently solemn ceremony or ritual[17] and b) refraining from interacting with her in various ways for some time. Because any attitude of moral blame seems expressible through behavior that is not significantly harmful, punishing a criminal never seems to be the least harmful means of expressing such an attitude toward her.

Third, suppose for the sake of argument that a warranted attitude of moral blame does not contain a warranted demand to undertake a punishment. Suppose also that punishing a criminal is the least harmful means of expressing such an attitude toward her. It is still not clear why she would deserve the expression of the attitude. There is no problem with our merely feeling the attitude toward her. Our merely feeling it would not

---

15. *Cf.* Baldwin, *supra* note 14, 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).

16. As S.I. Benn asserts, expressing an attitude of moral blame or "denunciation does not imply the deliberate imposition of suffering, which is the feature of punishment usually felt to need justification." S.I. Benn, *An Approach to the Problems of Punishment* 33 PHIL. 328 n.1 (1958).

17. As Feinberg writes: "Such a ritual might condemn so very emphatically that there could be no doubt of its genuineness, thus rendering symbolically superfluous any further hard physical treatment." FEINBERG, *supra* note 10, at 116.

harm her and so would not violate her rights. However, if punishing her is the least harmful means of expressing the attitude, then the state's expressing the attitude would involve intentionally harming her. Because the state's expressing the attitude would involve intentionally harming, it would prima facie violate her rights. Unless expressivists can explain why criminals do not have a right to be free from the intentionally harmful expression of a warranted attitude of moral blame, their theory does not explain why they deserve to be punished.

Although E does not justify PP, it yields an important insight into the type of theory that could. We can show that a criminal deserves to be punished for her crime by showing that others would be warranted in demanding her to undertake a punishment for her crime. To show this, we must show that she is obligated to undertake a punishment for her crime.[18] If she is, then she has no right to be free from the punishment.[19] So if she is obligated to undertake a punishment, and refuses to do so voluntarily, then the state would not violate her rights by imposing the punishment on her against her will. The challenge is to explain why criminals are obligated to undertake punishments for their crimes.

B. An Actual Consent Theory

Consider an actual consent theory, which I call 'AC.'[20] If someone consents to certain treatment, and her consent is irrevocable, then imposing the treatment on her

18. *Cf.* David Copp, *'Ought' Implies 'Can', Blameworthiness, and the Principle of Alternate Possibilities*, *in* MORAL RESPONSIBILITY AND ALTERNATIVE POSSIBILITIES 265, 271-75 (David Widerker & Michael McKenna eds., 2003) (noting the relation between what a person is obligated to do and what others are warranted in demanding her to do or can fairly demand her to do); DARWALL, *supra* note 13, at 96-99 (same).

19. I take this point to be basic for the purposes of this dissertation. In general, if someone is obligated to undertake a burden, she has no right to be free from it. In other words, she has no claim or entitlement to be free from the burden. *Cf.* MILL, *supra* note 14, 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.").

20. *See* C.S. Nino, *A Consensual Theory of Punishment*, 12 PHIL. & PUB. AFF. 289, 297-300 (1983) (defending the actual consent theory I describe).

against her will would not violate her rights. According to AC, someone consents to a consequence if she freely performs an act under conditions in which she knows the act would have the consequence. When someone freely commits a crime, she knows her crime will have the consequence of forfeiting her legal right against being punished. So by freely committing her crime, she consents to forfeit this legal right. In general, if someone consents to forfeit her legal right against certain treatment, she consents to forfeit her moral right against it. Thus, by freely committing her crime, the criminal consents to forfeit not only her legal right, but also her moral right against being punished. Therefore, criminals deserve to be punished.

In response, AC is objectionable for at least three reasons. First, its main principle about consent is false. As a conceptual truth, a person consents to something only if she intends to consent to it in the sense of intending to authorize or agree to it. As a consequence, a person consents to something only if she believes she consents to it. A person cannot unintentionally consent to something or consent to it by accident.[21] So the mere fact that someone freely performs an act knowing it will have a particular consequence does not entail that she consents to the consequence, because this fact does not entail that she intends to consent to the consequence or believes she consents to it.

To illustrate, suppose I know someone will destroy my home if I attend a political rally. The mere fact that I freely attend the rally does not entail that I consent to the destruction of my home, because this fact does not entail that I intend to consent to this consequence or believe I consent to it. Hence, even if someone commits a crime knowing her crime will the consequence of forfeiting her legal right against being punished, this fact does not entail that she consents to forfeit the right. Because criminals do not intend to consent to forfeit any of their rights when they commit their crimes, they do not consent to forfeit their legal or moral rights against being punished.

---

21. *See* A. JOHN SIMMONS, MORAL PRINCIPLES AND POLITICAL OBLIGATIONS, 75-79 (1981).

Second, suppose for the sake of argument that AC's main principle about consent is correct. Still some criminals do not know that they forfeit their legal right against being punished when they commit their crimes.[22] They are mistaken about what the criminal laws of the state prohibit. For example, someone might kill her spouse's lover on the mistaken assumption that paramour killings are not crimes.

Third, AC does not account for PP's ordinal or cardinal rankings.[23] For example, suppose everyone knows that the state punishes some mildly serious crimes extremely severely, but it punishes moderately serious crimes only moderately severely. Then when people freely commit such mildly serious crimes, AC entails that they consent to forfeit their legal and moral rights against being punished extremely severely for them. So AC unacceptably entails that those mildly serious criminals deserve to be punished more severely than moderately serious criminals, and they deserve to be punished extremely severely.[24] Under PP's ordinal ranking, mildly serious criminals deserve to be punished less severely than more serious criminals, and under its cardinal ranking, they deserve no more than mildly severe punishments for their crimes.

Although AC does not justify PP, it yields an important insight into the type of theory that could. AC tries to prove that criminals deserve to be punished because they actually consent to forfeit their moral right against being punished. Although criminals do not actually consent to forfeit any of their rights by committing their crimes, the concept of consent might still play a role in explaining why they deserve to be punished. Even though criminals do not actually consent to be punished, they might be obligated to

22. *See* T.M. SCANLON, *Punishment and the Rule of Law*, *in* THE DIFFICULTY OF TOLERANCE: ESSAYS IN POLITICAL PHILOSOPHY 219, 219-33 (2003).

23. *See* Larry Alexander, *Consent, Punishment, and Proportionality*, 15 PHIL. & PUB. AFF. 178 (1986).

24. The consequences of punishing mildly serious criminals extremely severely might be worse than the consequences of punishing them less severely. Hence, punishing them so severely might be unjustified on the grounds that it would violate the value requirement of my general theory. But whether or not punishing them so severely would be unjustified on these grounds is irrelevant to whether they deserve to be punished so severely.

consent to be punished. More precisely, they might be obligated to consent to undertake a punishment. If so, then they deserve to be punished. For in general, if someone is obligated to consent to undertake certain burdens, then she has no right to be free from them.[25] The challenge is to explain why criminals are obligated to consent to undertake punishments for their crimes.

C. An Unfair Advantage Theory

Consider an unfair advantage theory, which I call 'UA.'[26] UA assumes a "principle of fair play."[27] According to this principle, if people undertake a burden to maintain a just scheme of social cooperation from which they benefit, then anyone else who accepts benefits from the scheme is obligated to do her fair share to maintain it.[28] Doing her fair share to maintain the scheme involves her undertaking a fair share of the burdens that maintain it. Given the principle of fair play, UA assumes that in the state, a group of cooperators maintains a just scheme of social cooperation by undertaking the burden of restraining themselves from committing crimes. And each of the state's citizens accepts from the scheme the benefit of the cooperators' not committing crimes against her. As a consequence, each of the state's citizens is obligated to do her fair share to maintain the scheme. And doing her fair share to maintain the scheme involves her assuming the burden of self-restraint, restraining herself from committing crimes.

---

25. On one interpretation, if criminals are obligated to consent to undertake a punishment, then this 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, chapters 4-5 (forthcoming) (discussing interpretations of this formulation of Kant's categorical imperative and its relation to various principles of consent).

26. 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).

27. JOHN RAWLS, *Legal Obligation and the Duty of Fair Play*, *in* COLLECTED PAPERS 117, 122 (Samuel Freeman ed., 1999).

28. John Rawls endorses the "principle of fair play" in *id.* at 122. H.L.A. Hart endorses a similar principle in *Are There any Natural Rights?*, 64 PHIL. REV. 175, 185 (1955).

By committing a crime, a criminal obtains an unfair advantage over the cooperators that consists in the benefit of her renouncing the burden of self-restraint. Unless the benefit is removed, the benefit will constitute an unjust enrichment. For the criminal obtains the benefit by violating her obligation not to commit the crime. And in general, if someone obtains a benefit from violating an obligation to others, the benefit constitutes an unjust enrichment. To remove the benefit, the criminal must undertake a burden that is as severe as the burden of self-restraint she renounced by committing her crime. Because she would be unjustly enriched unless she undertakes such a burden, the state may impose the burden on her as a punishment against her will without violating her rights. For in general, if someone would be unjustly enriched by not undertaking a burden, and refuses to undertake it voluntarily, then the state would not violate her rights by imposing the burden on her against her will.[29] According to the main principle of UA, a criminal deserves a punishment that is no more severe than the burden of self-restraint she renounced by committing her crime.

UA does not justify PP. On the most natural interpretation of the burden of self-restraint that a criminal renounces by committing her crime, the severity of the burden is proportional to the strength of the average cooperator's desire to commit the crime.[30] On this interpretation, UA does not justify PP's cardinal or ordinal ranking. UA does not justify PP's cardinal ranking because for very serious crimes, such as murder and rape, the average cooperator has no desire to commit them, so restraining herself from committing them is no burden at all. Thus, UA does not explain why many serious criminals deserve

29. This principle reflects the norm that a person may not profit from her own crime. *See, e.g.*, Riggs v. Palmer, 22 N.E. 188, 190 (N.Y. 1889) (claiming that "[n]o one shall be permitted to profit by his own fraud, or to take advantage of his own wrong, or to found any claim upon his own iniquity, or to acquire property by his own crime").

30. Richard Burgh discusses this interpretation of the burden of self-restraint and its objectionable consequences for UA in *Do the Guilty Deserve Punishment?*, 79 J. PHIL. 193, 209-10 (1982). *See also* 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).

any punishment at all. UA does not justify PP's ordinal ranking because the average cooperator has a stronger desire to commit less serious crimes, like tax evasion and theft, than more serious crimes, like murder and rape. So the average cooperator assumes a greater burden in restraining herself from committing relatively less serious crimes. Hence, UA does not explain why criminals deserve to be punished more severely for committing more serious crimes.

Although UA does not justify PP, it yields an important insight into the type of theory that could. UA tries to prove that a criminal deserves a punishment because she would be unjustly enriched by not undertaking it. But unlike UA, the unjust enrichment does not consist in a benefit that she obtains from committing her crime. Contrary to UA, there might not be such a benefit. Rather the unjust enrichment consists in a benefit that she stands to obtain from violating an obligation she incurs from committing her crime. The challenge is to identify this obligation.

D. An Annulment Theory

Consider an annulment theory, which I call 'AN.'[31] According to AN, someone's committing a crime expresses the false proposition that her victim lacks the right to be free from the crime. By expressing the false proposition about her victim, the criminal imposes on him a moral harm that consists in her falsely representing him as not having the right. The moral harm will persist until the false proposition is annulled. Because the criminal caused her victim the harm, she is obligated to annul it by annulling the false proposition about him. So she is obligated to undertake any means necessary to annul the false proposition. 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.[32] Hence, the

31. Hegel seems to endorse a similar annulment theory in HEGEL'S PHILOSOPHY OF RIGHT §§ 90-103 (T.M. Knox trans., 1942). My interpretation of Hegel follows David Cooper's interpretation of him in *Hegel's Theory of Punishment*, *in* HEGEL'S POLITICAL PHILOSOPHY: PROBLEMS AND PERSPECTIVES 151 (Z.A. Pelczynski ed., 1971). Jean Hampton endorses such a theory in *Correcting Harms Versus Righting Wrongs: The Goal of Retribution*, 39 UCLA L. REV. 1659-702 (1991-92).

32. This is a transmission principle concerning obligations. For similar transmission principles concerning

criminal is obligated to undertake any burdens that are necessary to annul the false proposition. So assuming she must undertake certain burdens to annul it, the state may impose those burdens on her as a punishment against her will without violating her rights. According to the main principle of AN, a criminal deserves a punishment for her crime that is no more severe than the burdens she must undertake to annul the false proposition she expressed about her victim by committing it.

AN does not justify PP. A criminal need not undertake any burdens to annul the false proposition about her victim. To annul it, a criminal need only comply with a non-punitive policy: a) assert the negation of the false proposition, b) apologize to her victim, and c) compensate her victim for any material or psychological harm she caused him. A criminal's complying with this policy would annul the false proposition by expressing the fact that her victim had the right to be free from the crime.[33] Because a criminal can annul the false proposition by complying with a non-punitive policy, AN does not justify PP's cardinal or ordinal ranking.

AN does not justify PP's cardinal ranking because complying with the non-punitive policy does not require many serious criminals to undertake any significant burdens. Neither asserting the negation of the false proposition nor merely apologizing to her victim requires a criminal to undertake any significant burdens. And requiring a criminal to compensate her victim for any material or psychological harm she caused him does not require many serious criminals to undertake any burdens because many serious crimes, like some unsuccessful attempts at murder, do not cause much or any such harm to their victims.[34] Thus, AN does not explain why many serious criminals deserve any

---

"oughts," see Mark Schroeder, *Means-end Coherence, Stringency, and Subjective Reasons*, PHIL. STUD. (forthcoming).

33. *See* Deidre Golash, *The Retributive Paradox*, 54 ANALYSIS 72, 75 (1994); Dolinko, *supra* note 30, at 545-49.

34. For example, consider a case in which someone unsuccessfully tries to murder a person under conditions in which he is unaware of the attempt and would not experience much, if any, trauma upon learning about it.

punishment at all.

AN does not justify PP's ordinal ranking because complying with the non-punitive policy does not require many serious criminals to undertake more severe burdens than less serious criminals. As we noted, the policy does not require many serious criminals, like some who unsuccessfully attempt to commit murder, to undertake any significant burdens. However, the policy might require some less serious criminals to undertake significant burdens. For some less serious crimes, like vandalism, cause their victims significant material harms, and the criminals who commit them might need to undertake significant burdens to compensate their victims for those harms.[35] Thus, AN does not explain why criminals deserve to be punished more severely for committing more serious crimes.

Although AN does not justify PP, it yields an important insight into the content of an obligation that could. AN tries to prove that a criminal deserves to be punished because she is obligated to undertake a punishment as means to annulling something that her crime expresses. However, what justifies PP is not a criminal's obligation to annul a false proposition about her victim, rather it seems to be her obligation to annul something her crime indicates about herself that is of greater social importance. The challenge is to identify what her crime indicates about herself that she has an obligation to annul by undertaking a punishment.

E. A Prevention Theory

Consider a prevention theory, which I call 'P.' According to P, someone's committing a crime indicates that she is disposed to commit crimes in the sense that she is willing and able to commit them. Unless she annuls her disposition to commit crimes in the sense of extinguishing it, she will commit crimes in the future. Everyone is

35. Vandalism is a less serious crime than an unsuccessful attempt at murder because an attempted murderer disrespects to a worse degree the rights of others. For the right to life protects more important interests than the right to property.

obligated to undertake the means that are necessary to prevent herself from committing crimes. Hence, a criminal is obligated to undertake the means that are necessary to annul her disposition to commit crimes. So she is obligated to undertake any burdens that are necessary to annul it. Thus, assuming she must undertake certain burdens to annul it, the state may impose those burdens on her as a punishment against her will without violating her rights. According to the main principle of P, a criminal deserves a punishment that is no more severe than the burdens she must undertake to annul her disposition to commit crimes.

Assuming a criminal is disposed to commit crimes, there are three reasons why she might need to undertake certain burdens to annul the disposition at least partially. First, if she is incapable of annulling her willingness to commit crimes, she might need to incapacitate herself, placing herself under conditions that would hinder her ability to commit crimes.[36] Second, if she would annul her willingness to commit crimes by having a sufficiently strong prudential incentive not to commit them, she might need to follow a policy of undertaking certain burdens after she commits crimes. By following such a policy, she might deter herself from committing them out of fear of suffering the burdens.[37] Third, if she would annul her willingness to commit crimes by undertaking a course in moral education, she might need to do so to develop a good will, which is a stable disposition to be appropriately motivated by the moral reasons against violating the rights of others.[38] To be effective, such a course might need to be burdensome, in some

36. Conditions of incapacitation as such are not necessarily burdensome. Nevertheless, their burdensome nature is standardly unavoidable to some degree.

37. If a person were to follow a policy of undertaking certain burdens after she commits crimes, she might deter herself from committing a crime in the present out of fear that her future self would otherwise be harmed. To prevent her future self from opting out, she might need to implement the policy through a device that would automatically impose the burdens on her future self if she were to commit a crime in the present. For the related concept of an "automatic retaliation device," see Lawrence Alexander, *The Doomsday Machine: Proportionality, Punishment and Prevention*, 63 THE MONIST 199, 209 (1980); Warren Quinn, *The Right to Threaten and the Right to Punish*, 14 PHIL. & PUB. AFF. 327, 337-38 (1985); Claire Finkelstein, *Threats and Preemptive Practices*, 5 LEGAL THEORY 311, 315-16 (1999).

38. *See* Jean Hampton, *The Moral Education Theory of Punishment*, 13 PHIL. & PUB. AFF. 208 (1984);

way, to prevent her from opting out and to encourage her to take the content of the course seriously.

In response, if a criminal must undertake certain burdens to annul her disposition to commit crimes, then P does explain why she deserves some punishment. However, P does not justify PP's cardinal or ordinal ranking. It does not justify the cardinal ranking because some criminals who deserve some punishment need not undertake any burdens to annul their dispositions to commit crimes. For example, consider immediately reformed criminals who perform serious crimes, like murder, but who immediately reform themselves afterward. They immediately annul their dispositions to commit crimes by simply recognizing, appreciating, and becoming appropriately motivated by the moral reasons against committing them.[39] Immediately reformed criminals deserve some punishment. But because they annulled their dispositions to commit crimes without undertaking any burdens, they are not obligated to undertake any as a means to annulling such a disposition. Thus, P does not explain why some serious criminals deserve any punishment.

P does not justify the ordinal ranking because some less serious criminals must undertake more severe burdens than more serious criminals to annul their dispositions to commit crimes. As we noted, very serious criminals, like murderers, who immediately reform themselves need not undertake any burdens to annul their dispositions to commit crimes. However, less serious criminals, like non-violent thieves, who do not immediately reform themselves might need to undertake some burdens to annul theirs. Nevertheless, very serious criminals who are immediately reformed still deserve to be punished more severely than less serious criminals who are not reformed. Thus, P does not generally explain why criminals deserve to be punished more severely for committing

---

Herbert Morris, *A Paternalistic Theory of Punishment*, 18 AM. PHIL. Q. 263 (1981).

39. Cases of immediately reformed criminals are not prevalent. Nevertheless, they are not fantastic or beyond the realm of relevant possibilities.

more serious crimes.

Although the theory does not justify PP, it yields an important insight into the content of an obligation that could. P points out that someone's committing a crime is strong evidence that she is disposed to commit crimes. However, what justifies PP is not a criminal's obligation to annul her disposition to commit crimes; it is her obligation to annul something that even immediately reformed criminals retain an obligation to annul.

V. A Restorative Signaling Theory of Punitive Desert

In light of the insights from the previous theories, 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. 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. Second, she is obligated to consent to undertake the punishment.[40] In general, if someone 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.[41] 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 punishment necessary to fulfill the obligation. In general, if someone obtains a benefit from violating an obligation to others,

40. 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. 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.

41. Insofar as an unjust enrichment constitutes an unfair advantage, RS could be considered a type of unfair advantage theory of punitive desert. Unlike UA, 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 UA, RS can account for the proportionality of punitive desert, as I argue below.

the benefit constitutes an unjust enrichment. Each of these related claims entails that she has no right to be free from the punishment. 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 deserve to be punished by explaining why they must undertake a punishment to fulfill an obligation they incur from committing their crimes. 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.[42] 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

42. *Cf.* THOMAS HOBBES, LEVIATHAN chap. 13 (Richard Tuck ed., Cambridge 1996) (noting the costs of insecurity that someone rationally must incur in response to being justified in believing that others are disposed to engage in acts of aggression).

takes to restore them the longer others must incur the costs of insecurity.[43]

When someone commits a crime, she undermines conditions of trust.[44] Her crime is strong evidence that she is disposed to commit crimes; therefore, others are not justified in believing otherwise. So she is 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.[45]

A criminal 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, 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

---

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

44. 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 L. & PHIL. 37 (1997); David A. Hoekema, *Trust and Obey: Toward a New Theory of Punishment*, 25 ISRAEL L. REV. 332 (1991).

45. 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 it. 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 it.

will.[46] 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.[47] 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.[48] For the fact that someone acts with a sufficiently high degree of benevolence for a sufficiently long time is strong evidence that she has a stable disposition to care highly about the

---

46. 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).

47. 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 13, 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. On 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. On 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, on 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).

48. 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 47. On 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.

interests of others.[49] 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. Thus, to fulfill her obligation of restoration, a criminal must undertake some burdens.[50]

At this point, we have shown that RS explains why criminals deserve some punishment for their crimes. Now we can show that it plausibly justifies PP's ordinal and cardinal rankings. With respect to its ordinal ranking, RS explains why criminals deserve to be punished more severely for committing more serious crimes. Suppose a criminal commits a more serious crime and, thus, disrespects the rights of others to a worse degree in committing it. Then her crime is strong evidence that she has a worse disposition to commit crimes. Her crime is strong evidence that she has a worse disposition to disrespect the rights of others. Hence, the crime is strong evidence that there is a worse deficiency in the degree to which she cares about the interests of others. On reflection, to annul this evidence and demonstrate that she has a stable disposition to care highly about the interests of others, she must demonstrate that she has acted with a higher degree of benevolence for a longer time after committing her crime. To demonstrate this, she must sacrifice more of her important personal interests for a longer time for the sake of benefiting others. So she must undertake a more severe burden to restore the conditions

---

49. *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).

50. 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, like 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.

of trust she undermined by committing her crime. Thus, she must undertake a more severe burden to fulfill her obligation of restoration. So she deserves a more severe punishment. In general, someone who commits a more serious crime deserves a more severe punishment for her crime because she must undertake a more severe burden to fulfill the obligation of restoration she incurs from committing the crime.

With respect to PP's cardinal ranking, RS accounts for the absolute severity of the punishments that criminals deserve. 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. In general, the absolute severity of such burdens does correspond to the absolute severity of the most severe punishment that a criminal seems to deserve for her crime.

To illustrate, consider someone who commits a moderately serious crime 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 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 deserves no more than a mildly severe punishment for her crime, and an extremely serious criminal deserves an extremely severe punishment for hers.

VI. Critical Discussion

A. A Specific Deterrence Theory

Critics might argue that RS is at best only one among two theories that each justifies PP. They might concede that criminals have an obligation of restoration. But contrary to RS, they might argue that criminals need not demonstrate that they have a good will to fulfill it. Rather they can fulfill it merely by demonstrating that they have a deterred will in the sense that they have a stable disposition not to commit crimes out of fear of being punished for committing them. The critics might claim that a criminal deserves a punishment that is no more severe than the one she must undertake to demonstrate that she has a deterred will. Call this specific deterrence theory 'SD.'

In response, SD does identify an important potential value of imposing a deserved punishment on a criminal: the punishment might deter her from committing some crimes in the future. But SD does not justify PP. A criminal will be deterred from committing a crime by its expected punishment. The crime's expected punishment is the product of its severity and probability of punishment. Its severity of punishment is the severity of the punishment that the state would impose on her for committing it. Its probability of punishment is the criminal's subjective probability that the state would detect and punish her if she were to commit it. Given that a criminal will be deterred by a crime's expected punishment, it is not practically feasible to use specific deterrence as the basis for determining how much to punish her because there is no practically feasible way to control precisely a crime's probability of punishment across all the situations in which she will find herself. As a consequence of the inevitable variability in a crime's probability of punishment, SD does not justify PP's cardinal ranking. In many situations, a crime's probability of punishment will be too low to justify a deserved punishment on specific deterrence grounds, whereas in many others, it will be too high. Contrary to SD, how much a criminal deserves to be punished does not vary with any crime's probability of punishment.

To illustrate, consider a moderately serious criminal who deserves to be punished no more than moderately severely. On the one hand, she will commonly find herself in situations in which her crime's probability of punishment is insignificant. In these situations, she will be significantly confident that she can commit the crime without being detected or punished. Such situations are frequent and unavoidable because states do not have the resources to maintain a significant probability of punishment for criminals across all the situations in which they will find themselves.[51] Maintaining such a probability is especially difficult given that a criminal might ignore or be overly optimistic about her ability to avoid detection. In such situations, there is no strong reason to believe that the threat of a moderately severe punishment will be sufficient to deter the criminal from committing the crime: it is likely that only the threat of a much more severe punishment will deter her if any will. So with respect to these situations, SD entails that the criminal deserves either no punishment at all or a punishment that is much more than moderately severe. Either way SD's consequences for her are either too lenient or too harsh.

On the other hand, the criminal will also commonly find herself in situations in which her crime's probability of punishment is significant. In these situations, she will be significantly confident that she cannot commit the crime without being detected and punished. In these situations, there is no strong reason to believe that the threat of a moderately severe punishment will be necessary to deter her from committing the crime: it is likely that the threat of a much less severe punishment will be sufficient. In general, criminals have little incentive to commit a crime in situations in which its probability of punishment is significant given that they will inevitably find themselves in situations in which its probability of punishment is insignificant. So with respect to situations in

51. *See* Paul H. Robinson & John M. Darley, *Does Criminal Law Deter? A Behavioral Science Investigation*, 24 OXFORD J. LEGAL STUD. 173, 184 (2004); Paul H. Robinson & John M. Darley, *The Role of Deterrence in the Formulation of Criminal Law Rules: At Its Worst When Doing Its Best*, 91 GEO. L. J. 949, 954, 992-94 (2003).

which a crime's probability of punishment is significant, SD entails that the criminal deserves at most a punishment that is much less than moderately severe. Here SD's consequences for her are too lenient.

The critics might concede that SD does not justify PP's cardinal ranking but argue that it still justifies the ordinal ranking. They might argue that to reduce the costs of insecurity that others must incur because she undermined conditions of trust, a criminal must demonstrate that she has a marginally deterred will in the sense that she has a stable disposition to minimize the harm she would cause in pursuing her criminal objectives. In general, more serious crimes are more harmful than less serious crimes. So to demonstrate that she is marginally deterred, a criminal must demonstrate that she is disposed to choose a less serious crime over a more serious crime in pursuing her criminal objectives. In general, a criminal will be marginally deterred if the punishments for more serious crimes are more severe than those for less serious crimes. So to demonstrate that she is marginally deterred, a criminal must undertake more severe punishments for committing more serious crimes. Thus, a criminal deserves to be punished more severely for committing more serious crimes.

In response, SD does not even justify PP's ordinal ranking for two reasons. First, to reduce the costs of insecurity that others must incur because she undermined conditions of trust, a criminal at most must demonstrate that she is marginally deterred only with respect to crimes that are substitutes.[52] Two crimes are substitutes just in case increasing the expected punishment of one will increase the demand for the other.[53] For example, armed robbery and unarmed robbery are substitutes because increasing the expected punishment of unarmed robbery beyond that of armed robbery will increase the demand

52. RICHARD A. POSNER, ECONOMIC ANALYSIS OF LAW 245 (5th ed. 1998); David Friedman & William Sjostrom, *Hanged for a Sheep: The Economics of Marginal Deterrence*, 22 J. LEGAL STUD. 345, 345-46, 357-60 (1993).

53. *See* Friedman & Sjostrom, *supra* note 52, at 358.

for armed robbery, as criminals will have no incentive not to use firearms to facilitate their robberies. For another example, suppose "a murder-theft" consists in someone's murdering another in the course of a theft, and "a mere theft" consists in someone's committing a theft without murdering anyone.[54] Murder-theft and mere theft are substitutes because increasing the expected punishment of mere theft beyond that of murder-theft will increase the demand for murder-theft, as criminals will have no incentive not to use murder to facilitate their thefts. Not all crimes, though, are substitutes. For example, rape is much more serious than theft, but presumably they are not substitutes because increasing the expected punishment of one will not increase the demand for the other, since a criminal's decision to commit one does not depend on the expected punishment of the other. If two crimes are not substitutes, then a criminal would not choose between them, and so no one would rationally incur any costs of insecurity on the assumption that she would choose between them. Hence, if two crimes are not substitutes, a criminal need not demonstrate that she is disposed to choose the less serious one over the more serious one. And so if two crimes are not substitutes, SD does not explain why a criminal who performs the more serious one deserves to be punished more severely than a criminal who performs the less serious one.

Second, even if a criminal must demonstrate that she is marginally deterred with respect to all crimes, SD still does not justify PP's ordinal ranking. As we have noted, a criminal is deterred by a crime's expected punishment, not by its severity of punishment considered by itself. So to demonstrate that she is marginally deterred, a criminal need only undertake higher expected punishments for committing more serious crimes. But she need not undertake more severe punishments to undertake higher expected punishments. Assuming the probability of punishment is higher for more serious crimes, such crimes can have higher expected punishments than less serious crimes even though

54. I assume a crime can be a combination of more distinct crimes.

their severity of punishment is lower.[55] Because states devote more resources to investigating and prosecuting more serious crimes, their probability of punishment is generally higher than the probability of punishment of less serious crimes. As a consequence, a criminal need not undertake more severe punishments for committing more serious crimes to demonstrate that she is marginally deterred. For even if the severity of punishment for two crimes were the same, the lower probability of punishment of the less serious crime would be an incentive to choose it over the more serious crime. Thus, SD does not explain why a criminal deserves to be punished more severely for committing more serious crimes.

B. A General Deterrence Theory

Critics might argue that RS is implausible because its obligation of restoration is too narrow. When someone commits a crime, she does undermine conditions of trust regarding herself. However, she also undermines conditions of trust between others. For her crime is strong evidence that others are also disposed to commit crimes. As a consequence of her crime, others are not justified in believing of each other that they are not disposed to commit crimes. More precisely, they are not justified in believing this with as high a credence as they would have been justified in believing it if she had not committed her crime. As a result, others will incur additional costs of insecurity. To avoid causing others to incur these additional costs, the criminal is obligated to restore not only the conditions of trust regarding herself, but also the ones between others that she undermined by committing her crime. Because RS claims that she is obligated to restore only the former conditions, its obligation of restoration is too narrow. To restore the latter conditions, the criminal must undertake a punishment to signal to others that they too stand to be punished if they commit crimes. Undertaking the punishment would restore the conditions of trust between others by justifying others in believing of each

55. *See* POSNER, *supra* note 52, at 245 n.4; Friedman & Sjostrom, *supra* note 52, at 363.

other that they are generally deterred from committing crimes. Call this general deterrence theory 'GD.'[56]

In response, GD does identify an important potential value of imposing a deserved punishment on a criminal: the punishment might deter others from committing some crimes in the future. But as a theory of punitive desert, GD is implausible for at least four reasons. First, when someone commits a crime, her crime is strong evidence that she is disposed to commit crimes. Because her dispositions are in some sense causally responsible for her acts, she was disposed to commit crimes at the time of her crime. And the fact that she was so disposed at the time of her crime is strong evidence that she is similarly disposed at later times. Contrary to GD, though, her crime is not strong evidence that others are disposed to commit crimes. Their dispositions were not causally responsible for her crime. And the mere fact that one person has a particular disposition is not strong inductive evidence on which to infer that others are similarly disposed. So by committing her crime, the criminal did not undermine conditions of trust between others to any significant degree. Her crime did not significantly reduce the credence with which others are justified in believing of each other that they are not disposed to commit crimes. So in response to her crime, others are not rationally required to incur any significant costs of insecurity on the grounds that they are no longer warranted in trusting each other not to commit crimes.

Second, suppose for the sake of argument that everyone knows others will commit future crimes unless a criminal is punished for her crime. The criminal is still not obligated to undertake a punishment as a means to deterring them. In general, a person is obligated to prevent herself only from causing others harm; a person is not obligated to

56. GD is similar to the trust based theories of punitive desert defended by both Dimock and Hoekema. They suggest that a criminal deserves to be punished because punishing her is necessary to restore or maintain conditions of trust regarding herself and conditions of trust between others. They suggest that punishing the criminal restores or maintains such conditions by deterring people generally from committing crimes. *See* Dimock, *supra* note 44, at 51-54; Hoekema, *supra* note 44, at 345-50. Unlike RS, neither suggests that a criminal must demonstrate that she has a good will to fulfill her obligation of restoration.

prevent herself from merely allowing others harm.[57] Presumably, if the criminal were not to undertake a punishment, she would merely allow the potential criminals to commit the future crimes, and she would merely allow everyone to incur additional costs of insecurity in response to the fact that they would not be justified in believing that the potential criminals are not disposed to commit crimes. The potential criminals would be the sole cause of both harms. For in general, a person is responsible only for whether she commits crimes and only for what others are justified in believing about her own dispositions. In general, a person is not responsible for whether someone else commits crimes or for what others are justified in believing about the dispositions of someone else.

Third, suppose for the sake of argument that a criminal does cause conditions of trust between others to be undermined or would cause them to be undermined by not undertaking a punishment. GD still does not justify PP for the same reasons SD does not. People are generally deterred from committing a crime by its expected punishment. Because it is not practically feasible to control precisely a crime's probability of punishment across all situations, it is not practically feasible to use general deterrence as the basis for determining how much to punish a criminal. In many situations, a crime's probability of punishment will be too low to justify a deserved punishment on general deterrence grounds, whereas in many others, it will be too high.

Fourth, GD violates a conceptual truth about punitive desert. According to the constraint on penal substitution, the fact that a criminal deserves a punishment cannot be annulled by others undertaking burdens on her behalf.[58] Thus, a criminal's obligation to undertake certain burdens that makes her deserving of punishment cannot be fulfilled

---

57. *See, e.g.*, Warren Quinn, *Actions, Intentions, and Consequences: The Doctrine of Doing and Allowing*, 98 PHIL. REV. 287 (1989); Samuel Scheffler, *Doing and Allowing*, 114 ETHICS 215 (2004); BERNARD WILLIAMS, *A Critique of Utilitarianism*, *in* UTILITARIANISM: FOR AND AGAINST 75, 93-95, 99 (J.J.C. Smart & Bernard Williams eds., 1973).

58. *Cf.* DAVID LEWIS, *Do We Believe in Penal Substitution?*, *in* PAPERS IN ETHICS AND SOCIAL PHILOSOPHY 128 (2000).

vicariously by others undertaking burdens on her behalf. RS satisfies the constraint on penal substitution. When a criminal undermines conditions of trust regarding herself, only she can restore those conditions by sacrificing some of her own personal interests for the sake of benefiting others. Only by making such a sacrifice can she demonstrate to others that she has a good will. So only she can fulfill her obligation of restoration under RS by undertaking the relevant burdens on her own. GD, however, violates the constraint on penal substitution. For suppose a criminal does undermine conditions of trust between others, and suppose she is obligated to restore them. Even so, others could restore those conditions on her behalf by sacrificing some of their own personal interests for the sake of benefiting others. By doing so, they would demonstrate to each other that they each have a good will. So others could vicariously fulfill the criminal's wider obligation of restoration under GD by undertaking burdens on her behalf. As a consequence, even if a criminal is obligated to restore conditions of trust between others, this obligation does not make her deserving of punishment because the obligation could be fulfilled by penal substitution.[59]

C. The Gradational Character of Conditions of Trust

Critics might argue that RS is too lenient. Consider a repeat criminal who commits two crimes at two different times. Because she undermined conditions of trust by committing her first crime, she did not undermine any by committing her second. They were already undermined when she committed her second. So her second crime does not generate an obligation of restoration. Thus, RS unacceptably entails that she does not deserve to be punished for her second crime.

In response, RS does entail that the repeat criminal deserves to be punished for her second crime because she did undermine conditions of trust by committing it. Under the

59. For similar reasons, among others, a criminal's obligation to compensate her victims for the harm she caused them does not itself make her deserving of punishment. Others could fulfill this obligation on the criminal's behalf by themselves providing her victims with the required compensation.

conditions of trust regarding someone, others are justified in believing with a particular credence that she is not disposed to commit crimes.[60] As such, conditions of trust have a gradational character and can be undermined to various degrees in various ways. By committing her first crime, the criminal undermined conditions of trust only to a degree: her first crime justifies others in believing with an unduly high credence that she has a particularly bad disposition to commit crimes. So her first crime provides others with particularly strong evidence that she has a particularly bad deficiency in the degree to which she cares about others. By committing her second crime, she undermined conditions of trust to an even worse degree: her second crime justifies others in believing with an even higher credence that she is disposed to commit crimes. Depending on the seriousness and timing of her second crime, it might also justify others in believing that she has an even worse disposition to commit crimes. For example, she might be disposed to commit more serious crimes, and she might be disposed to commit crimes with a higher frequency, in a broader range of situations, and against a broader range of people. Hence, her second crime provides others with even stronger evidence that she has the same or worse deficiency in the degree to which she cares about others. Because the repeat criminal did undermine conditions of trust by committing her second crime, RS entails that she is obligated to restore them and so deserves to be punished for her second crime.[61]

---

60. A person's credence in a belief is her subjective probability or degree of confidence that the belief is true. For discussion on the concept of subjective probabilities and their relation to objective probabilities, see DAVID LEWIS, *A Subjectivist's Guide to Objective Chance*, *in* PHILOSOPHICAL PAPERS: VOLUME II 83 (1986).

61. Assuming repeat criminals deserve to be punished more severely than first time criminals, the repeat criminal deserves to be punished more severely for her second crime than a first time criminal would deserve for committing the same crime. *Cf.* Andrew von Hirsch, *Desert and Previous Convictions in Sentencing*, 65 MINN. L. REV. 591, 593 (1981). RS might explain this by, for example, assuming the increasing marginal difficulty of restoring conditions of trust: the worse someone undermines conditions of trust the more burdens she must undertake to restore them to a degree. In this paper, though, I focus primarily on first time criminals. I set aside for further analysis the issue of how much repeat criminals deserve to be punished relative to first time criminals.

The critics might still argue that RS is too lenient. Suppose someone commits an extremely serious crime late in life. By doing so, she undermines conditions of trust to an extremely bad degree. To restore them, she must sacrifice some of her extremely important personal interests for an extremely long time for the sake of benefiting others. However, suppose she cannot make such a lengthy sacrifice given the unavoidable constraints on the duration of her life. Hence, she cannot restore the conditions of trust and so is not obligated to restore them. For if someone cannot do something, she is not obligated to do it.[62] Thus, RS unacceptably entails that she does not deserve to be punished.

In response, RS does entail that the older extremely serious criminal deserves to be punished. Although she cannot restore the conditions of trust completely, she can restore them partially and so is obligated to restore them as much as she can. Just as someone can undermine conditions of trust to a degree, she can restore them to a degree. To fulfill her obligation of restoration, she must sacrifice some of her extremely important personal interests for as long as she can for the sake of benefiting others. By doing so, she will provide others with strong evidence that she has partially rectified the extreme deficiency in the degree to which she cares about others. As a consequence, others will rationally reduce the costs of insecurity they incur in response to her crime. Given her obligation of restoration, RS entails she does deserve to be punished.

D. Timing

Critics might still suspect that RS is too lenient. In a standard case, a person's crime is detected shortly after she commits it. To restore the conditions of trust as quickly as is reasonably possible, she must demonstrate that she has acted with a high degree of benevolence for a long time after her crime. To do so, she must sacrifice some

62. This is similar to the more general principle that someone ought to do something only if she can do it. *See* 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").

of her important personal interests for a long time for the sake of benefiting others. However, suppose someone commits a crime that goes undetected for a very long time. In the interim, she avoids any criminal activity but does not act with a high degree of benevolence and so does not undertake any significant burdens. By merely avoiding criminal activity for so long, though, she restores the conditions of trust she undermined by committing her crime.[63] Thus, RS unacceptably entails that she does not deserve to be punished when her crime is eventually detected.

In response, either RS does not have this implication or it is acceptable. On the one hand, if the criminal committed a very serious crime, she cannot restore the conditions of trust by merely avoiding criminal activity for a very long time. Her crime is such strong evidence that she has such a serious deficiency in the degree to which she cares about others that she cannot annul such evidence unless she acts with a high degree of benevolence for a long time after committing her crime. To demonstrate that she has a stable disposition to care highly about others, she must sacrifice some of her important personal interests for a long time for the sake of benefiting others. Thus, if she is a very serious criminal, RS entails that she does deserve to be punished because she must undertake some burdens to restore the conditions of trust.

On the other hand, if the criminal did not commit a very serious crime, she might have restored the conditions of trust by merely avoiding criminal activity for a very long time. If so, then RS does entail that she does not deserve to be punished when her crime is eventually detected. In this case, she is merely obligated to compensate her victims for any harm she caused them. Contrary to the critics, this implication is acceptable. Its intuitive plausibility is reflected in the widespread adoption of statutes of limitation, which generally bar states from prosecuting someone for a mildly or moderately serious

63. But to emphasize, even if a criminal could restore the conditions of trust without undertaking any burdens by merely avoiding criminal activity for a very long time, doing so would standardly take too long to fulfill her obligation of restoration. For the longer a criminal takes, the longer others must incur the costs of insecurity.

crime that she committed a sufficiently long time before the prosecution would begin. RS reflects one rationale behind such statutes: they prevent states from punishing criminals who have restored the conditions of trust by avoiding criminal activity for a sufficiently long time.[64]

E. Restorative Signaling

Critics might again argue that RS is too lenient. 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.[65] The critics might contend that some restorative signals are costless and so do not require criminals to undertake any burdens. For example, a restorative signal might consist in one's merely apologizing verbally for her crime and claiming to care highly about others for a long time. If some restorative signals are costless, then RS unacceptably entails that criminals do not deserve to be punished.

In response, a restorative signal cannot be costless. By definition, a restorative signal is credible; it justifies others in believing that the criminal has acted with the

64. *See* WAYNE R. LAFAVE, JEROLD H. ISRAEL & NANCY J. KING, CRIMINAL PROCEDURE: HORNBOOK SERIES 875 (4th ed. 2004) (suggesting this rationale when they write that statutes of limitation "prevent the prosecution of those who have been law abiding for some years"). States do have other more pragmatic reasons for adopting statutes of limitation. *See id.* at 875-77. But if they are defensible on the basis of punitive desert, they are not defensible merely on the basis of "purely public policy arguments." *But cf.* Paul H. Robinson, *Criminal Law Defenses: A Systematic Analysis*, 82 COLUM. L. REV. 199, 229-30 (1982) (classifying statutes of limitation as providing "nonexculpatory public policy defenses").

65. 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 49.

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,"[66] 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 desires not to obtain.[67] According to "the non-pooling condition,"[68] 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.[69]

Any apparent restorative signal that is costless would not be credible because it would violate the non-pooling condition. If an act is costless, it is not too costly for people who are not at all benevolent. Any costless act might be performed by criminals who do not place any weight on the interests of others relative to their own in deliberating

---

66. Bacharach & Gambetta, *supra* note 49, at 160.

67. I take the concept of all-in benefits and costs from Bacharach & Gambetta, *supra* note 49, 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.

68. Bacharach & Gambetta, *supra* note 49, at 160.

69. 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 49, at 160.

about what to do. For example, criminals who do not care at all about others might merely apologize verbally for their crimes and claim to care highly about others. Thus, restorative signals cannot be costless. Because they must be costly and, hence, burdensome for the criminals obligated to send them, RS does not entail that criminals deserve no punishment.

Exactly 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>>M.[70] 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.[71] 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 others.[72] 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

---

70. 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.

71. 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.

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.

72. 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.

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≤f"<f. The all-in benefits to her would equal f"tM. The all-in costs to her would equal tS-E. Because f"≥(tS-E)/tM, tS-E≤f"tM.

Given why and how much a restorative signal must be burdensome, we can see what could and what could not standardly constitute a restorative signal. On the one

hand, a restorative signal could standardly consist in labor intensive community service in which the criminal labors in gratis to benefit others. Such labor could consist in the burdensome production, maintenance, or distribution of resources that benefit others. Such resources could range from large scale public resources, such as buildings, parks, and roads, to small scale private resources, such as food, health care, or education. The direct beneficiaries of such labor could range from the direct victims of her crime to third parties. So under RS, punishments should ideally take the form of labor intensive community service performed under reasonable conditions of incapacitation to minimize the costs of insecurity that others must rationally must incur while the service is performed.

On the other hand, a restorative signal could not standardly consist in a criminal's merely resisting the temptation to commit crimes. In a standard case of resisting temptation, a person sacrifices some of her own interests for the sake of promoting the interests of others whom she would otherwise harm. So by resisting temptation, a criminal could demonstrate that she cares about the interests of others to some degree and has the ability to restrain herself to some degree from promoting her own interests at the expense of others. Hence, resisting temptation could restore to some degree the conditions of trust she undermined by committing her crime. But that said, merely resisting temptation would standardly be neither a practical nor an effective means of fulfilling a criminal's obligation of restoration.

It would be impractical for a very serious criminal in the state's custody because the conditions under which her resisting temptation might constitute a restorative signal would standardly be too risky to others or too difficult to devise and implement once or repeatedly. For resisting temptation to constitute a restorative signal, the criminal must have an apparent opportunity to commit other crimes with impunity. To provide her with one, the state must put her in a position in which she believes can commit a crime with impunity. But given that she has already undermined conditions of trust to a very bad

degree, putting her in a position in which she actually can commit a crime with impunity would standardly pose an unacceptably high risk to her potential victims. Alternatively, devising and implementing conditions under which she cannot commit a crime with impunity but nevertheless believes she can would standardly be too difficult to do once or repeatedly, especially for very serious crimes.

Merely resisting temptation would also be an ineffective means of fulfilling a criminal's obligation of restoration in a timely manner because it would standardly not demonstrate that she has acted with a sufficiently high degree of benevolence. In a standard case of resisting temptation, a person sacrifices only her minor interests that she cares about to a minor degree. Consider standard cases of resisting the temptation to steal from, cheat, or assault another. Her sacrifice will standardly seem especially minor to her because it is a mere "foregone gain" as opposed to a loss of something to which she has been entitled. Other things being equal, a person tends to care much less about foregoing goods to which she has not been entitled than losing goods to which she has.[73]

Given that merely resisting temptation could not standardly constitute a restorative signal, critics might argue that labor intensive community service also could not. An act is a restorative signal only if it justifies others in believing that in performing it, the actor sacrificed some of her important personal interests for the sake of benefiting others. Unless the actor made the sacrifices with the right motive, her act would not justify others in believing that she is disposed to care highly about others. In performing labor intensive community service, a criminal does sacrifice some of her important personal interests. However, the critics might argue that the service could not be a restorative signal because there is no way for others to verify that the criminal performed it with the right motive. For all they know, the criminal might have performed it merely

73. This is an instance of the "endowment effect." *See* Daniel Kahneman, Jack L. Knetsch & Richard H. Thaler, *Experimental Tests of the Endowment Effect and the Coase Theorem*, 98 J. POL. ECON. 1325 (1990).

for the sake of convincing others that she is trustworthy in order to obtain the personal benefits of being trusted. Such benefits might include close friendships and highly valuable employment. Assuming labor intensive community service cannot constitute a restorative signal, then nothing can. So RS does not explain why any criminals deserve to be punished.

In response, labor intensive community service could be a restorative signal because there are two ways in which others could be justified in believing that it was performed with the right motive. First, people might have a mechanism for directly detecting the motive with which the service was performed. For example, people seem to have a mechanism for directly detecting the intention with which others perform their acts. As Oliver Wendell Holmes said, even a dog can distinguish between being kicked and merely tripped over.[74] Assuming people can directly detect the intention with which a criminal performs labor intensive community service, they might be able to detect directly the motive with which she performs it. In others words, they might be able to detect directly whether she performs it for the sake of benefiting others or for the sake of something else, like her own long-term personal interests.

Second, even if people cannot detect directly the motives with which a criminal performs labor intensive community service, they can still be justified in believing that she performed it with the right motive. For the conditions under which she performs the service might be structured such that she does not know whether others would extend her the personal benefits of trust if she were to perform the service. Assuming the criminal performs the service under such conditions of ignorance, then her performing it for the sake of benefiting others can be the inference to the best explanation of what motivated her in performing it.[75] To facilitate such conditions, others should not treat a criminal's

74. *See* O.W. HOLMES, THE COMMON LAW 3 (1881) (stating that "even a dog distinguishes being stumbled over and being kicked").

75. *Cf.* Gilbert H. Harman, *The Inference to the Best Explanation*, 74 PHIL. REV. 88 (1965).

punishment as a quid pro quo in which they promise to extend her the personal benefits of trust in exchange for her undertaking the punishment.

Suppose, though, that a criminal performs the required labor intensive community service, but she does so without the right motive. She sacrifices some of her important personal interests not for the sake of benefiting others, but for the sake of promoting her own long-term personal interests in anticipation of receiving the benefits of trust. Her service might still be a weak restorative signal in the sense that it is still some evidence she is not disposed to commit crimes. Criminals tend to be impulsive, lacking self-control.[76] They tend to act on the basis of what would best promote their own short term personal interests. So in deliberation, they tend to discount highly the value of future benefits and harms relative to the value of those more present. They tend not to make significant short term sacrifices for the sake of promoting long term gains even for themselves. Their impulsiveness significantly weakens the deterrent effect of the threat of punishment since they are much more concerned with realizing the more immediate gains of crime than avoiding the more distant harm of punishment. And as a result of their impulsiveness, they give priority to their own present interests over the interests of others.

When a criminal performs labor intensive community service even for the sake of benefiting her own long term personal interests, she demonstrates greater self-control and a lack of impulsiveness. By making significant short term sacrifices for the sake of promoting long term gains even for herself, she demonstrates that she no longer tends to act merely on the basis of what would best promote her own present interests. As a consequence, her service is still some evidence that she is no longer disposed to commit crimes. For it is evidence that the threat of punishment will have a greater deterrent effect

76. *See, e.g.*, ROBERT H. FRANK, PASSIONS WITHIN REASON: THE STRATEGIC ROLE OF THE EMOTIONS 161-62 (1988) (noting that criminals are typically impulsive); MICHAEL R. GOTTFREDSON & TRAVIS HIRSCHI, A GENERAL THEORY OF CRIME 85-120 (1990) (noting that criminals typically lack self-control).

on her, and it is evidence that she is in a better position to resist promoting her own short term interests at the expense of others. So labor intensive community service performed even from self-interested motives might restore partially the conditions of trust a criminal undermined by committing her crime.

F. The Character of Criminals

Critics might argue that RS's assumptions about the character of criminals are implausible because they are in tension with the situationist research tradition in social psychology.[77] Situationists contend that most people do not have global character traits.[78] A global character trait is such that if a person possesses it, she is highly likely to engage in trait-relevant behavior in each trait-relevant eliciting situation in which she might be.[79] Situationists also contend that most people's character is not evaluatively integrated.[80] An evaluatively integrated character consists of almost all good traits or almost all bad ones.[81] Unlike an evaluatively integrated character, most people's character contains many good traits and many bad ones.[82]

Assuming situationism, the critics might infer that most criminals do not have a global character trait of criminal deviance: a criminal is not highly likely to commit any type of crime in any situation in which she could do so. Assuming most criminals do not have a global character trait of criminal deviance, the critics might argue that no one could rationally incur the costs of insecurity to any significant degree in response to

77. On situationism, see, e.g, JOHN DORIS, LACK OF CHARACTER: PERSONALITY AND MORAL BEHAVIOR (2002); WALTER MISCHEL, PERSONALITY AND ASSESSMENT (1968); LEE ROSS & RICHARD E. NISBETT, THE PERSON AND THE SITUATION (1991); Peter B. M. Vranas, *The Indeterminacy Paradox: Character Evaluations and Human Psychology*, 39 NOUS 1 (2005).

78. *See* DORIS, *supra* note 77, at 24-25.

79. *See id.* at 19, 22.

80. *See id.* at 25.

81. *See id.* at 22.

82. *See* Vranas, *supra* note 77, at 1-3.

someone's committing a crime. So criminals do not have an obligation of restoration.

In response, RS does not rest on the implausible assumption that most criminals have a global character trait of criminal deviance. It assumes only that most criminals have a general character trait of criminal deviance: for a broad range of crimes and common situations, a criminal is too likely to perform any of them over a significant run of such situations. Unlike a global character trait of criminal deviance, a general one does not range over all possible crimes and situations. It ranges over only a broad subset of crimes and situations in which a criminal will commonly find herself.[83] Unlike a global character trait of criminal deviance, a general one does not assume that a criminal is highly likely to commit a crime in any particular situation. Indeed, it does not even assume that a criminal is highly likely to commit a crime in any run of situations. It merely assumes that a criminal is too likely to commit one in a significant run of common situations.[84] What is too likely is what is so likely that others are rationally required to incur the costs of insecurity to a significant degree. Because crimes are so harmful to others, what is too likely could be significantly less than what is highly likely, assuming a high probability is greater than 0.5. Even if the probability of a criminal's committing a crime over a significant run of common situations is less than 0.5, the expected harm of her doing so could still be fearful.

Assuming most criminals have a general character trait of criminal deviance, criminals have an obligation of restoration because they will cause others to incur the costs of insecurity to a significant degree by not restoring the conditions of trust they

---

83. 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 full exculpatory defense, her specific crime is strong evidence that she is disposed to commit a broader range of crimes comparable in seriousness.

84. Unlike a global character trait of criminal deviance, a general one concerns the probability of a person's committing a crime over a run or aggregate of situations rather than in a single situation. On the role of aggregation in determining a person's general character traits, see, e.g, Seymour Epstein, *Aggregation and Beyond: Some Basic Issues on the Prediction of Behavior*, 51 J. PERSONALITY 360 (1983).

undermined by committing their crimes. Three empirical findings support RS's assumption about the character of criminals. First, there is a high rate of recidivism among criminals: a high percentage of criminals commit multiple crimes at different times.[85] 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.[86] Third, studies indicate that a high percentage of criminals have a stable and broad serious deficiency in the degree to which they care about others.[87]

Critics might concede that most criminals have a general character trait of criminal deviance. But they still might deny that a criminal has an obligation of

---

85. *See, e.g.*, Patrick A. Langan & David J. Levin, *Recidivism of Prisoners Released in 1994*, Bureau of Justice Statistics Special Report, *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); GOTTFREDSON & HIRSCHI, *supra* note 76, at 107-08, 177, 230-31, 253 (noting the high stability of criminals' dispositions to commit crimes, and citing numerous research studies in support); *cf.* Dan Olweus, *Stability of Aggressive Reaction Patterns in Males: A Review*, 86 PSYCHOL. BULL. 852 (1979) (finding a high stability in people's dispositions to engage in antisocial aggressive behavior).

86. *See, e.g.*, GOTTFREDSON & HIRSCHI, *supra* note 76, 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); Alex Piquero, Raymond Paternoster, Paul Mazerolle, Robert Brame & Charles W. Dean, *Onset Age and Offense Specialization*, 36 J. RES. CRIME & DELINQ. 275, 275-76 (1999) (same, stating "[r]esearchers investigating the sequencing of offense types over time in criminal offending have generally found that offenders exhibit some specialization amid a great deal of versatility"); 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).

87. *See, e.g.*, GOTTFREDSON & HIRSCHI, *supra* note 76, at 89 (stating that criminals tend to be acutely "self-centered, indifferent, or insensitive to the suffering and needs of others"); Joshua D. Miller & Donald Lynam, *Structural Models of Personality and their Relation to Antisocial Behavior: A Meta-Analytic Review*, 39 CRIMINOLOGY 765 (2001) (finding that criminals tend to be acutely unconcerned about the interests of others); Michael J. Vitacco & Craig S. Neumann, Angela A. Robertson & Sarah L. Durrant, *Contributions of Impulsivity and Callousness in the Assessment of Adjudicated Male Adolescents: A Prospective Study*, 78 J. PERSONALITY ASSESSMENT 87 (2002) (same); *cf.* Larry Alexander, *Insufficient Concern: A Unified Conception of Criminal Culpability*, 88 CAL. L. REV. 931 (2000) (arguing that the mens rea of a crime involves having an insufficient concern for the interests of others); Joan McCord, *Understanding Motivations: Considering Altruism and Aggression*, *in* FACTS, FRAMEWORKS, AND FORECASTS 115, 126 (Joan McCord ed., 1992) (stating that "crime is a consequence of motives to injure others or to benefit oneself without a proper regard to the welfare of others").

restoration because there is nothing she could do to demonstrate that she no longer has such a trait. Assuming situationism, most people have a character that is evaluatively integrated to an extremely low degree. So the fact that someone has one good trait is not strong evidence that she has any other good trait. As a consequence, there is nothing a criminal could do that would be strong evidence that she has a good will. She can demonstrate that she has a stable disposition to care highly about others by sacrificing some of her sufficiently important personal interests for a sufficiently long time for the sake of benefiting others. But the critics might argue that even this is not strong evidence of a stable disposition to be appropriately motivated by the moral reasons against violating the rights of others.

In response, the critics overestimate the limits on the evaluative integration of most people's character. There are limits. But none undermines the claim that a stable disposition to care highly about others is strong evidence of a good will. So none undermines the claim that a criminal has an obligation of restoration. Consider two limits. First, a stable disposition to care highly about others is not strong evidence of a perfect will in the sense of a stable disposition to govern oneself perfectly in every possible situation. Almost anyone is disposed to be moderately rude or unkind to others in certain stressful situations. And almost anyone is disposed to perform acts of very serious wrongdoing in certain extraordinary situations in which almost anyone would "lose her moral compass."[88] These situations involve an exculpatory defense. But a stable disposition to care highly about others is still strong evidence of a good will in the sense of a stable disposition not to disrespect the rights of others in situations not involving an exculpatory defense. For violating the rights of others would standardly cause them serious harm, and a stable disposition to care highly about others naturally

88. John Sabini & Maury Silver, *Lack of Character? Situationism Critiqued*, 115 ETHICS 535, 560 (2005). One such situation might involve the state's demanding the performance of wrongful acts with the widespread support of its citizens.

generates a stable disposition not to harm them so seriously. Moreover, the state normally issues a standing authoritative demand not to violate people's rights through its laws. For someone with a stable disposition to care highly about others, the state's demand only strengthens her disposition not to violate them by anchoring more securely her proper moral bearing.[89] So although a stable disposition to care highly about others is not identical with a stable disposition not to disrespect the rights of others, the former is still strong evidence of the latter.[90]

Second, a stable disposition to care highly about some is not necessarily strong evidence of a stable disposition to care highly about others or, therefore, not to commit crimes against them. For example, consider someone who commits a hate crime against those of a particular race, sex, religion, sexual orientation, nationality, or socioeconomic status. Although she might have a stable disposition to care highly about some, she does not care highly about those in her disfavored group. She can, though, still demonstrate that she has a good will because she can still demonstrate that she has a stable disposition not to commit crimes against those in her disfavored group. To do so, she must demonstrate that she has a stable disposition to care highly about them. She can demonstrate this by sacrificing some of her sufficiently important personal interests for a sufficiently long time for the sake of benefiting those in the disfavored group.

G. The Death Penalty

Critics might argue that RS is too harsh because it unacceptably entails that criminals standardly deserve the death penalty. According to the critics, to fulfill her

89. *Cf.* STANLEY MILGRAM, OBEDIENCE TO AUTHORITY 179-89 (1974) (discussing the strong influence of a state's demands on the motivations of its citizens).

90. Thus, it is logically possible that someone could be disposed both to care highly about someone and to disrespect her rights. Such a person might be in the grip of an overly consequentialist or paternalistic conception of morality. In practice, though, criminals tend not to manifest such a motivational deficiency. In practice, the disposition to care highly about someone naturally generates a disposition to respect her rights, especially those rights protected by the state's law. Further confirmation for this assumption must await the results of longitudinal studies in the social sciences that track the long term behavior of criminals who manifest a high degree of benevolence after committing their crimes.

obligation of restoration, a criminal must justify others in believing with certainty that she is no longer disposed to commit crimes. To do so, she must undertake the death penalty.

In response, RS does not entail that a criminal must justify others in believing with certainty that she is no longer disposed to commit crimes. In other words, RS does not entail that a criminal must restore the conditions of trust to a degree of certainty. To fulfill her obligation of restoration, she must restore the conditions of trust only to the baseline degree. In other words, she must justify others in believing that she is no longer disposed to commit crimes only with the credence with which they would have been justified in believing this if she had not committed her crime. For if others were justified in believing with the baseline credence that she is no longer disposed to commit crimes, they would no longer be rationally required to incur additional costs of insecurity in response to her crime. Although the baseline credence must be sufficiently high, it is not certainty. No one is justified in being certain that any living person is not disposed to commit crimes. So to restore the conditions of trust to the baseline degree, the criminal need only justify others in believing with the baseline credence that she has a good will. And to do so, she need not undertake the death penalty.

The critics might still argue that RS unacceptably entails that criminals standardly deserve the death penalty. After a criminal demonstrates that she has a good will, she would no longer cause others to incur the costs of insecurity. However, during the time required to demonstrate that she has a good will, she would cause others to incur such costs even if she were incapacitated. For example, if she were incapacitated in a state prison, she would cause her fellow inmates and prison officials to incur such costs. To avoid imposing such costs on others, she must undertake the death penalty immediately after she commits her crime. Thus, she must undertake the death penalty to fulfill her obligation of restoration.

In response, to mitigate the costs of insecurity that she causes others to incur, a criminal must incapacitate herself for the time required to demonstrate that she has a good

will. But RS does not entail that she is obligated to undertake the death penalty to prevent herself from causing others to incur the costs of insecurity while she is incapacitated. A person's obligation not to cause others harm has a limit. Under the limit, a person is obligated to undertake the means necessary to prevent herself from causing others harm only if the means are not extremely worse for her than the harm she would otherwise cause them. Under standard conditions, the death penalty would be extremely worse for a criminal than the individual or aggregate costs of insecurity she would otherwise cause others to incur while she is incapacitated. As a further consequence, a criminal does not standardly deserve to be incapacitated under the conditions that would literally minimize the degree to which others would be vulnerable to her. Such conditions, which might involve straightjackets and various forms of sensory deprivation, would also standardly run afoul of the limit at issue.

H. The Entailment

Critics might concede that RS explains why a criminal has an obligation of restoration. But they might deny that it explains why she deserves to be punished. To explain why a criminal deserves to be punished, a theory must explain why the state would not violate her rights by punishing her against her will. Punishing a criminal against her will, though, would not fulfill her obligation of restoration because to fulfill it, she must undertake the punishment voluntarily. Because punishing a criminal against her will would not fulfill her obligation of restoration, RS does not explain why she deserves to be punished.

In response, it is true that punishing a criminal against her will would not fulfill her obligation of restoration. But RS still explains why a criminal deserves to be punished on the basis of her obligation of restoration. RS explains why a criminal must undertake certain burdens as a means to fulfilling the obligation of restoration she incurs from committing her crime. For three related reasons, which we discussed earlier, the criminal deserves these burdens as a punishment for her crime. First, she is obligated to

undertake the burdens. She has a derivative obligation to undertake the means necessary to fulfill her obligation of restoration. Second, she is obligated to consent to undertake the burdens. Third, she would be unjustly enriched by not undertaking the burdens. For by not undertaking them, she would not only violate her obligation of restoration, but also obtain a benefit from the violation consisting in her freedom from the burdens necessary to fulfill it. Each of these related claims entails that the criminal has no right to be free from the burdens necessary to fulfill her obligation of restoration. So if she refuses to undertake the burdens voluntarily and violates her obligation of restoration, the state would not violate her rights by imposing them on her as a punishment against her will.

I. One Problem with Defiant Criminals

Critics might argue that RS does not explain why a defiant criminal deserves a punishment that the state could impose on her. A defiant criminal unconditionally refuses to undertake any punishment of any form. So the state cannot impose on her a punishment with a restorative form, such that her voluntarily undertaking it with the right motive would fulfill her obligation of restoration. Because a punishment with a restorative form consists in labor intensive community service, the state can impose such a punishment only on a criminal who is at least minimally cooperative. Assuming a defiant criminal deserves only a punishment with a restorative form, RS does not explain why she deserves a punishment that the state could impose on her.

In response, RS does entail that fully cooperative criminals deserve only punishments with a restorative form. Because they are willing to undertake such punishments to fulfill their obligation of restoration, they have a right to undertake them instead of punishments with a non-restorative form. Unlike fully cooperative criminals, however, defiant criminals deserve even punishments with a non-restorative form, which might consist in a mere burdensome form of incapacitation. Because a punishment with a restorative form is not possible for a defiant criminal, she will be unjustly enriched unless she receives a punishment with a non-restorative form whose severity is proportional to

the severity of the burdens she must undertake to fulfill her obligation of restoration. Imposing such a punishment on her is the only means of preventing her from receiving the benefit of being free from the burdens necessary to fulfill her obligation of restoration. Because a defiant criminal would be unjustly enriched unless she receives a punishment with a non-restorative form, she deserves such a punishment.

J. A Second Problem with Defiant Criminals

Critics might argue that RS is too harsh because it unacceptably entails that any defiant criminal deserves an endless stream of punishment. Consider a defiant moderately serious criminal. Suppose the state punishes her as severely as the burdens she must undertake to fulfill her obligation of restoration. She did not fulfill the obligation by undertaking the punishment because she did not undertake it voluntarily. So even after undertaking it, she retains the same obligation of restoration and, therefore, retains an obligation to undertake an additional punishment. More generally, she will always retain an obligation to undertake an additional punishment no matter how much the state punishes her because she will always refuse to undertake the punishments voluntarily and so will always retain the same obligation of restoration. As a consequence, she is obligated to undertake an endless stream of punishment. Thus, RS unacceptably entails that she deserves an endless stream of punishment even though as a moderately serious criminal, she deserves no more than a moderately severe punishment.

In response, RS does not entail that every defiant criminal deserves an endless stream of punishment. Reconsider the defiant moderately serious criminal. For at least three reasons, she does not deserve an endless stream of punishment under RS. First, her obligation of restoration only entails that she lacks a right to be free from a punishment that is no more severe than the least severe burdens she must undertake to fulfill it. As a moderately serious criminal, she can fulfill her obligation of restoration by undertaking only a moderately severe burden. So she deserves no more than a moderately severe punishment on the basis of her obligation of restoration. Because an endless stream of

punishment, considered collectively, is more severe than a moderately severe burden, she does not deserve it on the basis of the obligation.

Second, after the state punishes her moderately severely, the defiant moderately serious criminal no longer retains the obligation of restoration she incurred from committing her crime. For assuming the state does not compensate her for the harm she suffered from the punishment, the state is not warranted in demanding her to undertake any more burdens to restore the conditions of trust she undermined by committing her crime. If the state were to demand more, then collectively it would demand the criminal to undertake more burdens than necessary to restore the conditions of trust. And the state is warranted in demanding her to undertake only the burdens necessary to restore them. As a consequence, the state is not warranted in further demanding her to restore the conditions of trust. Therefore, she is no longer obligated to restore them. She merely ought to.

Third, suppose for the sake of argument that the defiant moderately serious criminal must undertake an endless stream of punishment to restore the conditions of trust she undermined by committing her crime. Unless she undertakes such a punishment, she will cause others to incur additional costs of insecurity. Nevertheless, presumably she is not obligated to prevent herself from causing others to incur such costs by undertaking an endless stream of punishment. Such an endless stream would be extremely harmful to her, whereas the individual or aggregate costs of insecurity to others will presumably be far from constituting an extremely severe harm because she committed only a moderately serious crime. Thus, an endless stream of punishment would presumably be extremely worse for her than the individual or aggregate costs of insecurity to others. So she is not obligated to undertake an endless stream of punishment to restore the conditions of trust. Therefore, she does not deserve such a punishment for her crime.

Although RS does not entail that any defiant criminal deserves an endless stream of punishment, it does entail that defiant criminals deserve to be punished more severely

than fully cooperative criminals who committed the same crimes. Assuming a defiant criminal and a fully cooperative criminal committed the same crimes, they undermined conditions of trust to the same degree by committing their crimes. Therefore, they do deserve the same punishment for their crimes. However, the defiant criminal undermined conditions of trust to an additional degree by being defiant in response to her crime. So she deserves an additional punishment for being defiant. But the additional punishment she deserves would not amount to an endless stream of punishment because being defiant in response to a crime is not itself an extremely serious crime and so it does not undermine conditions of trust to an extremely bad degree.[91]

The hard question now is what to do with a defiant criminal after she receives all the punishment she deserves. RS is consistent with two important policies for dealing with such cases. One the one hand, if the defiant criminal is so untrustworthy that the state knows she will commit additional very serious crimes unless she is incapacitated, then the state might be justified in incapacitating her longer on grounds of self-defense or preventative detention. The additional term of incapacitation, though, would not be an additional punishment. It would be a term of civil commitment, like the quarantine of someone with an infectious, dangerous disease. As such, the state would be obligated to minimize the degree to which the additional term would be burdensome for her. On the other hand, if she is not so untrustworthy, then incapacitating her longer might not be justified on grounds of self-defense. In this case, the state must provide her with at least a conditional release. For the state would no longer be warranted in demanding her to sacrifice anymore to prevent herself from causing others to incur additional costs of insecurity. So upon her release, others might need to monitor her on a regular basis. They might need to forgo some valuable activities that would leave them too vulnerable

91. In refusing to fulfill her obligation of restoration, a defiant criminal does disrespect to some degree the rights of others to be free from the costs of insecurity that they rationally must incur because she refuses to fulfill it. But because her fulfilling the obligation would require her to undertake a severe burden, the disrespect manifested is not extremely bad.

to her. And they might need to live with the additional fear of her committing crimes against them.

Some might endorse a third policy for dealing with a defiant criminal after she receives all the punishment she deserves. Under this policy, the state and its citizens would simply trust her not to commit more crimes. The state would not incapacitate her longer, and others would not incur any additional costs of insecurity. In defense of the policy, people are positively responsive to trust because they desire to preserve others' esteem.[92] By trusting her, the state and its citizens would express an attitude of esteem toward her that she might desire to preserve by not committing more crimes. So by simply trusting her not to commit more crimes, the state and its citizens would provide her with a prudential incentive not to commit more.

This third policy is indefensible. In general, people are positively responsive to trust, and partly for this reason, people are justified in trusting someone not to commit crimes if they know she has not committed any. But someone's committing a crime is strong evidence not only that she is not appropriately motivated by the moral reasons against committing crimes, but also that her desire for others' esteem is an insufficient incentive for her not to commit crimes. Thus, the fact that a defiant criminal might have some desire for others' esteem does not justify others in simply trusting her not to commit more crimes. Unless she fulfills her obligation of restoration, the state and its citizens are rationally required to incur the costs of insecurity or incapacitate her longer.

K. The Authority to Punish

Critics might contend that RS does not explain why the state has the exclusive authority to punish criminals in the following sense. First, the state would not, but a private citizen would, violate a criminal's rights by punishing her. Second, a criminal may not resist the state's attempt to punish her, but may resist a private citizen's attempt.

---

92. *See, e.g.*, Philip Pettit, *The Cunning of Trust*, 24 PHIL. & PUB. AFF. 202, 212-17 (1995).

Third, a private citizen may not interfere with the state's attempt to punish a criminal, but the state may interfere with a private citizen's attempt.

In response, it is not clear that the plausibility of a theory of punitive desert depends on whether it can account for the state's exclusive authority to punish. Nevertheless, RS has the resources to account for it by justifying three related principles. First, the state would not violate a criminal's substantive right against intentional harm by punishing her. Second, the state's criminal procedure would not, but a private citizen's criminal procedure would, violate a criminal's procedural rights. Third, the state's criminal procedure would, but a private citizen's criminal procedure would not, be fair to all its citizens. A criminal procedure is a procedure for determining how much a criminal deserves to be punished and for imposing a punishment on her.

Consider the first principle. If someone is obligated to others to undertake certain burdens, then the others' agent can impose those burdens on her without violating her substantive right against intentional harm. Given that a criminal owes her obligation of restoration to her fellow citizens, she is obligated to them to undertake the burdens necessary to fulfill it. Assuming the state acts on behalf of its citizens in a way that is fairly responsive to their demands, the state is their agent. Thus, the state can impose such burdens on the criminal as a punishment without violating her substantive right against intentional harm.

Consider the second principle. Because a criminal procedure risks imposing on a criminal an undeserved punishment, a criminal has a right to a procedure that is reliable in both a minimal and relative sense.[93] A procedure is minimally reliable as applied to a criminal if and only if its risk of imposing on her an undeserved punishment is sufficiently low. A procedure P1 is relatively reliable as applied to a criminal if and only if there is no other available procedure P2 such that a) P2's risk of imposing on her an

93. *See* NOZICK, *supra* note 6, at 96-101.

undeserved punishment is lower than P1's risk, and b) she has a right to P2 over P1. On standard assumptions, the state's criminal procedure would, but a private citizen's procedure would not, be both minimally and relatively reliable. Thus, the former would not, but the latter would, violate a criminal's procedural rights.[94]

Two aspects of the state's criminal procedure standardly make it minimally reliable, and their absence from a private citizen's procedure standardly make it not minimally reliable. First, only the state's procedure is supported by a robust set of resources to determine the facts of a case. To determine how much a criminal deserves to be punished and the form of punishment she deserves, the state has ready access to the beliefs of many experts and a jury, which represents the beliefs of a wide cross-section of its citizens. The state also has a robust set of resources to provide a criminal with the form of punishment she deserves, such as a restorative form. Second, only the state's procedure is governed by the rule of law.[95] Under the rule of law, the state actors who administer the state's procedure are generally impartial and constrained by sentencing guidelines. Given that some state actors will inevitably be prejudiced, the guidelines limit the negative effects of their prejudice on their contribution to the state's procedure.

Even if a private citizen's criminal procedure is minimally reliable, it is not relatively reliable. The state's procedure has a lower risk of imposing on a criminal an undeserved punishment because it is governed by the rule of law and supported by a more robust set of resources. A criminal has a right to the state's procedure over a private citizen's because the former's being less risky generates a prima facie claim to it over the latter that is not defeated by any countervailing considerations. A consideration that could defeat such a prima facie claim would be the fact that the less risky procedure

---

94. The state's procedure is not necessarily minimally and relatively reliable. In arguing that it is, I presuppose the state is a well-working democracy operating under favorable conditions.

95. *See* LON L. FULLER, THE MORALITY OF LAW 33-94 (rev. ed. 1964) (describing aspects of the rule of law).

would be infeasible to implement or use to impose on a criminal a deserved punishment. Because a criminal is obligated to others to undertake a deserved punishment, she does not have a claim to a less risky criminal procedure if it would be so infeasible. The state's procedure, though, need not be infeasible.

Consider the third principle. Under a criminal's obligation of restoration, she is obligated to the state's citizens as a group to undertake just one punishment as a means to restore the conditions of trust. For two reasons, she is not obligated to each citizen to undertake a separate punishment. First, doing so would be unnecessary to restore the conditions of trust. One punishment would suffice. Second, if she were to undertake a separate punishment for each citizen, the resulting harm to her could be extremely worse than the individual or aggregate costs of insecurity she stands to cause others to incur. Thus, a criminal owes her obligation to undertake a punishment to the state's citizens as a group. So the right to punish is held jointly by the state's citizens.[96] As a jointly held right, only the state, as the agent of all its citizens, can exercise it in a way that would be fair to all its citizens. A private citizen would exercise the right unilaterally, whereas the state would exercise it on behalf of all its citizens in a way that is fairly responsive to all their demands. Hence, the state's criminal procedure would, but a private citizen's procedure would not, be fair to all its citizens.

Together these three principles entail that the state has the exclusive authority to punish criminals. First, the state would not violate a criminal's rights by punishing her because it would not violate her substantive or procedural rights by doing so. However, a private citizen would violate a criminal's procedural rights by punishing her. Second, a criminal may not resist the state's attempt to punish her because the state would not violate her rights by doing so, and she is obligated to the state and its citizens to undertake a punishment. However, a criminal may resist a private citizen's attempt to

96. *Cf.* NOZICK, *supra* note 6, at 139 (claiming that the right to punish is held jointly rather than individually by people in the state of nature).

punish her in order to protect her procedural rights and to prevent him from unfairly exercising the right to punish unilaterally. Third, a private citizen may not interfere with the state's attempt to punish a criminal because his doing so would interfere with the fair exercise of the right to punish held jointly by the state's citizens. However, the state may interfere with a private citizen's attempt to punish a criminal in order to protect her procedural rights and to prevent him from unfairly exercising the right to punish unilaterally.[97]

L. Punishable Crimes

RS explains why someone deserves to be punished for committing crimes against others, which involve disrespecting their rights. However, critics might contend that RS does not explain why someone does not deserve to be punished for committing crimes against herself or harmless crimes.[98] A crime against self involves the criminal flouting only the reasons that count against harming herself.[99] A crime against self might consist in the state's prohibition on the use of certain drugs. A harmless crime involves the criminal not flouting any reasons that count against harming anyone.[100] A harmless crime might consist in the state's prohibition of certain acts violating certain sexual taboos.

In response, RS does explain why someone does not deserve to be punished for crimes against herself or harmless crimes because it explains why such a criminal is not obligated to others to undertake a punishment. When someone commits such a crime, she does provide others with strong evidence that she is disposed to commit such crimes.

---

97. There also might be more pragmatic reasons to vest the state with the exclusive authority to punish. Doing so minimizes the risk of private citizens' engaging in endless cycles of retaliation in response to their endless disagreement over what constitutes an excessive use of force. *See* JOHN LOCKE, SECOND TREATISE OF GOVERNMENT, secs 13, 124, 125 (C.B. Macpherson ed., Hackett 1980); NOZICK, *supra* note 6, at 10-12.

98. *See, e.g.*, J.S. MILL, ON LIBERTY 9 (Elizabeth Rapaport ed., Hackett 1978) (suggesting that people have a right against being punished for such crimes).

99. *Cf.* JOEL FEINBERG, HARMS TO SELF (1986).

100. *Cf.* JOEL FEINBERG, HARMLESS WRONGDOING (1988).

However, she does not provide others with strong evidence that she is disposed to commit crimes against others. For in committing such a crime, she does not disrespect the rights of others. So she would not cause others to incur the costs of insecurity or any other harm by not demonstrating that she is no longer disposed to commit such crimes. Thus, she does not incur an obligation to undertake any punishment from committing those crimes.

The critics might argue that someone who commits a crime against herself deserves to be punished because she is obligated to herself to undertake a punishment as a means to deterring herself from committing more crimes against herself. Unless she deters herself, she will harm herself. And she is obligated to herself not to harm herself.

In response, all the critics' assumptions here are objectionable. But even if we concede that a person who commits a crime against herself is obligated to herself to undertake a punishment, this obligation does not entail that she deserves to be punished because it does not give the state the authority to punish her against her will. Because she owes the obligation only to herself, only she and her agent have the authority to punish her. If the state were to punish her against her will, it would not be acting as her agent because it would not be acting in a way that is responsive to her demands. As a consequence, the state would violate her rights by punishing her for the crime against herself.

Critics might argue that someone who commits a crime against herself deserves to be punished because she is obligated to others to undertake a punishment as a means to deterring them from committing crimes against themselves. Unless she undertakes a punishment, others will commit crimes against themselves because they will be undeterred from doing so. So unless she undertakes a punishment, she will cause others harm. And she is obligated to others not to cause them harm.

In response, the critics' argument is unsound for at least three reasons. First, by not undertaking a punishment, someone who commits a crime against herself would not

cause others to commit crimes against themselves. At most, she would merely allow them to. In general, she is not responsible for whether others commit crimes against anyone. So by not undertaking a punishment, she would not cause others harm. And so she is not obligated to others to undertake a punishment as a means to deterring them from committing crimes against themselves.

Second, suppose by not undertaking a punishment, she would cause others to commit crimes against themselves. Even so, she is still not obligated to them to undertake a punishment as a means to deterring them from committing such crimes. For she is only obligated to them not to cause them harm against their will. Even if they were undeterred from committing crimes against themselves, they would not commit such crimes against their will.

Third, suppose she is obligated to others to undertake a punishment as a means to deterring them from committing crimes against themselves. In the standard case, this obligation does not entail that she deserves to be punished because it does not give the state the authority to punish her against her will. Because she owes the obligation only to those willing to commit crimes against themselves, only they and their agent have the authority to punish her against her will. If the state were to punish her, it would not standardly do so as their agent because it would not be acting in a way that is responsive to their demands. For those willing to commit crimes against themselves do not standardly demand the state to punish others for such crimes. They standardly demand the contrary.

M. Two Constraints

Critics might argue that RS is implausible because it violates two constraints on any plausible theory of punitive desert. According to the first constraint, a plausible theory must explain why someone deserves to be punished for committing a crime. According to the second constraint, a plausible theory must explain why a criminal deserves to be punished on the basis of the properties that essentially make her crime

wrongful. Any theory violating either constraint is too far removed from the essence of crimes to be plausible. RS violates the first constraint because it explains why someone deserves to be punished only for violating her obligation of trust. RS violates the second constraint because it explains why a criminal deserves to be punished only on the basis of the properties that make her crime a violation of her obligation of trust.[101]

In response, RS satisfies both constraints. First, RS does explain why someone deserves to be punished for violating her obligation of trust. Such a violation, though, consists in her committing a crime. So RS explains why someone deserves to be punished for committing a crime. Second, RS does explain why a criminal deserves to be punished on the basis of the properties that make her crime a violation of her obligation of trust. Such properties, though, consist in part in the properties that essentially make her crime wrongful, namely her disrespecting the rights of others. Hence, RS explains why a criminal deserves to be punished on the basis of the properties that essentially make her crime wrongful. It is true that RS's explanation of why a criminal deserves to be punished for her crime is mediated by facts that go beyond the crime's essence. However, this does not make RS too far removed from it. Any plausible theory of punitive desert must be mediated by such facts on pain of being explanatorily vacuous.

VII. Two Concluding Remarks

First, in addition to explaining why criminals deserve to be punished, RS also explains why innocent people, who have not performed any wrongful acts, do not deserve any punishment. In virtue of being innocent, they have not undermined any conditions of trust. So they do not have an obligation of restoration. More precisely, they have not undermined any conditions of trust to the sufficiently bad degree necessary to incur an obligation of restoration whose fulfillment would require them to undertake a

101. R.A. Duff raises a similar objection against unfair advantage theories of punitive desert. *See* R.A. DUFF, TRIALS AND PUNISHMENTS 211-13 (1986).

punishment.[102] Assuming they are not obligated to undertake a punishment, they do not deserve one.

Second, RS determines whether a punishment satisfies the desert requirement of my general theory of the justification of punishment. However, it does not play a comprehensive role in determining whether a punishment satisfies the general rights requirement or the value requirement. RS does identify a valuable aim of imposing a deserved punishment on a fully cooperative criminal. But it does not identify any valuable aims of imposing a deserved punishment on a defiant criminal. Thus, a deserved punishment might or might not be justified. On the one hand, it might be justified because it would not violate the rights of others and would promote some sufficiently valuable aims, like deterrence.[103] On the other hand, it might not be justified on either ground.[104] So RS plays an important but also limited role in determining whether a punishment is all things considered justified.

---

102. *See* GOTTFREDSON & HIRSCHI, *supra* note 76, at 259-61 (noting the serious inaccuracies in the best available strategies for predicting future criminal behavior when applied before any criminal acts take place).

103. Although considerations of deterrence do not explain why or how much criminals deserve to be punished, they do identify an important potential value of imposing deserved punishments on them. *Cf.* Douglas N. Husak, *Why Punish the Deserving?*, 26 NOUS 447, 459-62 (1992) (arguing that the state is justified in imposing a deserved punishment on a criminal only if the punishment results in a sufficiently valuable reduction in crime).

104. For example, someone might deserve to be punished for committing the crime of defamation, consisting in libel or slander. However, punishing her might not be justified because doing so might violate the rights of others or promote other bad consequences by having a chilling effect on people's freedom of expression. *See* Garrison v. Louisiana, 379 U.S. 64 (1964) (holding that a Louisiana criminal defamation statute is unconstitutional under the First Amendment, and noting its general chilling effect on people's freedom of expression). To avoid such a chilling effect, defamation might be a mere tort under which a person who commits it is obligated to compensate her victims for any reputational harm.

Chapter 2

On the Optimal Enforcement of a Criminal Law

I. Introduction

For any crime, the state must set its probability and severity of punishment.[1] A crime's probability of punishment is the probability that the state would detect and punish someone who commits it.[2] Its severity of punishment is the severity of the punishment that the state would impose on someone for committing it.[3] On some fairly plausible assumptions, the state has strong reason to adopt an extreme enforcement policy for all crimes, even a moderately serious one. Under such a policy, the crime's probability of punishment would be extremely low, and its severity of punishment would be extremely high. Call such a policy as applied to a moderately serious crime 'the EEP.' Although the EEP seems strongly unreasonable, it is not immediately obvious why given the reasons in favor of it. Herein lies a problem.

To see the reasons in favor of the EEP, consider four assumptions. First, suppose deterrence is an important valuable aim of a system of punishment, which consists in the state's threatening to punish anyone who violates its criminal laws and the state's actually punishing those it finds to have violated those laws. A system of punishment promotes

1. More precisely, the state must set a crime's probability of punishment in those situations where this is feasible. Although the state cannot control a crime's probability of punishment in all the situations in which people will find themselves, it might be able to control it in some.

2. At this level of description, a crime's probability of punishment is doubly ambiguous. First, it could refer to an objective or a subjective probability. Second, if it refers to a subjective probability, it could refer to either the state's or the individual's subjective probability that the state would detect and punish her if she were to commit it. Throughout the paper, I use the term to refer to the subjective probability of both the state and the individual. I assume they are the same. I also assume that a crime's probability of punishment corresponds to the rate at which the state would detect and punish those who commit it.

3. I assume the severity of a punishment corresponds to how much someone would find it burdensome.

deterrence by providing people with a prudential incentive not to commit crimes for fear of being detected and punished for committing them.

Second, suppose people are expected value maximizers in the sense that they perform an act only if its expected value for them is at least as high as the expected value for them of any other available act. To determine the expected value for someone of a particular act, determine the possible sets of consequences of the act and the value for her of those sets of consequences. Call the value for her of each set of consequences, respectively, V1, V2, ... and Vn. Next determine her subjective probability that each set of consequences will actually result from the act. Call her subjective probability that each set of consequences will actually result from the act, respectively, P1, P2, ... and Pn. The expected value for her of the act is the sum of the products V1*P1, V2*P2, ... and Vn*Pn.

Third, suppose a crime's probability of punishment varies directly with the amount of resources that the state devotes to enforcing the law against it. Specifically, suppose the state can lower a crime's probability of punishment by devoting fewer resources to enforcing the law, and the state can raise the crime's probability of punishment only by devoting more resources to enforcing it. For example, the state might lower the crime's probability of punishment by hiring fewer police officers to search for those who might commit it, and the state might raise the probability of punishment by hiring more police officers.

Fourth, suppose a scarcity of important resources unavoidably obtains in the state. So there are not enough resources to satisfy all the critical needs of everyone, like the relief of suffering. Given the scarcity, if one enforcement policy requires less resources to implement than another, then the resources saved under the former policy could be used to satisfy some people's critical needs that would otherwise go unsatisfied. For example, if one policy requires fewer police officers than another, then under the former policy, the people who would otherwise work as police officers might work in health care to relieve the suffering of those who would otherwise continue to suffer due to a scarcity of health

care services.

Given these assumptions, the state has strong reason to adopt the EEP rather than a more moderate enforcement policy for the moderately serious crime at issue.[4] To achieve the crime's optimal level of deterrence, the state must ensure that committing it has the optimal expected value for potential criminals. To ensure this, the state must ensure that the crime's expected punishment equals the optimal magnitude, where its expected punishment is the product of its probability and severity of punishment. To set the crime's expected punishment at the optimal level, the state can set its probability of punishment as low as possible so long as it sufficiently raises its severity of punishment. Hence, assuming the crime's expected punishment equals the optimal magnitude under the EEP, the EEP would be the most efficient means to achieving the crime's optimal level of deterrence.[5] For the EEP would require the least resources to set the crime's probability of punishment because it would be the lowest under the EEP. And overall the EEP would not require any additional resources to set the crime's severity of punishment because although the EEP would require the most severe punishments, it would also require the fewest punishments since the fewest criminals would be punished under it. Given the conditions of scarcity, the resources saved under the EEP could be used to satisfy some people's critical needs that would otherwise go unsatisfied under a more moderate enforcement policy.[6]

---

4. Under a more moderate policy, the crime's severity of punishment would be lower and its probability of punishment would be higher than they would be under the EEP.

5. As Jeremy Bentham suggests, one enforcement policy might be better than another insofar as it would be a more efficient means of deterring people from committing the crime at issue: "The last object is, whatever the mischief be, which it is proposed to prevent, to prevent it at as *cheap* a rate as possible." JEREMY BENTHAM, AN INTRODUCTION TO THE PRINCIPLES OF MORALS AND LEGISLATION 165 (J.H. Burns & H.L.A. Hart eds., 1970).

6. For more discussion on the considerations of efficiency that favor an extreme enforcement policy for a crime, see Gary S. Becker, *Crime and Punishment: An Economic Approach*, 76 J. POL. ECON. 169 (1968); George J. Stigler, *The Optimum Enforcement of Laws*, 78 J. POL. ECON. 526 (1970); ROBERT COOTER & THOMAS ULEN, LAW AND ECONOMICS 427-54 (3d ed. 2000); JEFFRIE G. MURPHY & JULES L. COLEMAN, PHILOSOPHY OF LAW: AN INTRODUCTION TO JURISPRUDENCE 211-13 (rev. ed. 1990); MITCHELL A. POLINSKY, AN INTRODUCTION TO LAW AND ECONOMICS 75-

In spite of the strong reasons in favor of the EEP, it still seems strongly unreasonable to punish moderately serious criminals extremely severely even though doing so might be the most efficient means to achieving the crime's optimal level of deterrence.[7] The problem, though, is to explain why.[8] In this paper, I explain why the EEP would be strongly unreasonable on the assumption that some moderately serious criminals would inevitably be undeterred and punished extremely severely under it. Initially, I consider explanations grounded in the value requirement of my general theory of the justification of punishment. I argue that none succeeds. Then I consider

---

86 (2d ed. 1989); RICHARD A. POSNER, ECONOMIC ANALYSIS OF LAW 242-50 (5th ed. 1998).

7. Throughout the paper, I assume the moderately serious criminals at issue are first time criminals who have no exculpatory defenses. If they were repeat criminals, it might seem reasonable to punish them extremely severely. If they had a defense, it might be unreasonable to punish them at all.

8. At the outset, I dismiss three pragmatic objections to the EEP based on how people might behave in response to it. Some might argue that the state could not adopt the EEP. If it tried, its officials and private citizens would nullify the policy because punishing moderately serious criminals extremely severely would seem so unreasonable. *See* Becker, *supra* note 6, at 184; Jerome Michael & Herbert Wechsler, *A Rationale of the Law of Homicide II*, 37 COLUM. L. REV. 1261, 1264-65 (1937). To nullify the EEP, judges and juries might refuse to find anyone guilty of the crime; prosecutors might refuse to charge anyone with it; police officers might refuse to arrest anyone for the crime; and private citizens might refuse to cooperate in the investigation and trial of anyone charged with the crime. I dismiss this objection for two reasons. First, the EEP might not be nullified. Second, even if the EEP were nullified because it would seem strongly unreasonable, we seek to understand precisely why the EEP would be strongly unreasonable.

Some might argue that the EEP would not deter anyone from committing the crime because whatever the crime's severity of punishment, no one would be deterred by the crime's extremely low probability of punishment. *See* Michael & Wechsler, *supra*, at 1264. I dismiss this objection for two reasons. First, people might be deterred by the crime's extremely low probability of punishment if its severity of punishment were sufficiently high. Second, the EEP would still seem strongly unreasonable even if it were an effective means of deterrence. We seek to understand why it would be so unreasonable even under such conditions.

Some might argue that the EEP would result in the commission of more serious crimes because the state could not achieve "marginal deterrence" under it. POSNER, *supra* note 6, at 245; *see* BENTHAM, *supra* note 5, at 165. Because the state would punish the moderately serious crime extremely severely under the EEP, people would have no incentive to choose it over more serious crimes as a means to fulfilling their objectives. I dismiss this objection because there are two ways the state could achieve marginal deterrence under the EEP. First, there are a gradation of extremely severe punishments. Even though the state would punish the moderately serious crime extremely severely under the EEP, it could still punish more serious crimes even more severely. Second, even if the state were to impose the most severe punishment for the moderately serious crime under the EEP, it could raise the probability of punishment for more serious crimes. Either way the state would ensure that the expected punishment for more serious crimes is higher than the expected punishment for the moderately serious crime under the EEP. In doing so, the state would provide people with an incentive to choose the latter over the former as a means to fulfilling their criminal objectives.

explanations grounded in the desert requirement of my general theory. I argue that my restorative signaling theory of punitive desert, RS, best explains why the EEP would be strongly unreasonable. I conclude that RS should play an important role in constraining the means by which the state may enforce its laws against any crime.

II. Two Requirements of Justified Punishment

As we discussed in Chapter 1, the state is justified in imposing a punishment on someone against her will only if the punishment would satisfy the desert requirement and the value requirement of my general theory of the justification of punishment. According to the desert requirement, the person must deserve the punishment. 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 any other available act that would not violate anyone's rights.

Assuming the state would punish some against their will under the EEP, both requirements are prima facie plausible grounds on which to explain why the EEP would be strongly unreasonable. The EEP might violate the desert requirement because moderately serious criminals might not deserve to be punished extremely severely. The EEP might violate the value requirement because the expected value of the consequences of punishing moderately serious criminals extremely severely under it might be lower than the expected value of the consequences of punishing them less severely under a more moderate enforcement policy. We will examine each requirement in turn.

III. The Value Requirement

To determine whether the EEP would violate the value requirement, call a moderate enforcement policy as applied to the moderately serious crime at issue 'the MEP.' Under the MEP, the crime's probability of punishment would be moderately high, and it would be punished only moderately severely. Assume the crime's expected punishment would be the same under both the EEP and the MEP. So assume both policies would achieve the same level of deterrence. Finally, assume the consequences of

the EEP and the MEP would be better than the consequences of any other available enforcement policy for the crime. On these assumptions, to show that the EEP would violate the value requirement, we must show that the EEP's consequences would be all things considered worse than the MEP's consequences. To do so, we must identify some respect in which the former would be worse than the latter.

A. An Egalitarian Argument

Egalitarians might argue that the EEP would violate the value requirement because the EEP would be worse than the MEP with respect to inequality.[9] They assume inequality is intrinsically bad in the sense that it would be intrinsically bad if any two like criminals were punished with different severities or if one were punished but the other were not.[10] Assuming inequality is intrinsically bad, they might argue that the EEP would be all things considered worse than the MEP because the EEP's inequality would be all things considered worse than the MEP's inequality. To see why, note that the EEP's consequences would differ from the MEP's consequences in two significant ways. First, relative to the MEP, fewer criminals would be punished and more criminals would go unpunished under the EEP because the crime's probability of punishment would be lower. Second, the criminals who would be punished under the EEP would be punished much more severely than the criminals who would be punished under the MEP because the crime's severity of punishment would be much higher under the EEP. Egalitarians might conclude that these differences would make the EEP's inequality all things considered worse than the MEP's inequality.[11]

---

9. Isaac Ehrlich emphasizes some of the inequalities that would result from extreme enforcement policies in *The Optimum Enforcement of Laws and the Concept of Justice: A Positive Analysis*, 2 INT'L REV. L. & ECON. 3, 10-11 (1982).

10. I refer only to "telic egalitarians" who claim we should prevent inequalities because they would be intrinsically bad. Derek Parfit, *Equality or Priority?*, The Lindley Lecture, University of Kansas 3-4 (Nov. 21, 1991). I do not refer to "deontic egalitarians" who claim we should prevent inequalities for moral reasons other than the fact that they might be intrinsically bad. *Id.* at 4.

11. For a more detailed defense of this claim, see Appendix.

In response, the egalitarian argument is unsound for at least three reasons. First, inequality is not intrinsically bad.[12] To illustrate the implausibility of assuming otherwise, suppose several like criminals each commit a moderately serious crime. The state apprehends all of them and imposes on them moderately severe punishments in the form of moderately long periods of incarceration. However, at the end of one criminal's period of incarceration, the state mistakes her for a much more serious criminal and tortures her. At this point, the state must choose between a) torturing the others or b) releasing them after their moderately long periods of incarceration.

Given the choice, there are two reasons the state should release the others rather than torture them. First, the state's torturing them would violate their right not to be tortured even if torturing them would be optimific. In short, they do not deserve to be tortured. Second, even if the others deserve to be tortured, we can plausibly assume that the consequences of torturing them would be all things considered worse than the consequences of releasing them. For we can plausibly assume that the extreme suffering caused by torturing them would make the consequences of doing so all things considered worse than the consequences of releasing them.

However, although the state should release the others for these reasons, if inequality were intrinsically bad, then there would be something bad about releasing them rather than torturing them. For if inequality were intrinsically bad, there would be something bad about punishing like criminals with different severities. Because the state has already tortured the one, it would punish the others with a different severity by simply releasing them, whereas it would punish all of them the same by torturing the others.

But nothing seems bad about releasing the others rather than torturing them: releasing them would not be in any way worse than torturing them. To support this claim, we might appeal to our considered intuitions or a plausible principle. When we reflect on

12. My argument against the claim that inequality is intrinsically bad parallels "the Levelling Down Objection" that Parfit raises against telic egalitarians more generally. *See id.* at 17-18.

the possibility of the state's releasing the others, intuitively nothing seems bad about its doing so: nothing seems bad about benefiting them under conditions in which benefiting them would not harm anyone else. Releasing the others would ensure that not all like criminals receive punishments of the same severity. But nothing seems bad about the mere pattern of unequally severe punishments that would result.

According to a plausible principle, if a) two states of affairs S1 and S2 contain exactly the same people, and b) no one in S2 is in any way worse off than she would be in S1, then S2 is not in any way worse than S1. Call this welfare principle 'WP'. WP's plausibility stems from the fact that if its antecedent conditions are satisfied, and no other states of affairs are obtainable, then there seems nothing for anyone to regret about moving from S1 to S2. If there is nothing for anyone to regret about moving from S1 to S2, S2 is not in any way worse than S1.

Given WP, we may assume that the states of affairs that would result from either torturing or releasing the others would contain exactly the same people. And we may assume that releasing the others would not make anyone in any way worse off than she would have been if they had been tortured. Most notably, releasing the others would make them significantly better off than they would have been if they had been tortured. And, presumably, releasing the others would not in any way worsen the condition of the one whom the state has already tortured. So WP entails that the state of affairs that would result from releasing the others would not be in any way worse than the state of affairs that would result from torturing them even though the former would contain an inequality absent from the latter. As a consequence, inequality is not intrinsically bad.

In summary, the egalitarian argument is unsound because inequality is not intrinsically bad. So even if the EEP would realize more inequality than the MEP, the additional inequality by itself would not make the EEP worse in any way than the MEP.

Second, even if inequality were intrinsically bad, the egalitarian argument is still unsound because it does not show that the EEP's overall inequality would be all things

considered worse than the MEP's overall inequality. The argument focuses only on inequalities between the criminals whom the EEP and the MEP specifically target. However, both policies could also affect the inequalities between people outside the targeted class of criminals. These inequalities would likely be all things considered worse under the MEP than under the EEP. For the resources saved under the EEP could be used to improve inequalities between people outside the targeted class of criminals. Given conditions of scarcity, such inequalities would likely persist under the MEP. Because the resources saved under the EEP could improve inequalities that would persist under the MEP, the EEP's overall inequality might be all things considered better than the MEP's overall inequality even though the inequality among the targeted criminals would be worse under the EEP than under the MEP.

Third, even if the EEP's overall inequality were all things considered worse than the MEP's overall inequality, the egalitarian argument still does not show that the EEP would be all things considered worse than the MEP. Even if inequality is intrinsically bad, it is not the only thing that is intrinsically bad. Suffering is also intrinsically bad, and increases in suffering can offset improvements in inequality. For two reasons, the EEP would likely realize less overall suffering than the MEP. First, the EEP and the MEP would realize the same overall amount of suffering among the targeted class of criminals. For they would realize the same overall amount of punishment among them because the crime's expected punishment would be the same under both.[13] Second, the EEP would likely realize less suffering than the MEP among people outside the targeted class of criminals. For the resources saved under the EEP could be used to prevent the suffering of people outside the targeted class. Thus, even if the EEP's overall inequality were all

13. The policies would differ insofar as the overall amount of punishment under the EEP would consist in a relatively small number of criminals receiving extremely severe punishments, whereas the overall amount under the MEP would consist in a relatively large number of criminals receiving moderately severe punishments. However, such differences would only affect the way that the overall amount of punishment and suffering would be distributed among the targeted criminals: they would not affect the overall amount itself.

things considered worse than the MEP's overall inequality, the EEP might be all things considered better than the MEP because the EEP might realize sufficiently less overall suffering than the MEP.

B. A Prioritarian Argument

Prioritarians might concede that inequality is not intrinsically bad, and the EEP would realize less overall suffering than the MEP. But they might argue that the EEP would still violate the value requirement because collectively people's suffering would be worse under the EEP than under the MEP. According to prioritarians, for any particular amount of suffering a person experiences, her suffering that amount is worse the worse her life considered as a whole.[14] To illustrate, suppose two people suffer the same amount at some time. Suppose also that one person's life considered as a whole is worse than the other person's life considered as a whole. According to prioritarians, the former's suffering is worse than the latter's suffering even though they suffer the same amount.

To show that collectively people's suffering would be worse under the EEP than under the MEP, prioritarians might focus on showing that the suffering of the affected class would be worse under the EEP than under the MEP. The affected class consists in the people whose suffering would be affected by the choice between the EEP and the MEP. Other things being equal, the affected class consists in the targeted criminals who would be punished under either policy, people closely related to them, and the people whose suffering would be prevented by the resources saved under the EEP.

Prioritarians might assume that among the affected class, the lives of the criminals who would receive extremely severe punishments under the EEP would be the worst. Assuming their lives would be the worst, prioritarians might argue that the suffering of the affected class would be worse under the EEP than under the MEP. Even though the affected class would suffer less under the EEP than under the MEP, the criminals who

14. *See* Parfit, *supra* note 10, at 19-22.

would be punished extremely severely under the EEP would have the worst lives, and therefore, their suffering any particular amount under the EEP would be worse than anyone in the affected class suffering that amount under the MEP.

In response, prioritarianism is a plausible view about the badness of suffering. However, the argument does not show conclusively that the EEP would violate the value requirement because it does not show conclusively that the suffering of the affected class would be worse under the EEP than under the MEP. Among the affected class, the people who stand to have the worst lives might not be the criminals who would be punished extremely severely under the EEP: they might be the people under the MEP whose suffering would have been prevented by the resources saved under the EEP. Even if not, the mere fact that a person has a worse life than another does not make the former's suffering some amount worse than any amount of suffering that the latter might experience. The former's suffering some amount might be better than the latter's suffering a sufficiently larger amount. Similarly, the mere fact that a person has a worse life than others does not make her suffering some amount worse than any amount of suffering that the others might experience collectively. Her suffering some amount might be better than the others' suffering collectively a sufficiently larger amount. Thus, even if the criminals who would be punished extremely severely under the EEP would have the worst lives among the affected class, the suffering of the affected class might still be better under the EEP than under the MEP because the affected class might suffer sufficiently less overall under the EEP than under the MEP.

C. An Absolutist Argument

Absolutists concede that the mere fact that a person has a worse life than others does not make her suffering some amount worse than any amount of suffering that the others might experience individually or collectively. But they might still argue that the suffering of the affected class would be worse under the EEP than under the MEP. According to absolutists, some types of suffering are so much more severe than other

types that the former types are absolutely worse than the latter types in the following sense: a person's suffering some amount of the former types would be worse than any number of other people's suffering any amount of the latter types.

To show that the suffering of the affected class would be worse under the EEP than under the MEP, absolutists might assume that the criminals punished extremely severely under the EEP would experience an extremely severe type of suffering that would be much more severe than the types of suffering anyone in the affected class would experience under the MEP. They might infer that the extremely severe type of suffering would be absolutely worse than the latter types of suffering. So they might conclude that the suffering of the affected class would be worse under the EEP than under the MEP no matter how much more the affected class would suffer as a whole under the MEP.

In response, the absolutist argument is also unpersuasive for at least two reasons. First, the people who stand to experience the worst type of suffering might not be the criminals who would be punished extremely severely under the EEP: they might be the people under the MEP whose suffering would have been prevented by the resources saved under the EEP. Second, the argument rests on the unacceptable assumption that some types of suffering are absolutely worse than less severe types of suffering. The assumption is unacceptable because if it were true, it would unacceptably entail that the relation "is all things considered worse than" is intransitive.[15] To see why, consider an extreme type of suffering that might seem absolutely worse than a mild type of suffering. For example, suppose the extreme type of suffering consists in several years of extremely intense pain. Call this type 'A.' Suppose the mild type of suffering consists in just a few minutes of mildly intense pain. Call this type 'Z.' Suppose also that A is absolutely worse than Z such that one person's suffering A is all things considered worse than any number

15. My argument for why the absolutist assumption has this implication parallels an argument that Larry Temkin formulates in *A Continuum Argument for Intransitivity*, 25 PHIL. & PUB. AFF. 175, 179-81 (1996).

of people's suffering Z. Now consider someone's suffering A. Presumably, for some number x>1, x people's suffering B, which is only slightly less severe than A, would be all things considered worse than one person's suffering A. For example, suppose B consists in pain that is only slightly less intense than A and lasts for only slightly less time than A. Presumably, for some number x_2>x, x_2 people's suffering C, which is only slightly less severe than B, would be all things considered worse than x people's suffering B. For example, suppose C consists in pain that is only slightly less intense than B and lasts for only slightly less time than B. And presumably, for some number x_3>x_2, x_3 people's suffering D, which is only slightly less severe than C, would be all things considered worse than x_2 people's suffering C. For example, suppose D consists in pain that is only slightly less intense than C and lasts for only slightly less time than C.

At this point, we can see that by repeating the previous line of reasoning enough times, we can construct a chain from one person's suffering A to some large number of people's suffering Z such that each group's suffering in the chain would be all things considered worse than the group's suffering that immediately precedes it in the chain.[16] Given the chain, if A were absolutely worse than Z, then the relation "is all things

16. In reply, absolutists might argue that no matter how many times we repeat the previous line of reasoning, we could never extend the chain from one person's suffering A to some number of people's suffering Z. Ken Binmore and Alex Voorhoeve suggest such an argument in *Defending Transitivity against Zeno's Paradox*, 31 PHIL. & PUB. AFF. 272 (2003). Following their argument, absolutists might contend that after repeating the previous line of reasoning innumerable times, the chain would converge to some limit of suffering that would be more severe than Z. As the chain approaches the limit, the number of people's experiencing the types of suffering close to the limit would increase to infinity. And no number of people's experiencing the types of suffering at the limit or after it would be all things considered worse than the relevant number of people's experiencing the types of suffering before the limit. As a consequence, the chain would not extend to Z no matter how many times we repeat the previous line of reasoning.

In response, the claim that the chain might converge to some limit seems implausible. If the chain converges to some limit of suffering of type M, then there is a type of suffering N such that N is only slightly less severe than M but no number of people's experiencing N would be all things considered worse than a particular number of people's experiencing M. But to the contrary, it seems that for any two types of suffering M and N, if N is only slightly less severe than M, then for any number x of people's experiencing M, there is some number y>>x such that y people's experiencing N is all things considered worse than x people's experiencing M. As a consequence, it seems implausible to assume that the chain converges to some limit. Thus, it seems that by repeating the previous line of reasoning enough times, we could extend the chain from one person's suffering A to some large number of people's suffering Z.

considered worse than" would be intransitive. For if the relation were transitive, then A would not be absolutely worse than Z because the large number of people's suffering Z would be all things considered worse than the one person's suffering A. Because presumably we could construct a similar chain between one person's experiencing any type of severe suffering and a much larger number of people's experiencing any type of less severe suffering, the absolutist assumption entails that the relation "is all things considered worse than" is intransitive.

This relation, though, does not seem intransitive.[17] Because the absolutist assumption unacceptably entails that it is, the assumption itself is unacceptable. Contrary to the absolutist argument, the most severe type of suffering experienced in the affected class under the EEP would not be absolutely worse than the types of suffering experienced in the affected class under the MEP. Hence, the suffering of the affected class might be all things considered better under the EEP than under the MEP if the affected class were to suffer sufficiently less overall under the EEP.

D. A Retributive Argument

Given the objections to the prioritarian and absolutist arguments, retributivists might contend that the EEP would violate the value requirement on very different grounds.[18] Both the prioritarian and absolutist arguments assume that anyone's suffering

17. Assuming the relation "is all things considered worse than" is intransitive would conflict with a canon of practical reasoning. According to the canon, given any set of acts available to an agent, at least one act in the set is choiceworthy in the sense that the agent has at least as much reason to perform it as she has to perform any of the other acts in the set. However, if the relation "is all things considered worse than" were intransitive, then given some set of acts available to an agent, none of the acts might be choiceworthy. For example, consider a case in which among the acts available to an agent, an act would be choiceworthy if and only if its consequences would be at least as good as the consequences of any other available act. Not every case is like this, but some are, especially when no one's rights are at stake. If the relation "is all things considered worse than" were intransitive, then it is possible that for any available act, its consequences would be all things considered worse than the consequences of some other available act. Hence, contrary to the canon, none of the acts available to the agent would be choiceworthy if the relation "is all things considered worse than" were intransitive.

18. Ehrlich mentions some of the retributive costs of an extreme enforcement policy in *supra* note 9, at 18-20.

in the affected class under either policy would be intrinsically bad. However, retributivists assume that a criminal's suffering could be intrinsically good. Depending on the seriousness of her crime, they assume there is some optimal amount of suffering that it would be intrinsically good that she experience.[19] To the extent she suffers any less or any more, it would be intrinsically bad.

Retributivists might assume that a moderate amount of suffering would be the optimal amount for moderately serious criminals. So for three reasons, the suffering of the targeted class of criminals would be worse under the EEP than under the MEP. First, fewer criminals would suffer the optimal amount under the EEP than under the MEP. Second, more criminals would suffer less than the optimal amount under the EEP than under the MEP. Third, more criminals would suffer more than the optimal amount under the EEP than under the MEP. All three claims hold given that a) fewer criminals would be punished under the EEP than under the MEP because the crime's probability of punishment would be lower under the EEP, and b) all the criminals punished under the EEP would suffer too much because they would be punished extremely severely, whereas all the criminals punished under the MEP would suffer only the optimal amount because they would be punished only moderately severely. Assuming the suffering of the targeted class of criminals would be worse under the EEP than under the MEP, retributivists might infer that the EEP would be all things considered worse than the MEP.

In response, the retributive argument is unpersuasive for three reasons. First, it rests on the repugnant assumption that it is intrinsically good that criminals suffer. On reflection, no one's suffering any amount could be intrinsically good; anyone's suffering any amount under any conditions would be intrinsically bad.[20] Suffering could be

19. *See* Michael S. Moore, *Justifying Retributivism*, 27 ISRAEL 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).

20. *See* T.M. SCANLON, WHAT WE OWE TO EACH OTHER 274 (1998).

instrumentally good in the sense that it would have good effects. For example, someone's suffering might prevent her from boarding an airplane destined to crash. A criminal's suffering might have the good effect of deterring herself or others from committing crimes in the future. But no one's suffering could be good in itself.

Second, at best, if a criminal's suffering some amount were intrinsically good, it would be so only if she were strongly responsible for her crime.[21] A criminal would be strongly responsible for her crime only if no factors beyond her control by themselves causally determined her committing the crime in the sense of making her committing it inevitable. But on a plausible view of the relation between any actual criminal and the natural world, factors beyond her control by themselves did causally determine her committing the crime. According to the view, for any actual criminal, states of the world prior to her birth in conjunction with the laws of nature causally determined her committing the crime. Because the laws of nature and the states of the world prior to someone's birth are beyond her control, factors beyond any actual criminal's control causally determined her committing the crime. Hence, no actual criminal is strongly responsible for her crime. Therefore, no actual criminal's suffering any amount is intrinsically good.[22]

Third, even if the retributive assumption is true, the argument still does not show conclusively that the EEP would violate the value requirement. It would show only that the suffering of the targeted class of criminals would be worse under the EEP than under the MEP. However, the resources saved under the EEP might be used to prevent enough suffering among innocent people outside the targeted class of criminals to make the suffering of the affected class considered as a whole better under the EEP than under the

21. Parfit alludes to this point and the following problem with it in *supra* note 10, at 32-33.

22. In clarification, this objection to retributivism is not a defense of hard determinism; it is not meant to undermine the possibility that criminals are morally responsible for their crimes in the sense that they are blameworthy and deserve to be punished for them. It only undermines the claim that it is intrinsically good that criminals suffer.

MEP.

At this point, we may conclude that the value requirement does not provide a robust explanation of why the EEP would be strongly unreasonable. On the one hand, the EEP might very well satisfy the requirement. No considerations decisively prove otherwise. On the other hand, even if the EEP would violate it, the violation would not explain why the EEP would be *strongly* unreasonable. For the EEP to be strongly unreasonable on the basis of the value requirement, the EEP must not only violate it, but also be a particularly egregious violation of it such that its consequences would be not just worse, but *much* worse than the consequences of the MEP. However, even if the EEP's consequences would be worse than the consequences of the MEP, they would not be much worse in light of the fact that a) the resources saved under the EEP could be used to prevent a substantial amount of suffering among innocent people outside the targeted class of criminals and b) the targeted class of criminals would suffer substantially under both policies.

IV. The Desert Requirement

On its face, the desert requirement is a more promising basis on which to explain why the EEP would be strongly unreasonable because moderately serious criminals seem to deserve no more than moderately severe punishments. If so, then the EEP would be not only a violation of the desert requirement, but also a particularly egregious violation of it. For under the EEP, the state would punish moderately serious criminals not just more severely than they deserve, but much more. The challenge is to explain why moderately serious criminals deserve to be punished no more than moderately severely.[23]

A. An Expressive Theory

According to expressivists, a criminal does not deserve a punishment if the punishment would express an unwarranted attitude of moral blame toward her. They

23. In considering theories of punitive desert here, I make the same presuppositions I defended in Chapter 1.

might argue that moderately serious criminals do not deserve the EEP's extremely severe punishments because such punishments would express too much blame toward them. They might defend the claim in two ways.

According to the first argument, under the EEP the state would punish moderately serious criminals just as severely as it punishes criminals who commit very serious crimes, such as very violent crimes. In fact, the state does tend to punish very serious criminals extremely severely, and under the EEP, the state would also punish moderately serious criminals extremely severely. Because under the EEP the state would punish moderately and very serious criminals extremely severely, expressivists might infer that under the EEP the state would punish them equally severely. By punishing them equally severely, the state would express toward moderately serious criminals the attitude that they are just as blameworthy as very serious criminals.[24] However, moderately serious criminals are not as blameworthy as very serious criminals. Thus, punishing moderately serious criminals extremely severely under the EEP would express too much blame toward them.

In response, the first expressive argument does not show that the EEP's extremely severe punishments would express too much blame toward moderately serious criminals. The argument falsely assumes that under the EEP the state must punish moderately and very serious criminals equally severely. Although under the EEP the state would punish both extremely severely, it need not punish them equally severely. For there is a gradation of extremely severe punishments: some extremely severe punishments are less severe than other extremely severe punishments. To avoid expressing toward moderately serious criminals the unwarranted attitude that they are just as blameworthy as very serious criminals, the state could impose on moderately serious criminals under the EEP

24. H.L.A. Hart suggests this point when he asserts that if a state were to punish with equal severity crimes that differ in seriousness, then "there is a risk of either confusing common morality or flouting it and bringing the law into contempt." H.L.A. HART, *Prolegomenon to the Principles of Punishment, in* PUNISHMENT AND RESPONSIBILITY: ESSAYS IN THE PHILOSOPHY OF LAW 1, 25 (1968).

extremely severe punishments that are less severe than the extremely severe punishments it imposes on very serious criminals.

Expressivists might argue that even if the state were to punish moderately serious criminals less severely than very serious criminals under the EEP, it would still express too much blame toward moderately serious criminals by punishing them extremely severely. They might claim that the absolute severity of a criminal's punishment expresses an absolute degree of blame toward her. Punishing a criminal extremely severely expresses an extremely high degree of blame toward her. By punishing moderately serious criminals extremely severely under the EEP, the state would express too much blame toward them because they are only moderately blameworthy, not extremely blameworthy. Hence, moderately serious criminals do not deserve to be punished extremely severely under the EEP.

In response, the second expressive argument is circular. As we discussed in Chapter 1, an attitude of moral blame consists at least partly in certain demands. One demand that is constitutive of moral blame is the demand to undertake a punishment. Punishing someone expresses blame toward her by expressing its constitutive demand on her to undertake the punishment. Under the EEP, punishing moderately serious criminals extremely severely would express an extremely high degree of blame toward them by expressing a demand on them to undertake an extremely severe punishment. To claim that the EEP's extremely severe punishments would express too much blame toward moderately serious criminals is to claim that the punishments would express an unwarranted demand on them to undertake such extremely severe punishments. But to claim that a demand on the criminals to undertake an extremely severe punishment would be unwarranted presupposes that they do not deserve such a punishment. The claim that moderately serious criminals do not deserve the EEP's extremely severe punishments is precisely the claim to be explained. Because the second expressive argument

presupposes the claim it aims to explain, it is circular.[25]

B. A Restorative Signaling Theory

As I argued in Chapter 1, RS explains why criminals deserve to be punished by explaining why they must undertake a punishment to fulfill the obligation of restoration they incur from committing their crimes. According to RS, when someone commits a crime, she undermines certain conditions of trust in the sense that she undermines the conditions that are necessary for others' being justified in believing that she is not disposed to commit crimes. She is obligated to restore those conditions because unless she restores them, she will cause others to incur the costs of insecurity.

To restore the conditions of trust, she must demonstrate to others that she has a good will in the sense of a stable disposition to be appropriately motivated by the moral reasons against violating the rights of others. To demonstrate that she has a good will, she must demonstrate that she has a stable disposition to care highly about the interests of others. To demonstrate this, she must demonstrate that she has acted with a sufficiently high degree of benevolence for a sufficiently long time after committing her crime. 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

---

25. To avoid this circularity objection, expressivists might contend that punishing moderately serious criminals extremely severely would be inappropriate because such punishments would express the unwarranted assertion that they disrespected the rights of others to an extremely bad degree in committing their crimes. In fact, they disrespected the rights of others only to a moderately bad degree. *Cf.* SCANLON, *supra* note 20, at 267 (claiming that "[i]nsofar as punishment involves an assertion that the agent governed him-or herself in a way that was faulty, it is appropriate only when this is true").

For at least three reasons, this is not a good explanation of why the criminals would not deserve to be punished extremely severely under the EEP. First, the EEP's extremely severe punishments would not express the false assertion that the moderately serious criminals disrespected the rights of others to an extremely bad degree. Unlike a mere utterance, an assertion is the expression of a belief. Because the state does not believe that the moderately serious criminals disrespected the rights of others to an extremely bad degree, it does not assert this in punishing them extremely severely under the EEP. At most, the EEP's punishments would express a mere utterance of the false proposition. Second, it is not clear people have a right against the mere utterance of a false proposition, especially under conditions in which no believes it. Assuming the state is transparent about the reasons of efficiency for which it adopts the EEP, these conditions could obtain under the EEP. Third, even if people have a right against the mere utterance of a false proposition, the violation of such a right does not account for the gravity of an undeserved punishment.

benefiting others.

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.[26]

Given RS, we can now explain why moderately serious criminals do not deserve to be punished extremely severely under the EEP. 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. When someone commits a moderately serious crime, she 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 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. As a consequence, she would not deserve to

26. 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 it. 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 it.

be punished extremely severely under the EEP.

V. Three Objections

A. First Objection

Critics might concede that RS correctly explains why moderately serious criminals deserve moderately severe punishments. But drawing on the form of RS's explanation, they might construct a similar argument for why moderately serious criminals also deserve the EEP's extremely severe punishments. Because the EEP would likely realize less overall suffering than more moderate enforcement policies, critics might claim that the EEP would be optimific in the sense that its consequences would be all things considered better than the consequences of any other available enforcement policy. Assuming the EEP would be optimific, punishing moderately serious criminals extremely severely under the EEP would be optimific in the sense that the consequences of doing so would be all things considered better than the consequences of treating them in any other way available. Assuming people have an obligation of extreme beneficence to undertake any optimific treatment, critics might infer that moderately serious criminals are obligated to undertake the EEP's extremely severe punishments. So contrary to RS, moderately serious criminals deserve them.

In response, people do not have an obligation of extreme beneficence. Such an obligation would demand too much from them. If a person had such an obligation, she would be obligated to undertake extremely severe burdens merely because the consequences of her doing so would be all things considered *slightly* better than the consequences of her not doing so. For example, suppose a person could prevent someone from suffering years of intense pain only by undertaking a course of treatment that would cause her the same intense pain but for slightly less time. Other things being equal, if she had an obligation of extreme beneficence, she would be obligated to undertake this painful course of treatment merely because the consequences of her doing so would be all things considered slightly better than the consequences of her not doing so. On reflection,

though, no one seems to have such a strong obligation to sacrifice her own critical interests whenever doing so would be merely optimific. In general, an individual seems justified in caring about her own critical interests out of proportion to their objective value. So within limits, an individual can be justified in promoting her own critical interests even when promoting them would not be optimific.[27] Thus, people do not have an obligation of extreme beneficence.[28] Therefore, the first objection does not show that moderately serious criminals are obligated to undertake the EEP's extremely severe punishments.

B. Second Objection

Critics might point out that a criminal's obligation of restoration is ultimately grounded in her more basic obligation not to cause others harm.[29] A person seems specially responsible for the consequences she causes rather than merely allows to occur.[30] So although a person is not generally obligated not to allow others harm, she is obligated not to cause others harm. The critics might argue that a moderately serious criminal is obligated to undertake the EEP's extremely severe punishment because unless she undertakes it, she will cause others harm.

---

27. *See, e.g.*, THOMAS NAGEL, THE VIEW FROM NOWHERE 171-75 (1986); SAMUEL SCHEFFLER, THE REJECTION OF CONSEQUENTIALISM 20 (rev. ed. 1994).

28. A person does have an obligation of minimal beneficence to make a trivial sacrifice to relieve a dire burden. But this obligation could not explain why someone would be obligated to undertake the EEP's extremely severe punishment because such a punishment would not be a *trivial* sacrifice. A person might have a more demanding obligation to undertake an extremely severe burden when the consequences of her doing so would be all things considered *extremely better* than the consequences of her not doing so. But this obligation would also not explain why a moderately serious criminal would be obligated to undertake the EEP's extremely severe punishment. For there is little reason to believe that the EEP's consequences would be all things considered *extremely better* than the consequences of a more moderate policy.

29. Unless a criminal restores the conditions of trust she undermined by committing her crime, she will cause others harm by causing them to incur the costs of insecurity.

30. *See, e.g.*, Warren Quinn, *Actions, Intentions, and Consequences: The Doctrine of Doing and Allowing*, 98 PHIL. REV. 287 (1989); Samuel Scheffler, *Doing and Allowing*, 114 ETHICS 215 (2004); BERNARD WILLIAMS, *A Critique of Utilitarianism, in* UTILITARIANISM: FOR AND AGAINST 75, 93-95, 99 (J.J.C. Smart & Bernard Williams eds., 1973).

Suppose the state adopts the EEP and apprehends a moderately serious criminal. If she does not undertake the EEP's extremely severe punishment, she will effectively lower the crime's severity of punishment for other potential criminals by causing them to believe that they too stand to suffer a less severe punishment for committing the crime. By effectively lowering the crime's severity of punishment for other potential criminals, she will cause others either of two harms.

On the one hand, the state might raise the crime's probability of punishment for others to compensate for their belief that they now stand to suffer a less severe punishment for the crime. By doing so, the state would ensure that they are deterred from committing the crime just as strongly as they would have been if she had not effectively lowered the crime's severity of punishment for them. To raise the crime's probability of punishment for others, though, the state must incur additional enforcement costs. So if the state were to raise the crime's probability of punishment for others, she would cause harm to the state by causing it to incur such costs.

On the other hand, to avoid incurring these additional costs, the state might not raise the crime's probability of punishment for others. As a consequence, others would commit the crime who would have been deterred from committing it if she had not effectively lowered the crime's severity of punishment for them. So if the state were not to raise the crime's probability of punishment for others, she would cause harm to the victims of their additional crimes.

In response, we might concede that if the moderately serious criminal does not undertake the EEP's extremely severe punishment, then either the state will incur additional enforcement costs or other criminals will commit crimes that they would have been deterred from committing it if she had not effectively lowered the crime's severity of punishment for them. However, the critics falsely assume that she would cause either harm to occur by not undertaking the EEP's extremely severe punishment. On reflection, by not undertaking it, she would merely allow either harm to occur. In either case, the

harm would be caused solely by other actual or potential criminals.

On the one hand, suppose the state does not raise the crime's probability of punishment for others, and some commit the crime who would have been deterred from committing it if she had not effectively lowered the crime's severity of punishment for them. Although more people are harmed as victims of these additional crimes, she did not cause them harm. In this case, the other criminals alone caused them harm by committing crimes against them. Because she is not responsible for their decisions to commit crimes against their victims, she did not cause the victims harm. She merely allowed them to be harmed.

On the other hand, suppose the state raises the crime's probability of punishment for others in order to deter them from committing the crime. Although the state does incur additional enforcement costs to raise the crime's probability of punishment for them, she again did not cause the state to incur such costs. In this case, the state's incurring the additional enforcement costs is caused solely by the dispositions of the other potential criminals. Because she is not responsible for their dispositions to commit crimes, she did not cause the state to incur the additional enforcement costs. She merely allowed the state to incur them.

Because the moderately serious criminal would not cause either harm to occur by not undertaking the EEP's extremely severe punishment, she is not obligated to undertake it on the basis of her obligation not to cause others harm. So the second objection does not show that she deserves the EEP's extremely severe punishment.

C. Third Objection

Critics might concede that by not undertaking the EEP's extremely severe punishment, the moderately serious criminal would not cause others to commit crimes and would not cause the state to raise the crime's probability of punishment for others. But they still might argue that unless she undertakes the EEP's extremely severe punishment, she will cause others harm. By undertaking a less severe punishment, she

would effectively lower the crime's severity of punishment for herself in the sense that she would cause herself to believe that she now stands to suffer a less severe punishment for committing the crime in the future. By effectively lowering the crime's severity of punishment for herself, she would cause others either of two harms.

On the one hand, the state might raise the crime's probability of punishment for her to compensate for her belief that she now stands to suffer a less severe punishment for committing the crime in the future. By doing so, the state would ensure that she is deterred from committing the crime just as strongly as she would have been if she had not effectively lowered the crime's severity of punishment for her. To raise the crime's probability of punishment for her, though, the state must incur additional enforcement costs, such as hiring an additional police officer to monitor her more frequently. So if the state were to raise the crime's probability of punishment for her, she would cause harm to the state by causing it to incur additional enforcement costs.

On the other hand, to avoid incurring additional enforcement costs, the state might not raise the crime's probability of punishment for her. As a consequence, she would commit the crime more than she would have if she had not effectively lowered the crime's severity of punishment for her. So if the state were not to raise the crime's probability of punishment for her, she would cause harm to the victims of her additional crimes.

Assuming she would cause either harm to others unless she undertakes the EEP's extremely severe punishment, she is obligated to undertake it on the basis of her obligation not to cause others harm. Thus, she deserves the punishment.

In response, we might concede that the moderately serious criminal would cause either harm to others by not undertaking the EEP's extremely severe punishment. But even so, the third objection does not plausibly explain why she would be obligated to undertake it. The obligation not to cause others harm has a limit. Under the limit, a person is obligated to undertake the means necessary to prevent herself from causing others harm only if the means are not extremely worse for her than the harm she would

otherwise cause them. Given the limit, the moderately serious criminal would be obligated to undertake the EEP's extremely severe punishment only if the punishment would not be extremely worse for her than the harm she would otherwise cause others. But presumably, the punishment would be extremely worse for her than this harm. To see why, suppose she undertakes only a moderately severe punishment.

On the one hand, the state might not raise the crime's probability of punishment for her because the additional enforcement costs required to do so would be worse than the additional crimes she would otherwise commit. As a consequence, she would commit the crime more often in the future. But because the crime is only moderately serious, presumably her being punished extremely severely for it would be extremely worse for her than the harm she would otherwise cause the victims of her additional crimes.

On the other hand, the state might raise the crime's probability of punishment for her because the additional crimes she would otherwise commit would be worse than the additional enforcement costs the state must incur to raise it for her. As a consequence, she would cause the state to incur additional enforcement costs, such as hiring an additional police officer to monitor her more frequently. But such costs would likely be relatively low: presumably, her being punished extremely severely would be extremely worse for her than those costs.

Thus, the EEP's extremely severe punishment would presumably be extremely worse for the moderately serious criminal than the harm she would otherwise cause others by undertaking a less severe punishment. So the third objection does not plausibly show that she is obligated to undertake it or, therefore, that she deserves it.[31]

31. We might worry that the limit on the obligation not to cause others harm would similarly undermine RS's explanation of why a moderately serious criminal deserves a moderately severe punishment. But presumably, the limit would not undermine it because the moderately severe burden she must undertake to restore the conditions of trust would not be extremely worse for her than the significant costs of insecurity that she would otherwise cause others to incur.

VI. Concluding Remarks

Our analysis of the EEP illustrates the importance of the desert requirement in constraining the types of enforcement policies that the state may adopt for crimes, particularly moderately and mildly serious ones. An enforcement policy is justified only if under it, the state would not punish anyone more than she deserves. So as a consequence of RS, an enforcement policy is justified only if under it, the state would not punish anyone more severely than the burdens she must undertake to fulfill her obligation of restoration. Without this constraint, there is no robust reason for the state not to adopt the kinds of extreme enforcement policies that would be the most efficient means to achieving the optimal level of deterrence for any crime. With this constraint, the state must adopt more costly alternatives. Rather than cause for complaint, the additional costs are simply the costs of respecting the rights of its citizens.

VII. Appendix

Given the main differences between the distributions of punishment under the EEP and the MEP, egalitarians might argue that the EEP's inequality would be all things considered worse than the MEP's inequality among the targeted criminals because on balance the different aspects of inequality would favor the latter over the former.[32] Egalitarians contend that inequality is a complex notion: it has different aspects consisting in the different considerations that could make the inequality in one distribution worse than the inequality in another. The inequality in one distribution might be worse than the inequality in another with respect to one aspect but not with respect to another. To determine whether the inequality in one distribution is all things considered worse than the inequality in another, we must determine the degree to which each distribution realizes the different aspects of inequality, and we must determine the relative importance of each aspect if some aspects favor one distribution, while other aspects

32. I follow Larry Temkin in my terminology and general approach to comparing inequalities. *See* Larry S. Temkin, *Inequality*, 15 PHIL. & PUB. AFF. 99 (1986); LARRY S. TEMKIN, INEQUALITY (1986).

favor the other.

Egalitarians might argue that the EEP's inequality would be all things considered worse than the MEP's inequality because several aspects of inequality would favor the latter over the former, whereas no aspect of inequality would favor the former over the latter. To demonstrate, consider different aspects of inequality that consist in different views about a) which criminals have individual complaints against an enforcement policy's inequality, b) the size of those individual complaints, and c) how to compare one policy's inequality with the inequality of another on the basis of those individual complaints.

First, egalitarians might endorse a "relative to the best off" view of individual complaints against an enforcement policy's inequality: for any enforcement policy E, a criminal has an individual complaint against E's inequality just in case she receives a punishment under E that is more severe than the punishment any other criminal receives under E.[33] According to the "relative to the best off" view of individual complaints, for any enforcement policy E, the size of a criminal's individual complaint against E's inequality is the difference between the severity of her punishment and the severity of the least severe punishment that any other criminal receives under E. Egalitarians also might endorse a maximin view of comparing one policy's inequality with the inequality of another: for any two enforcement policies E1 and E2, E1's inequality is worse than E2's inequality insofar as some criminal's individual complaint against E1's inequality is larger than any criminal's individual complaint against E2's inequality. Call the "relative to the best off" view of individual complaints combined with a maximin view of comparison 'View 1.'

According to View 1, the EEP's inequality would be worse than the MEP's

33. In clarification, when I speak of an enforcement policy, I mean an enforcement policy for one particular crime of one degree of seriousness. Only like criminals would be punished under a particular enforcement policy.

inequality. The difference between the most severe and least severe punishments that criminals would receive under the EEP would be larger than the difference under the MEP. Under either policy, the least severe punishment that some criminals would receive would be no punishment at all. However, under the EEP, the most severe punishment that other criminals would receive would be extremely severe, whereas under the MEP, the most severe punishment that other criminals would receive would be only moderately severe. On the "relative to the best off" view of individual complaints, some criminal's individual complaint against the EEP would be larger than any criminal's individual complaint against the MEP. So given its maximin view of comparison, the EEP's inequality would be worse than the MEP's inequality according to View 1.[34]

Second, egalitarians might endorse a "relative to all those better off" view of individual complaints: for any enforcement policy E, a criminal has an individual complaint against E's inequality just in case she receives a punishment under E that is more severe than the punishment any other criminal receives under E. According to the "relative to all those better off" view of individual complaints, for any enforcement policy E, the size of a criminal's individual complaint against E's inequality is the sum of the differences between the severity of her punishment and the severity of each less severe punishment that other criminals receive under E. Call the "relative to all those better off" view of individual complaints combined with a maximin view of comparison 'View 2.'

34. To illustrate the application of View 1, consider the following numerical example. Suppose a state would optimally deter people from committing a moderately serious crime by making its expected punishment 100 units of severity. Assuming the crime's expected punishment were 100 units of severity, 150 criminals would still commit it: such criminals either cannot be deterred or the means required to deter them would be much worse than the crimes they would otherwise commit. Suppose the EEP and the MEP both make the crime's expected punishment 100 units of severity. However, under the EEP, the crime's probability of punishment would be 0.04, and its severity of punishment would be 2500 units of severity. Whereas under the MEP, its probability of punishment would be 0.4, and its severity of punishment would be 250 units of severity. Under the EEP, six of the 150 undeterred criminals would receive a punishment of 2500 units of severity, and the other 144 would receive no punishment at all. Under the MEP, 60 of the 150 undeterred criminals would receive a punishment of 250 units of severity, and the other 90 would receive no punishment at all. On the "relative to the best off" view of individual complaints, the largest individual complaint of a criminal against the EEP would be 2500-0, which is 2500, whereas the largest individual complaint of a criminal against the MEP would be 250-0, which is 250.

According to View 2, the EEP's inequality would also be worse than the MEP's inequality. The sum of the differences between the most severe punishment that a criminal would receive under the EEP and each less severe punishment that other criminals would receive under the EEP would be larger than the sum of such differences under the MEP. The most severe punishment that a criminal would receive under the EEP would be more severe than the most severe punishment that a criminal would receive under the MEP, and the less severe punishments that other criminals would receive under either policy would be no punishment at all. Furthermore, relative to the MEP, more criminals would receive no punishment at all under the EEP because the crime's probability of punishment would be lower under the EEP. Hence, on the "relative to all those better off" view of individual complaints, some criminal's individual complaint against the EEP would be larger than any criminal's individual complaint against the MEP. So given its maximin view of comparison, the EEP's inequality would be worse than the MEP's inequality according to View 2.[35]

Third, egalitarians might endorse a "relative to the average" view of individual complaints: for any enforcement policy E, a criminal has an individual complaint against E's inequality just in case she receives a punishment under E that is more severe than the severity of the average punishment that criminals receive under E. According to the "relative to the average" view of individual complaints, for any enforcement policy E, the size of a criminal's individual complaint against E's inequality is the difference between the severity of her punishment and the severity of the average punishment that criminals receive under E. To determine the average punishment that criminals receive under E, divide the total amount of punishment that all criminals receive under E by the total number of criminals under E. The average punishment that criminals receive under E

35. To illustrate, apply View 2 to our numerical example. On the "relative to all those better off" view of individual complaints, the largest individual complaint of a criminal against the EEP would be 144*(2500-0), which would be 360,000, whereas the largest individual complaint of a criminal against the MEP would be 90*(250-0), which would be 22,500.

will equal the crime's expected punishment under E. Call the "relative to the average" view of individual complaints combined with a maximin view of comparison 'View 3.'

According to View 3, the EEP's inequality would also be worse than the MEP's inequality. The difference between the most severe punishment that a criminal would receive under the EEP and the crime's expected punishment under the EEP would be larger than the difference under the MEP. By hypothesis, the crime's expected punishment would be the same under both the MEP and the EEP. And the most severe punishment that a criminal would receive under the EEP would be more severe than the most severe punishment that a criminal would receive under the MEP. So on the "relative to the average" view of individual complaints, some criminal's individual complaint against the EEP would be larger than any criminal's individual complaint against the MEP. Given its maximin view of comparison, the EEP's inequality would be worse than the MEP's inequality according to View 3.[36]

Fourth, egalitarians might endorse an additive view of comparing one enforcement policy's inequality with the inequality of another: for any two enforcement policies E1 and E2, E1's inequality is worse than E2's inequality insofar as the sum of each criminal's individual complaint against E1's inequality is larger than the sum of each criminal's individual complaint against E2's inequality. Call the "relative to the average" view of individual complaints combined with an additive view of comparison 'View 4.' Call the "relative to all those better off" view of individual complaints combined with an additive view of comparison 'View 5.'

According to Views 4 and 5, the EEP's inequality would be worse than the MEP's inequality. On either the "relative to the average" or the "relative to all those better off" view of individual complaints, more criminals would have individual complaints against

36. To illustrate, apply View 3 to our numerical example. On the "relative to the average" view of individual complaints, the largest individual complaint of a criminal against the EEP would be 2500-100, which would be 2400, whereas the largest individual complaint of a criminal against the MEP would be 250-100, which would be 150.

the MEP's inequality than against the EEP's inequality; however, the size of each criminal's individual complaint against the EEP's inequality would be disproportionately larger than the size of each criminal's individual complaint against the MEP's inequality. As a consequence, on either the "relative to the average" or the "relative to all those better off" view of individual complaints, the sum of each criminal's individual complaint against the EEP would be larger than the sum of each criminal's individual complaint against the MEP. Given their additive view of comparison, the EEP's inequality would be worse than the MEP's inequality according to Views 4 and 5.[37]

On the basis of Views 1-5, egalitarians would infer that the EEP's inequality would be all things considered worse than the MEP's inequality. Our application of the views shows that several aspects of inequality would favor the MEP's inequality over the EEP's inequality.[38] And no aspect of inequality seems to favor the EEP's inequality over the MEP's inequality. The one salient aspect of inequality we have yet to consider is a "relative to the best off" view of individual complaints combined with an additive view of comparison. Call it 'View 6.' View 6 does not favor one over the other. Although more criminals would have individual complaints against the MEP's inequality, the individual complaints of criminals against the EEP's inequality would be proportionately larger, such

37. To illustrate Views 4 and 5, apply them to our numerical example. On the "relative to the average" view of individual complaints, six criminals would have individual complaints against the EEP, and each complaint would be 2400. So the sum of their complaints against the EEP would 14,400. Whereas 60 criminals would have individual complaints against the MEP, and each complaint would be 150. So the sum of their complaints against the MEP would be 9000.

On the "relative to all those better off" view of individual complaints, six criminals would have individual complaints against the EEP, and each complaint would be 360,000. So the sum of their complaints against the EEP would be 2,160,000. Whereas 60 criminals would have individual complaints against the MEP, and each complaint would be 22,500. So the sum of their complaints against the MEP would be 1,350,000.

38. In addition to Views 1-5, weighted variants of them would also favor the MEP's inequality over the EEP's inequality. Such variants would weight the individual complaints of criminals who receive more severe punishments disproportionately higher than the individual complaints of criminals who receive less severe punishments. Because the non-weighted versions of Views 1-5 favor the MEP's inequality over the EEP's inequality, the weighted variants of them would favor the MEP's inequality even more over the EEP's inequality since criminals would be punished more severely under the EEP.

that the sum of the individual complaints of criminals against the EEP's inequality would be the same as the sum of the individual complaints of criminals against the MEP's inequality.[39]

39. To illustrate, apply View 6 to our numerical example. On the "relative to the best off" view of individual complaints, six criminals would have individual complaints against the EEP, and each complaint would be 2500. So the sum of their complaints against the EEP's inequality would be 15,000. Whereas 60 criminals would have individual complaints against the MEP, and each complaint would be 250. So the sum of their complaints against the MEP's inequality would also be 15,000.

Chapter 3

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

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 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 she 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.[3] This presumption is rebutted,

---

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. 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

though, when the criminal actor has an exculpatory defense.[4] Defenses mitigate how much a criminal actor is blameworthy and liable to be punished.[5] 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 seek to spell out a taxonomy of the main types of defenses. Second, I aim to 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.

---

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).

4. 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.

5. 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, like 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).

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 vulnerable to her to a significant degree. 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 1-3 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 1-3 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.

Fourth, everyone knows all the external and internal facts about all the acts the criminal actor has performed, including the external and psychological causes of her acts and their consequences. 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 4-5. So if everyone's knowledge of the relevant facts justify 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 4-6 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 interest in them but are unaware of them. Presuppositions 4-6 entail that others fulfill such duties.

Also I make presuppositions 4-7 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 4-7 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 4-7 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 4-7 were to obtain. 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.[6] Similarly, these less idealized conditions also seem irrelevant to the issue of whether 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 4-7 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

6. *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, like blame, toward something depends on what she has evidence for believing about it).

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.[7] So in this sense, the fact that someone believes that her performing a particular act would directly cause another to die 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.[8] 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

---

7. 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 PHILOSOPHY: FROM ACTION TO VALUES 17, 17 (Dan Egonsson, Jonas Josefsson, Bjorn Petersson, & Toni Ronnow-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.

8. *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 (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).

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.[9] 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.

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.[10] 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

---

9. 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.

10. *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.

that make it a criminal act.[11] 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 caused him to die, I did not believe that doing it would cause him to die or harm him 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.[12] 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

---

11. *See* ARISTOTLE, *supra* note 10, 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 7. For discussion on this issue, see VICTOR TADROS, CRIMINAL RESPONSIBILITY 237-64 (2005).

12. *See* ARISTOTLE, *supra* note 10, 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 provide the criminal actor with a defense. *See supra* notes 7, 11.

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.[13] 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 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.[14] 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

---

13. 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.

14. *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 11, at 273-80 (same); Marcia Baron, *Justifications and Excuses*, 2 OHIO ST. J. CRIM. L. 387, 392-93 (2005) (same).

believed they obtained, then she would not have been motivated as strongly to perform it.[15]

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 killed 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 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

15. 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).

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.[16] 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 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

16. 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.

actual consequences might have been at least as good as the consequences of any other acts then available to her.[17] 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 killing 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 others under conditions in which doing so is necessary to prevent them from killing many others. Hence, expressing any blame toward the misanthrope for killing the one 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

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

causal consequences, but also the value of the act itself and the motives with which it is performed.[18] 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

18. *See* G.E. MOORE, PRINCIPIA ETHICA 198-99 (Thomas Baldwin ed., rev. ed. 1993); William Shaw, *The Consequentialist Perspective*, *in* CONTEMPORARY DEBATES IN MORAL THEORY 5, 6-7 (James Dreier ed., 2006); TADROS, *supra* note 11, at 278-79.

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.[19] And others' blaming her for the act would consist in their feeling an attitude of moral blame toward her for the act. Such an attitude might consist in resentment or indignation.[20] 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.[21] 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

19. The idealized conditions plausibly include the satisfaction of my presuppositions 4-7.

20. *See, e.g.*, STRAWSON, *supra* note 8, at 13-15.

21. *See* SCANLON, *supra* note 8, at 269.

toward her.[22] 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 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

22. *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").

attitude is warranted depends on a distinct and narrower set of considerations.[23] 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.[24] Thus, the consequences of having or not 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

---

23. 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 6; D'Arms & Jacobson, *supra* note 22; 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 7; Wlodek Rabinowicz & Toni Ronnow-Rasmussen, *The Strike of the Demon: On Fitting Pro-Attitudes and Value*, 114 ETHICS 397 (2004).

24. 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 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 8, 13-15 (discussing the types of reactive attitudes and their relations to each other).

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.[25] 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 or in undertaking a process of getting myself to feel it. But even if the indignation would be optimific, 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

25. *See* GIBBARD, *supra* note 23, 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 7, at 27 (same).

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 performed intentionally.[26] 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 that she has a justification for the intention she had to perform it.[27] 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

26. Strictly speaking, all acts are performed intentionally. Because I assume criminal acts can be performed unintentionally, I take criminal acts to be, strictly speaking, merely putative acts. *See supra* note 10.

27. 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.

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 indefeasible type 1 defense for it. She has an indefeasible 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) indefeasible type 1 defense for it. And she has a full (or partial) indefeasible 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.

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.[28] 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 indefeasible 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 indefeasible type 1 defense of consent for this intention. Furthermore, no third parties could have an objective justification for intending to strike

28. *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 L. & PHIL. 18 (1999) (same).

my opponent on my behalf because no third parties could have an indefeasible type 1 defense for this intention. In particular, no third parties could have an indefeasible 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.

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.[29] 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

29. 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.

equal.[30]

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.[31] 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 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 unliked, 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

30. 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 id*.

31. *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 28, at 489 (same).

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.[32] Similarly, the concept of a subjective justification is conceptually prior to the concept of an objective 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

32. To illustrate, consider again cases of compulsion.

justification for it is indefeasible. Thus, the fact 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.[33]

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.[34] A type 2 defense blocks this

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

34. 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 a motivational defect in her character. *See, e.g.,* Michael D. Bayles, *Character, Purpose, and Criminal Responsibility*, 1 L. & 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

inference primarily by blocking the 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.[35] 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.

Many different considerations can provide a criminal actor with a type 2 defense.

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 11, at 293-321; George Vuoso, *Background, Responsibility, and Excuse*, 96 YALE L.J. 1661 (1987); *cf.* SCANLON, *supra* note 8, 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 8, 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 8, at 123 (describing excuses that, according to Strawson, "present the other ... as acting uncharacteristically due to extraordinary circumstances").

35. 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.

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 disposed to commit crimes.[36] 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, 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,[37] which consists in all the norms she accepts regarding what to intend in particular situations.[38] 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.

---

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

37. *See* FLETCHER, *supra* note 28, at 802 (stating that an "excuse represent[s] a limited, temporal distortion of the actor's character").

38. 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.

However, she has a type 2 defense for her crime 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.[39] 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,

39. *Cf.* HORDER, *supra* note 1, at 139-90 (endorsing an excuse of diminished capacity).

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 in me a drug that makes me intoxicated. While intoxicated, 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.[40]

In the cases at issue, the selves might also be the same person even though they

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

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.[41] 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.[42] 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.[43] 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

41. *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).

42. *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).

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

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 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 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.[44] If a criminal actor has a full type 3 defense, she lacks the capacity to

44. *Cf.* SCANLON, *supra* note 8, at 280 (describing defenses that consist in an incapacity to respond appropriately to reasons); STRAWSON, *supra* note 8, at 8-10 (same); Watson, *supra* note 8, at 123 (same);

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 non-moral. 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.[45] So she might 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.[46] 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

MODEL PENAL CODE art. 4 (same).

45. 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.

46. 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.

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.[47] 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 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.[48] So if the criminal actor lacks the relevant

47. 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.

48. If a criminal actor performed her act due to a mere temporary incapacity at the time of the act, the

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.[49] 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 appropriately to them even if she has the capacity to.[50] Similarly, type 3 defenses

---

temporary incapacity would constitute a type 2 but not a type 3 defense.

49. *See* R.A. DUFF, TRIALS AND PUNISHMENTS 14-38 (1986) (emphasizing this possibility).

50. 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-

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.[51] 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 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

---

95 (David Widerker & Michael McKenna eds., 2003) (arguing that determinism is consistent with an actor's ability to do otherwise).

51. 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.

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.[52] 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 which 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 sub-types.

We have already seen that justifications are a subset of type 1 defenses whose salience warrants a separate exposition. Other salient sub-types include exemptions, failure of proof defenses, and excuses. An exemption is an enduring incapacity to respond appropriately to any moral reasons.[53] Under an exemption, a person would not

52. 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 8, at 131-37.

53. *See* TADROS, *supra* note 11, at 124-29.

be blameworthy or liable to punishment for any acts she might perform, including criminal acts. Exemptions are a subset of type 3 defenses.

Consider now failure of proof defenses.[54] In general, the state is justified in punishing someone for committing a crime only if a statute explicitly prohibits the crime.[55] 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.[56] 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.[57] 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, like type 3 ones.[58] 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.

---

54. *See* JOSHUA DRESSLER, UNDERSTANDING CRIMINAL LAW 181-82 (1995); Robinson, *supra* note 5, at 204-08.

55. *See* MODEL PENAL CODE sec. 1.05(1).

56. *See* DRESSLER, *supra* note 54, at 51-61; Robinson, *supra* note 5, at 250-64.

57. *See* MODEL PENAL CODE sec. 1.12(1).

58. *See* MODEL PENAL CODE secs. 1.12(2)-(4).

The concept of an excuse is open to broader and narrower conceptions.[59] 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, like justifications, exemptions, and failure of proof defenses.

Whether a defense falls under the extension of one of these more particular sub-types 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 sub-types of defenses, like justifications, exemptions, failure of proof defenses, and excuses.

IV. Two Requirements of Justified Punishment

My general theory of the justification of punishment specifies the general conditions under which the state is 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 has most reason to punish someone against her will. So the theory 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. As we discussed in Chapter 1, my general theory contains a desert requirement and a value requirement. The state is justified in imposing a punishment on someone against her will only if the punishment would satisfy both requirements. According to the desert requirement, the person must deserve the punishment. According to the value requirement, the expected value of the consequences of imposing the punishment on the person must be at least as

59. *See* Marcia Baron, *Excuses, excuses*, 1 CRIM. L. & PHIL. 21, 37 (2007) (crediting Antony Duff with pointing out broader and narrower senses of 'excuse').

high as the expected value of the consequences of any other available act that would not violate anyone's rights.

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.[60] 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 whom 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.[61] So punishing them would have no deterrence benefit. Therefore, the punitive costs would outweigh the benefits of punishing them.

60. 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.

61. *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.[62] 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

62. *See* H.L.A. HART, *Prolegomenon to the Principles of Punishment*, *in* PUNISHMENT AND RESPONSIBILITY: ESSAYS IN THE PHILOSOPHY OF LAW 1, 19 (1968).

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.[63] But if the state were to 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.[64] Such a punishment

63. *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 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.

64. 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*

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, 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.[65] 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

---

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).

65. *Cf.* BENTHAM, *supra* note 61, 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").

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 optimific. They are optimific because 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.[66] 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

66. *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).

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 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.[67]

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

67. *Cf.* PARFIT, *supra* note 42, at 20-23, 457-61 (explaining why it would probably be rational for a person to be transparently disposed to ignore threats).

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 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.[68] 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.[69]

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 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.[70]

---

68. 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 64, at 348, 352-53; Abbott v. The Queen, 1977 A.C. 755, (Lord Salmon); *cf.* PARFIT, *supra* note 42, at 20-23, 457-61.

69. 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.

70. H.L.A. Hart emphasizes this point in *supra* note 62, at 22-24 and in *Legal Responsibility and Excuses*, *in* PUNISHMENT AND RESPONSIBILITY: ESSAYS IN THE PHILOSOPHY OF LAW 28 (1968).

Hence, by punishing them, the state would restrict people's ability to plan on making choices that would safeguard them from being punished.[71] 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 them under duress. Second, they would avoid more valuable activities that would leave them too vulnerable to duress.

Both Hart and T.M. Scanlon emphasize the importance of providing people with a fair opportunity to avoid being punished. *See* HART, *supra*; SCANLON, *supra* note 8, at 263-67.

71. Both Hart and Scanlon emphasize the value of this ability. See their works cited in *supra* note 70.

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.[72] 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

72. *See* Sanford H. Kadish, *Excusing Crime*, 75 CAL. L. REV. 257, 263-64 (1987).

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 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.[73] 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.[74]

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.[75] On this sense, something is good only if everyone is warranted for agent-neutral reasons in desiring that it obtain.[76] Conversely,

---

73. *See* THOMAS NAGEL, THE VIEW FROM NOWHERE, 152-53 (1986) (distinguishing agent-relative from agent-neutral reasons).

74. *See* ROBERT NOZICK, ANARCHY, STATE, AND UTOPIA 30-33 (1974) (describing rights as side constraints).

75. 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).

76. 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.

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 justification. Without providing one, the argument is problematic.

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 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. There is no robust reason to believe one way or the other. 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 intervener."[77] 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.

---

77. 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 50, 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).

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 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.[78] 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

78. *See* SCANLON, *supra* note 8, at 268-69.

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 intervener. 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 in 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 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 in me the drug. 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 in me the drug. 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.

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 interveners or non-standard 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, like 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.

D. An Expressive Theory

An expressive theory might point out that punishing someone expresses an attitude of moral blame toward her.[79] Such an attitude might be resentment or indignation. Given the expressive aspect of punishment, an expressive 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

79. *See* JOEL FEINBERG, *The Expressive Function of Punishment*, *in* DOING AND DESERVING: ESSAYS IN THE THEORY OF RESPONSIBILITY 98 (1970).

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.[80] 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.[81] Punishing someone expresses

80. *See* DARWALL, *supra* note 23, at 17; STRAWSON, *supra* note 8, at 14-15, 21-22; Watson, *supra* note 8, at 121, 126-28.

81. 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 and the judgment that their objects can be properly

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.[82] As a consequence, the expressive theory is circular in virtue of the fact that it presupposes what it aims to explain.[83]

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

---

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).

82. 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 50, at 271-75; DARWALL, *supra* note 23, 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 23.

83. *Cf.* Baldwin, *supra* note 81, 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).

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 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.

E. A Restorative Signaling Theory

As I argued in Chapter 1, RS explains why and how much criminals with no defenses deserve to be punished by explaining why they must undertake a punishment to fulfill the obligation of restoration they incur from committing their crimes. According to RS, when someone commits a crime without a defense, she undermines certain conditions of trust in the sense that she undermines the conditions that are necessary for others' being justified in believing that she is not disposed to commit crimes. She is obligated to restore those conditions because unless she restores them, she will cause others to incur the costs of insecurity.

To restore the conditions of trust, she must demonstrate to others that she has a good will in the sense of a stable disposition to be appropriately motivated by the moral reasons against violating the rights of others. To demonstrate that she has a good will, she must demonstrate that she has a stable disposition to care highly about the interests of others. To demonstrate this, she must demonstrate that she has acted with a sufficiently high degree of benevolence for a sufficiently long time after committing her crime. 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.

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.[84]

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.

---

84. 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 it. 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 it.

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.[85] 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.

---

85. 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 73, at 171-75; SCHEFFLER, *supra* note 75, 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.

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 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.[86] 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 self-governance.[87] 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

---

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

87. *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).

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.[88] And others cannot fairly demand someone to do something that she lacks the capacity to do.[89] So a person is obligated to do something only if she has the capacity to do it.[90] 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

88. *Cf.* Copp, *supra* note 50, at 271-75; DARWALL, *supra* note 23, at 96-99.

89. *See* Copp, *supra* note 50, at 271-75.

90. 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").

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.[91] 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 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.[92]

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

---

91. 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. As we discussed in Chapter 1, criminals tend to exhibit a high degree of versatility.

92. 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.

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 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 necessary 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.[93] 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.[94] 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.[95] In other cases, though, childhood arguably provides a type 4 defense but not one of 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

93. *See* SCANLON, *supra* note 8, at 280.

94. *See id.* at 280-81.

95. *See id.* at 280.

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 non-punitive 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.[96] 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.[97] 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 have the capacity to respond appropriately to the moral

---

96. 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).

97. 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).

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 deprogramming or deconditioning, they are highly likely to reject the changes that were induced in them from the brainwashing.[98] 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.[99] Hence, the fact that a criminal was brainwashed can mitigate the severity of burdens she must undertake to fulfill her obligation of restoration.[100]

Consider a rotten social background.[101] 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.[102]

---

98. *See* Delgado, *supra* note 96, 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).

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

100. 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.

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

102. *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 L. & 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

Suppose 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.[103] 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.[104]

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

---

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).

103. 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.

104. 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.

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.

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, like 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.

Chapter 4

Forgiving Criminals: What It Means and When It is Warranted

I. Introduction

Suppose someone commits a crime by disrespecting the rights of others in performing a criminal act. She has no exculpatory defenses. As a consequence, she is blameworthy and deserves to be punished for the crime. We also standardly presume she could undertake some course of action that would warrant forgiving her for the crime. Our presumption generates two issues. One concerns what it means to forgive her. Another concerns the conditions under which forgiveness would be warranted. To appreciate these issues, we should distinguish forgiveness from other positive responses we might have to a criminal.

Forgiveness is distinct from mercy, which consists in not forcing someone to undertake the burdens she is obligated to undertake.[1] When someone commits a crime with no exculpatory defenses, she incurs an obligation to undertake a punishment for the crime.[2] If she caused harm to others, she also might incur an obligation to compensate them for the harm, and fulfilling this obligation might be burdensome for her. Having mercy on the criminal could consist in not forcing her to fulfill either obligation or to undertake the burdens necessary to fulfill them. As such, having mercy on her is consistent with continuing to blame her for the crime by continuing to feel an attitude of

1. *See, e.g.*, JEFFRIE G. MURPHY, *Forgiveness and resentment, in* FORGIVENESS AND MERCY 14, 20-21 (Jeffrie G. Murphy & Jean Hampton eds., 1988) (distinguishing mercy from forgiveness); Lucy Allias, *Wiping the Slate Clean: The Heart of Forgiveness*, 36 PHIL. & PUB. AFF. 36, 47-49 (2008) (same).

2. Throughout the paper, I refer only to criminals who commit their crimes with no full exculpatory defenses.

moral blame toward her.[3] Forgiveness, though, involves the suspension of blame.[4]

Forgiveness is not the mere suspension of blame.[5] People can suspend blame for no reason at all. For example, they can suspend blaming a criminal merely because they forgot about her crime. Or they can suspend blame merely because they lose the capacity to feel blame. Neither is forgiveness because forgiveness involves suspending blame for a reason.[6] It involves the suspension of blame in response to a judgment. In this sense, forgiveness not only is a judgment-sensitive attitude or state of mind, but consists at least partly in a judgment.[7]

Forgiveness, however, is not the mere suspension of blame for a reason.[8] People might suspend blame merely for reasons that bear only on whether they would be warranted in desiring to blame or in undertaking a process of getting themselves to blame.[9] Call these 'state-based reasons'.[10] For example, people might suspend blame for

---

3. Throughout the paper, I refer only to blame of a moral kind. An attitude of moral blame might consist in resentment, indignation, or guilt. *See*, *e.g.*, P.F. STRAWSON, *Freedom and Resentment*, *in* FREEDOM AND RESENTMENT AND OTHER ESSAYS 1, 13-15 (1974).

4. *See*, *e.g.*, MURPHY, *supra* note 1, at 22 (claiming that forgiveness involves forswearing resentment); Allias, *supra* note 1, at 41 (contending that forgiveness involves overcoming retributive emotions).

5. *See*, *e.g.*, MURPHY, *supra* note 1, at 22-23.

6. *See*, *e.g.*, MURPHY, *supra* note 1, at 23-24 (claiming that forgiveness involves forswearing resentment for moral reasons).

7. *See* Pamela Hieronymi, *Articulating an Uncompromising Forgiveness*, 62 PHIL. & PHENOMENOLOGICAL RES. 529, 530 (2001) (taking forgiveness to involve a judgment or change in view). For the idea of a judgment-sensitive attitude, see T.M. SCANLON, WHAT WE OWE TO EACH OTHER 20-24 (1998).

8. *See*, *e.g.*, JEAN HAMPTON, *Forgiveness, resentment and hatred*, *in* FORGIVENESS AND MERCY 35, 36-37 (Jeffrie G. Murphy & Jean Hampton eds., 1988).

9. *See* ALLAN GIBBARD, WISE CHOICES, APT FEELINGS 37 (1990) (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); Derek Parfit, *Rationality and Reasons*, *in* EXPLORING PRACTICAL PHILOSOPHY: FROM ACTION TO VALUES 17, 27 (Dan Egonsson, Jonas Josefsson, Bjorn Petersson, & Toni Ronnow-Rasmussen eds., 2001) (same).

10. I borrow this term from Derek Parfit. *See* John Broome, *Reason and Motivation*, 71 SUPPLEMENT TO THE PROCEEDINGS OF THE ARISTOTELIAN SOCIETY 137-38 (2003) (discussing Parfit's use the term).

mere prudential or altruistic reasons concerning its effects. Blaming can have harmful effects on the well-being of both those who blame and those who are blamed. When someone feels blame toward another, the feeling itself can impair her capacity to engage in valuable activities or relationships with others generally. When someone learns that others blame her, she can suffer a loss in self-esteem. So people might suspend blame merely to avoid the harmful effects of blaming itself. In this case, they would suspend blame for a reason, namely a state-based reason. But their suspension would not constitute forgiveness, which involves the suspension of blame for the right kind of reasons. The right kind concern only considerations that bear on whether others would be warranted in blaming.[11] Call these considerations 'object-based reasons'.[12] Others forgive someone only if they suspend blaming her in response to their judgment that certain considerations obtain which would make their continuing to blame her unwarranted.

Forgiveness, though, is not the mere suspension of blame for object-based reasons. Suppose someone commits a crime, and others blame her for it. However, they blame her too much; they feel too much resentment or indignation.[13] So their attitude of blame was never warranted. Now assume they reflect on the nature of her crime in a "cool hour" and realize their mistake. In response, they mitigate how much they blame her for the crime and subsequently blame her only to a warranted degree. In other words,

---

11. For discussion on the problem of identifying the considerations that are relevant to whether an attitude of blame is warranted or, in other words, fitting or rational, see, e.g., Justin D'Arms & Daniel Jacobson, *The Moralistic Fallacy: On the 'Appropriateness' of Emotions*, 61 PHIL. & PHENOMENOLOGICAL RES. 65, 77 (2000); Justin D'Arms & Daniel Jacobson, *Sentiment and Value*, 110 ETHICS 722, 745 (2000); STEPHEN DARWALL, THE SECOND-PERSON STANDPOINT: MORALITY, RESPECT, AND ACCOUNTABILITY 15-17 (2006); GIBBARD, *supra* note 9, at 36-40; Parfit, *supra* note 9, at 17-39; Wlodek Rabinowicz & Toni Ronnow-Rasmussen, *The Strike of the Demon: On Fitting Pro-Attitudes and Value*, 114 ETHICS 397, 397-423 (2004).

12. I borrow this term from Derek Parfit. *See* Broome, *supra* note 10, at 137-38 (discussing Parfit's use of it).

13. *See* JOSEPH BUTLER, *Sermon IX: Upon Forgiveness of Injuries*, *in* THE WORKS OF JOSEPH BUTLER 150, 151-52 (W.E. Gladstone ed., vol. 2, 1896) (noting this possibility).

they rectify their previously unwarranted attitude of blame toward her. In rectifying the attitude, they suspend for object-based reasons at least a degree of blame toward the criminal. However, the rectification does not involve forgiveness, not even of a partial kind. From the perspective of those who forgive a criminal, the criminal must earn their forgiveness by making their continued blame toward her unwarranted. But if their blame toward her was never warranted in the first place, there is nothing she must do to make its continuation unwarranted. So forgiveness involves the suspension of only a previously warranted attitude of blame. In other words, it involves suspending blame that was warranted prior to the considerations that make forgiveness warranted.[14]

But lastly, forgiveness is not the mere suspension of a previously warranted attitude of blame for object-based reasons. After committing her crime, suppose the criminal subsequently suffers from a mental illness that causes her to lose the capacity to respond appropriately to moral reasons or reasons of any kind. As a consequence, the mental illness provides her with a full exculpatory defense for the crime.[15] In response, others might suspend for this object-based reason their previously warranted attitudes of blame toward her. Their suspension, though, would still not involve forgiving her. Forgiveness involves suspending a previously warranted attitude of blame not just for any object-based reasons, but only for a relevant class of them. The challenge of demarcating the relevant class is the challenge of understanding forgiveness.

---

14. Relatedly, forgiving a criminal for her crime presupposes that she was blameworthy for it. *See* Allias, *supra* note 1, at 43. *But see* JOSEPH BUTLER, *Sermon VIII: Upon Resentment* & *Sermon IX: Upon Forgiveness of Injuries*, *in* THE WORKS OF JOSEPH BUTLER 136-67 (W.E. Gladstone ed., vol. 2, 1896) (contending that forgiveness could consist in forswearing revenge and moderating a previously unwarranted attitude of resentment). For this interpretation of Butler, see CHARLES L. GRISWOLD, FORGIVENESS: A PHILOSOPHICAL EXPLORATION 19-37 (2007).

15. In Chapter 3, I explain why such an incapacity provides an exculpatory defense. *Cf.* SCANLON, *supra* note 7, at 280 (describing defenses that consist in an incapacity to respond appropriately to reasons); STRAWSON, *supra* note 3, at 8-10 (same); Gary Watson, *Responsibility and the Limits of Evil: Variations on a Strawsonian Theme*, *in* PERSPECTIVES ON MORAL RESPONSIBILITY 123 (John Martin Fischer & Mark Ravizza eds., 1993) (same); R.A. DUFF, TRIALS AND PUNISHMENTS, 14-38 (2005) (same); VICTOR TADROS, CRIMINAL RESPONSIBILITY 124-29 (2005) (same); MODEL PENAL CODE art. 4 (same).

In this paper, I defend a theory of what it means and when it is warranted to forgive criminals.[16] In doing so, I demarcate the relevant class of reasons for which suspending blame toward them would constitute forgiveness. Ultimately, I defend the theory by deriving it from a restorative signaling theory of punitive desert, which I call 'RS'. I argue that RS illuminates the content of blame in a way that identifies the relevant class of reasons. Thus, my argument uncovers an important connection between the concepts of punitive desert, blame, and forgiveness. After defending the theory, I apply it to several potentially controversial issues concerning the forgiveness of criminals. The applications shed further light on the nature of forgiveness itself and the conditions under which it would be warranted.

II. A Restorative Signaling Theory of Punitive Desert

As I argued in Chapter 1, RS explains why criminals deserve to be punished by explaining why they must undertake a punishment to fulfill the obligation of restoration they incur from committing their crimes.[17] According to RS, when someone commits a crime, she undermines certain conditions of trust in the sense that she undermines the conditions that are necessary for others' being justified in believing that she is not disposed to commit crimes. She is obligated to restore those conditions because unless she restores them, she will cause others to incur the costs of insecurity.

To restore the conditions of trust, she must demonstrate to others that she has a good will in the sense of a stable disposition to be appropriately motivated by the moral reasons against violating the rights of others. To demonstrate that she has a good will, she must demonstrate that she has a stable disposition to care highly about the interests of others. To demonstrate this, she must demonstrate that she has acted with a sufficiently

16. I leave it an open question how my theory might apply to those who perform acts that are wrongful in some sense but not serious enough to be crimes. I focus only on forgiving criminals for two reasons. First, their forgiveness is especially difficult to warrant. Second, the practical consequences of their warranting forgiveness are especially important.

17. In summarizing RS here, I make the same presuppositions I defended in Chapter 1.

high degree of benevolence for a sufficiently long time after committing her crime. 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.

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.[18]

III. Blaming Criminals

RS illuminates the content of blame. A criminal not only deserves to be punished for her crime. She also is blameworthy for it in the sense that others would be warranted in blaming her for the crime if presuppositions 4-7 were to obtain.[19] On a standard view, blame consists partly in certain demands.[20] For example, it contains a demand to undertake a punishment or, more precisely, the burdens constitutive of a punishment.[21]

---

18. 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 it. 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 it.

19. See Chapter 1 for my exposition of these presuppositions. *Cf.* D'Arms & Jacobson, *Sentiment*, *supra* note 11, at 745 (noting that whether someone is warranted in feeling a particular emotion, like blame, toward something depends on what she has evidence for believing about it).

20. *See* DARWALL, *supra* note 11, at 17; STRAWSON, *supra* note 3, at 14-15, 21-22; Watson, *supra* note 15, at 121, 126-28.

21. 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 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); BUTLER, *Sermon VIII*, *supra* note 14, at 141 (suggesting that attitudes of resentment and indignation toward someone involve a desire that she be punished).

RS illuminates the content of blame by identifying some of the presuppositions and demands constitutive of it. According to RS, blame contains a demand to restore certain conditions of trust by undertaking certain burdens in order to demonstrate a good will. When others blame someone for a crime, they presuppose that she undermined certain conditions of trust by committing it, and they demand her to restore those conditions by undertaking certain burdens in order to demonstrate that she has a good will.[22] The higher the severity of burdens they demand her to undertake, the more they blame for the act. The lower the severity of burdens, the less they blame her for the act.

By illuminating its content, RS illuminates the conditions under which blame would be warranted. Suppose people blame a criminal for her crime. In doing so, they presuppose that she undermined certain conditions of trust by committing it, and they demand her to restore those conditions by undertaking certain burdens. Thus, their attitude is warranted only if a) they are justified in believing that the criminal really did undermine the relevant conditions of trust by committing her crime, and b) they are warranted in demanding her to undertake the relevant burdens to restore those conditions. Their demand is warranted only if they are justified in believing that the criminal really must undertake those burdens to restore the relevant conditions.[23]

IV. Forgiving Criminals

A. The Meaning of Forgiveness

Given the content of blame and the conditions under which it is warranted, we can identify what it means to forgive a criminal. Suppose others feel a warranted attitude of blame toward her for the crime. They are justified in believing that she undermined

22. When I speak of a demand to restore conditions of trust, I mean a demand to restore them by demonstrating a good will. When I speak of a criminal's restoring conditions of trust, I mean her restoring them in this way.

23. More generally, RS explains why the demand to restore conditions of trust is warranted insofar as the criminal will cause others to incur the costs of insecurity unless she restores them, and people are generally warranted in demanding someone not to cause others harm, like the costs of insecurity.

certain conditions of trust by committing it, and they are warranted in demanding her to restore those conditions by undertaking certain burdens. To forgive the criminal for her crime, they must suspend blaming her for it, and they must do so for the right reason. The right reason consists in their judging that the criminal has restored the conditions of trust she undermined by committing the crime.[24] In response to the judgment, they would cease demanding the criminal to restore those conditions, and so would cease demanding her to undertake any burdens as a means to restoring them. Thus, forgiving a criminal involves suspending a previously warranted attitude of blame toward her by suspending its constitutive demands in response to the right judgment.

B. Degrees of Forgiveness

Forgiveness comes in degrees. It can be full or partial. Before a criminal commits her crime, others are justified in believing with a particular baseline credence that she is not disposed to commit crimes.[25] Because her crime is strong evidence that she is so disposed, it undermines conditions of trust by lowering the credence with which others are justified in believing that she is not so disposed. Thus, conditions of trust are undermined to a degree, and they can be restored to a degree.

---

24. *Cf.* Aurel Kolnai, *Forgiveness*, 74 PROCEEDINGS OF THE ARISTOTELIAN SOCIETY, 91, 101 (1973-74) (taking forgiveness to involve the judgement that the forgiven has undergone a change in heart); MURPHY, *supra* note 1, at 24 (stating that one reason people forgive a wrongdoer is because she has repented or had a change of heart); HAMPTON, *supra* note 8, at 86 (stating that forgiving someone for a wrongful act is "a way of removing it as evidence of the state of his soul, so that one is able to judge him favorably without it," and forgiving her for the act involves the judgment that it "does not provide good evidence of the condition of the wrongdoer's soul"); GRISWOLD, *supra* note 14, at 50 (suggesting that forgiving a wrongdoer involves the judgement that she has shown through deeds and words a commitment "to becoming the sort of person who does not inflict injury"); Allias, *supra* note 1, at 56-57 (stating that "when you forgive the perpetrator your attitude towards her as a person is no longer the negative one that her wrongdoing supports; in other words, the act is disregarded in your ways of regarding and esteeming her, and in this sense, the slate is wiped clean, and the act is not held against her"); David Sussman, *Kantian Forgiveness*, 96 KANT-STUDIEN 85, 86 (2005) (pointing out that forgiveness involves a response to wrongdoing that "can repair significant violations of trust).

25. A person's credence in a belief is her subjective probability or degree of confidence that the belief is true. For discussion on the concept of subjective probabilities and their relation to objective probabilities, see DAVID LEWIS, *A Subjectivist's Guide to Objective Chance*, *in* PHILOSOPHICAL PAPERS: VOLUME II 83 (1986).

Fully forgiving the criminal for her crime involves judging that she has fully restored the conditions of trust she undermined by committing it. Given her crime's evidentiary significance, full forgiveness involves judging that she has provided others with equally strong countervailing evidence that she is no longer disposed to commit crimes. Other things being equal, it involves judging that she has justified others in believing that she is not so disposed with the credence with which they would have been justified in believing this if she had not committed the crime. Thus, fully forgiving the criminal for her crime involves ceasing to demand her to restore any of the conditions of trust she undermined by committing the crime, and so it involves ceasing to demand her to undertake any burdens to restore them. So full forgiveness involves fully suspending all blame toward the criminal for her crime.

Unlike full forgiveness, forgiving a criminal only partially for her crime involves judging that she has restored some but not all the conditions of trust she undermined by committing it. Given the crime's evidentiary significance, mere partial forgiveness involves judging that she has provided others with strong but not equally strong countervailing evidence that she is no longer disposed to commit crimes. Other things being equal, it involves judging that she has justified others in believing that she is not so disposed with a higher credence, but not a credence as high as the one with which they would have been justified in believing this if she had not committed the crime. Hence, partial forgiveness can involve ceasing to demand her to restore some but not all the conditions of trust she undermined by committing her crime, and so it can involve ceasing to demand her to undertake some but not all the burdens necessary to restore those conditions. Partial forgiveness can involve continuing to demand the criminal to restore the conditions of trust left undermined by her crime, and it can involve continuing to demand her to undertake the burdens necessary to restore them. So partial forgiveness can involve suspending some but not all blame toward the criminal for her crime.

C. Forgiving from Three Perspectives

Forgiveness can be given from three perspectives corresponding to three attitudes of blame.[26] Assume someone commits a crime against another. As a personal reactive attitude, the victim is warranted in resenting the criminal for her crime. As an impersonal reactive attitude, anyone is warranted in feeling indignation toward her for the crime. To forgive the criminal from the perspective of the victim or a third party, they must suspend their warranted resentment or indignation by making the required judgment and ceasing the relevant demands in response.[27] Because the criminal specifically disrespected the rights of the victim, he likely finds the prospect of forgiving her most challenging. For this reason, the criminal seeking forgiveness should be concerned primarily with the victim's perspective.

Forgiveness can also be given from the perspective of the criminal herself. As a self-reactive attitude, the criminal is warranted in feeling guilty for her crime. Mutatis mutandis the same presuppositions and demands that are constitutive of resentment and indignation are also constitutive of guilt. In feeling guilty for her crime, the criminal presupposes that she undermined certain conditions of trust by committing it, and she demands herself to restore those conditions by undertaking certain burdens in order to demonstrate to others that she has a good will.[28] To forgive herself for the crime, she must suspend her guilt by suspending its constitutive demands in response to her

---

26. *See* STRAWSON, *supra* note 3, at 13-15.

27. Although some writers assume that only the victim of a crime can forgive the criminal, there is no principled reason to hold such an exclusive view. *See* Eve Garrard & David McNaughton, *In Defence of Unconditional Forgiveness*, 104 PROCEEDINGS OF THE ARISTOTELIAN SOCIETY 39, 45 n. 6 (2003) (leaving room for forgiveness from third parties).

28. A person can be warranted in feeling guilty for committing a crime that has not been detected by others. In this case, she still presupposes that she undermined conditions of trust by committing her crime in the sense that if others knew she committed it, they would not be justified in believing that she is not disposed to commit crimes. In demanding herself to restore these conditions, she demands herself to undertake a course of action that would justify others in believing that she is not so disposed even if they were to learn that she committed the crime. Her demand is warranted given the inevitable risk that her crime will be detected and others will incur the costs of insecurity if she does not restore the conditions of trust.

judgment that she has restored the conditions of trust.

D. Warranted Forgiveness

The conditions under which forgiveness is warranted follow from its meaning. Suppose others fully forgive a criminal for her crime. Their full forgiveness is warranted if and only if they are warranted in judging that she has fully restored the conditions of trust she undermined by committing the crime. Assume they only partially forgive her. They judge that she has only partially restored the conditions of trust she undermined by committing the crime, and in response, they reduce the severity of the burdens they demand her to undertake to restore the other conditions of trust left undermined by her crime. Their partial forgiveness is warranted if and only if they are warranted in judging that she has restored the conditions of trust to the relevant degree, and they are warranted in judging that she need not undertake the more severe burdens to restore the other conditions of trust that they continue demanding her to restore. If others are not warranted in making these judgments, they might make them nonetheless. So they might forgive her nonetheless. But in this case, their forgiveness would not be warranted.

E. The Value of Warranted Forgiveness

When others are warranted in forgiving a criminal, their forgiveness would promote at least four values. First, they would rationally reduce the costs of insecurity they incur in response to her crime. They would fear the criminal less. They would engage in activities that would otherwise leave them too vulnerable to her. And they would not invest as heavily in costly precautionary measures to protect themselves from her. Second, the criminal's self-esteem would improve upon learning that others forgive her. People are ashamed to be regarded as untrustworthy. Third, forgiveness would promote reconciliation between the criminal and others. They would enter into relationships with each other, like friendships, that would not be rationally possible in the absence of warranted forgiveness. Fourth, forgiveness would benefit the relationships, projects, and commitments of others generally. For the feeling of blame itself can be an

all consuming experience that impairs one's capacity to sustain these important parts of one's life.

V. Implications

Given my theory of forgiving criminals, we can see where it stands on some potentially controversial issues concerning the subject. One issue concerns the relation between forgiveness, punishment, apology, and compensation. Others concern forgiving the untrustworthy, the unforgivable, forgiving the dead, and the elective nature of forgiveness.

A. Punishment, Apology, and Compensation

Being warranted in forgiving a criminal not only is consistent with, but also standardly requires her having undertaken a punishment. To be warranted in forgiving her, others must be justified in believing that she has restored the conditions of trust she undermined by committing her crime. As we discussed, to restore these conditions, she must demonstrate that she has a good will. To do so, she must undertake certain burdens for the sake of benefiting others. Such burdens would standardly constitute a punishment taking the form of labor intensive community service.

Providing an apology and compensation is also a necessary condition of warranted forgiveness.[29] To warrant others in judging that she has restored the conditions of trust, the criminal must apologize for her crime, and she must compensate her victims for the harm she caused them.[30] Neglecting to do either would indicate that she has not rectified the deficiency in the degree to which she cares about others that she manifested in committing her crime. More specifically, it would indicate a persisting insufficient

---

29. *Cf.* GRISWOLD, *supra* note 14, at 49-50 (suggesting that warranted forgiveness requires the wrongdoer to communicate contrition and her regret that she performed the act).

30. In addition, other steps might also be necessary to restore the conditions of trust and, hence, warrant forgiveness. For example, the criminal 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.

concern for the interests of her victims. While necessary, providing an apology and compensation might or might not be sufficient to warrant forgiveness. This depends on the form they take.

On the one hand, neither might be burdensome for the criminal to provide. For example, she might apologize through mere cheap talk; she might compensate her victims by effortlessly providing them with resources that she no longer needs or wants; or she might compensate them by directing a third party insurance company to compensate them on her behalf. Assuming an apology and compensation are not burdensome for her to provide, their provision would not demonstrate that she has rectified the deficiency in her concern for others.

On the other hand, providing an apology and compensation might be burdensome. For example, the criminal might express her apology and compensate her victims through engaging in labor intensive community service aimed at benefiting them. In this case, their provision would be part of her punishment for the crime.[31] Ultimately, the extent to which an apology and compensation warrant forgiveness depends on the extent to which their provision constitutes the punishment that the criminal undertakes to restore the conditions of trust.

Although warranted forgiveness standardly requires a past punishment, the issue of whether it is consistent with a warranted demand for further punishment depends on whether it is full or only partial. Warranted partial forgiveness is consistent with such a demand. When others are warranted in only partially forgiving a criminal, they are justified in believing that she has restored some but not all the conditions of trust she undermined by committing her crime. So they can be warranted in demanding her to undertake a further punishment to restore the conditions of trust left undermined. Warranted full forgiveness, however, is not consistent with a warranted demand for

31. *See* R.A. DUFF, PUNISHMENT, COMMUNICATION, AND COMMUNITY 106 (2001) (noting that a criminal's undertaking a punishment for her crime can constitute a forceful expression of her apology).

further punishment. When others are warranted in fully forgiving a criminal, they are justified in believing that she has restored all the conditions of trust she undermined by committing it. So they are not warranted in demanding her to undertake any further punishment to restore such conditions. As a consequence of RS, they would not be warranted in demanding her to undertake any further punishment for her crime. Moreover, it seems conceptually impossible for others fully to forgive a criminal for her crime and at the same time punish her for the crime themselves. For full forgiveness would involve suspending all blame toward her for the crime, and punishing her for it would express some blame.[32]

B. Forgiving the Untrustworthy

If a criminal restores some but not all the conditions of trust she undermined by committing her crime, then she remains untrustworthy at least to a degree. So others can be warranted in partially forgiving an untrustworthy criminal for her crime. There are also at least two situations in which others could be warranted in fully forgiving an untrustworthy criminal.

One involves a repeat criminal. Her initial crime undermines conditions of trust to a particular degree. It justifies others in believing with an unduly high credence that she is disposed to commit crimes, and there is a serious deficiency in the degree to which she cares about the interests of others. Assume she refuses to restore these conditions of trust. She refuses to apologize for the crime or to compensate any of her victims or to sacrifice any of her personal interests for the sake of benefiting anyone relevantly related to her victims. So she is untrustworthy, and others are not warranted in forgiving her even partially for the crime.

Assume she commits a second crime against different victims. Her second crime

32. *See* JOEL FEINBERG, *The Expressive Function of Punishment*, *in* DOING AND DESERVING: ESSAYS IN THE THEORY OF RESPONSIBILITY 95, 98 (1970) (discussing the conceptual connection between punishment and the expression of moral blame).

undermines additional conditions of trust. It justifies others in believing with an even higher credence that she has an even worse disposition to commit crimes and an even worse deficiency in the degree to which she cares about others.[33] Suppose, though, that she fully restores the conditions of trust she undermined by committing her second crime. She apologizes for it, compensates its victims, and sacrifices some of her sufficiently important personal interests for a sufficiently long time for the sake of benefiting people relevantly related to its victims. Because she fully restored the conditions of trust she undermined by committing her second crime, others are warranted in fully forgiving her for it even though her first crime still makes her untrustworthy. In general, others can be warranted in fully forgiving a repeat criminal for some but not all of her crimes, and the remaining crimes might make her untrustworthy at least to a degree.

A second situation concerns the self-fulfilling nature of trust.[34] People intrinsically desire to be trusted because they intrinsically desire others to hold them in esteem, and trusting them is one way of doing so. They instrumentally desire trust because it is necessary for the formation and flourishing of close relationships. They also desire trust because it reduces others' costs of insecurity. So trusting someone provides her with a good that she desires to maintain by not doing anything to undermine it. Hence, trusting someone not to commit crimes can provide her with additional prudential and altruistic incentives not to commit them. Thus, trusting someone can itself make her more trustworthy.

The self-fulfilling nature of trust entails that forgiving a criminal can itself make her more trustworthy.[35] Because forgiving her involves judging that she has become

33. Her disposition to commit crimes might be worse in the sense that she is willing to commit more serious crimes, and she is willing to commit crimes with a higher frequency, in a broader range of situations, and against a broader range of people.

34. *See, e.g.*, Kolnai, *supra* note 24, at 105; Philip Pettit, *The Cunning of Trust*, 24 PHIL. & PUB. AFF. 202, 212-17 (1995).

35. *See* Kolnai, *supra* note 24, at 102-03.

more trustworthy, it standardly involves placing a higher degree of trust in her by believing with a higher credence that she is not disposed to commit crimes. Assuming the additional trust provides her with a good that she desires to maintain, forgiving her can itself make her more trustworthy. The trust warranting effect of forgiveness entails that others can be warranted in fully forgiving an untrustworthy criminal for her crime when forgiving her would make her trustworthy or at least completely restore the conditions of trust she undermined by committing it.

To illustrate, suppose a criminal proceeds to restore the conditions of trust she undermined by committing her crime. She proceeds to undertake the required punishment and provide the required apology and compensation. In doing so, she evinces a significant desire for others' trust. Now suppose she succeeds in restoring the conditions of trust up to the point at which forgiving her would itself completely restore them. At this point, she has not fully restored the conditions of trust and so is still untrustworthy. However, others would be warranted in fully forgiving her because their doing so would itself complete the restoration and make her trustworthy.

For at least three reasons, though, the trust warranting effect of forgiveness cannot warrant full or even partial forgiveness unless the criminal undertakes a punishment and provides an apology and compensation to demonstrate that she has a good will.[36] First, the trust warranting effect of forgiveness is marginal at best. Although its evidentiary significance can be non-trivial, it can restore the conditions of trust only to a partial degree and make the criminal only marginally more trustworthy.

Second, forgiveness has a non-trivial trust warranting effect only if others are justified in believing that the criminal strongly desires their trust. Others would be justified in believing this only if they were justified in believing that the criminal is willing to restore the conditions of trust by undertaking a punishment. Unless the

36. *See* Kolnai, *supra* note 24, at 103 (stating that if a wrongdoer has provided no sign of a change of heart, then the reconciling or reforming effect of forgiveness would be "utterly dubious").

criminal makes such a sacrifice, others would not be justified in believing that her desire for their trust is sufficiently strong.

Third, forgiveness has a trust warranting effect in large part by providing the criminal with a prudential incentive not to commit crimes. But to restore the conditions of trust fully or to a significant degree, the criminal must demonstrate to others that she is appropriately motivated by the moral reasons not to commit crimes. She must demonstrate that she is disposed not to commit them even under conditions in which doing so would be in her best personal interests. To demonstrate this, she must demonstrate that she cares highly about the interests of others. To do so, she must undertake the relevant punishment and provide an apology and compensation. By demonstrating that she cares highly about others, she would thereby strengthen the trust warranting effect of forgiving her by indicating to others that she is appropriately motivated by the altruistic reasons not to undermine the trust inherent in their forgiveness.

C. The Unforgivable

A criminal might be unforgivable for her crime in two senses. On a weak sense, she is unforgivable for the crime if and only if she is blameworthy for it, and others would not be warranted in forgiving her for it if they knew all the facts about her capacities and what she has actually done. On the weak sense, the criminal might be fully or partially unforgivable. She is partially unforgivable for her crime only if others would not be warranted in fully forgiving her for it under the idealized conditions. She is fully unforgivable for the crime only if others would not be warranted in even partially forgiving her for it under those conditions. A defiant criminal is unforgivable in the weak sense because she refuses to restore the conditions of trust she undermined by committing her crime. At one extreme, she might refuse to undertake any punishment or to provide any apology or compensation in response to the crime.

Although unforgivable in the weak sense, defiant criminals need not be unforgivable in a stronger sense. On a strong sense, a criminal is unforgivable for her

crime if and only if she is blameworthy for it, and there is nothing she could do to warrant others in forgiving her for the crime if they were to know all the facts about her capacities and actions. On the strong sense, a criminal might also be fully or partially unforgivable depending on whether she could warrant others in fully or partially forgiving her for the crime under the idealized conditions. A defiant criminal need not be unforgivable in the strong sense because she might be able to restore the conditions of trust even though she chooses not to. Although no criminal is fully unforgivable in the strong sense, two considerations could make her partially unforgivable in this sense.

A criminal might lose her capacity to respond appropriately to some of the moral reasons against violating the rights of others. The seriousness of a crime justifies others in believing that the criminal is disposed to flout a particular class of such reasons. To restore the conditions of trust she undermined by committing her crime, the criminal must demonstrate that she is appropriately motivated by the relevant moral reasons. Assuming she has the capacity to respond appropriately to such reasons, she can demonstrate that she is appropriately motivated by them. But assume she later loses the capacity to respond appropriately to some of the relevant moral reasons.[37] If others know this, then she cannot justify their believing that she is appropriately motivated by all the relevant moral reasons. She can justify their believing that she is appropriately motivated by only some of them. So she can warrant their judging that she has restored some but not all the conditions of trust she undermined by committing her crime. Because she can warrant others in only partially forgiving her for the crime, she is partially unforgivable for it in the strong sense.[38]

---

37. She might lose the capacity in two ways. She might lose the epistemic capacity to recognize the reasons or to appreciate them as reasons. Or she might lose the volitional capacity to control what she intends in response to her judgments about what the reasons count for or against her intending.

38. After the criminal restores the conditions of trust as much as she can, she would no longer be blameworthy for her crime to any degree, as I later explain. At that point, others might suspend all blame toward her in response to their judgment that she lacks the capacity to restore the conditions any further. But their full suspension of blame would not constitute full forgiveness because it would not be in response to their judgment that she has fully restored the conditions of trust she undermined by committing her crime.

A worse incapacity, though, would not make the criminal fully unforgivable. Suppose she later loses her capacity to respond appropriately to any of the relevant moral reasons. If others know this, then she cannot justify their believing that she is appropriately motivated by any of these reasons. So she cannot warrant their judging that she has restored any of the conditions of trust she undermined by committing her crime. Hence, she cannot warrant their even partially forgiving her for the crime. But this does not entail that she is unforgivable for it because unforgivability presupposes blameworthiness. When the criminal loses her capacity to respond appropriately to any of the relevant moral reasons, she is no longer blameworthy for the crime because the incapacity would provide her with a full exculpatory defense for it. Others would no longer be warranted in blaming her for the crime because doing so would involve demanding her to restore conditions of trust that she lacks the capacity to restore. And in general, others are not warranted in demanding someone to do something that she lacks the capacity to do.[39]

If a criminal retains the capacity to respond appropriately to the relevant moral reasons, then she cannot be fully unforgivable for her crime in the strong sense. She can at least partially restore the conditions of trust by undertaking a punishment and providing an apology and compensation to her victims.[40] However, the seriousness of her crime combined with her age might still make her partially unforgivable for it in the strong sense. Consider two cases.

Mere partial forgiveness is consistent with the full suspension of blame.

39. This principle is similar to the principle that someone ought to do something only if she can do it. *See, e.g.*, David Copp, *'Ought' Implies 'Can', Blameworthiness, and the Principle of Alternate Possibilities*, *in* MORAL RESPONSIBILITY AND ALTERNATIVE POSSIBILITIES 265, 271-75 (David Widerker & Michael McKenna eds., 2003); 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").

40. *Cf.* Trudy Govier, *Forgiveness and the Unforgivable*, 36 AM. PHIL. Q. 59, 69-71 (1999) (emphasizing that even persons who commit atrocities can undergo positive moral changes in their character).

First, suppose the criminal commits a monstrous crime in which she disrespects the rights of others to a monstrously bad degree. Her crime is strong evidence that there is a monstrously bad deficiency in the degree to which she cares about the interests of others. To restore fully the conditions of trust, she must demonstrate that she has fully rectified this deficiency. To do so, she must act with an extremely high degree of benevolence for an extremely long time. Given the unavoidable constraints on the duration of her life, it might be impossible for her to do so since the required time would inevitably extend past her death. Second, suppose she commits a less serious crime, but does so very late in life. To restore fully the conditions of trust, she would need to act with a high degree of benevolence for a long time that would again inevitably extend past her death. In either case, she can restore the conditions of trust partially but not fully. So she can warrant others in partially but not fully forgiving her for either crime.

D. Forgiving the Dead

Death per se does not warrant forgiveness. When a criminal dies, she does restore the conditions of trust in some sense. If others know she died, they are no longer justified in believing she is still disposed to commit crimes. However, the judgment constitutive of forgiveness is not merely that the criminal has restored the conditions of trust. It is that she has restored them in the right way, demonstrating to others that she has a good will in the sense of being appropriately motivated by the moral reasons against violating the rights of others. A dead criminal has no will at all.

Assume, though, that the criminal restored the conditions of trust in the right way before she died. Whether others can be warranted in forgiving the criminal after her death depends on what they are justified in believing. On the one hand, suppose they are justified in believing she is still alive. Then they can still be warranted in blaming her for the crime. So when they learn that she restored the conditions of trust, they can be warranted in forgiving her because they can suspend their blame toward her in response to their warranted judgment that she restored the conditions.

On the other hand, suppose they know she died. Then they cannot be warranted in blaming her because a dead person lacks the capacities required to be blameworthy. So they cannot forgive her because forgiveness involves the suspension of an attitude of blame that was warranted prior to the considerations that make the judgment constitutive of forgiveness warranted. However, the concept of forgiveness is still relevant in assessing the moral quality of the criminal's life. Although the others cannot forgive her when they learn that she restored the conditions of trust, they are warranted in judging that she was forgivable in the sense that they would have been warranted in forgiving her when she was alive if they had known she restored those conditions.[41] When they learn that the criminal was forgivable in this sense, they learn that the moral quality of her life was better than they previously believed.

E. The Elective Nature of Forgiveness

The decision to forgive a criminal is elective in the sense that others are not obligated to forgive her even when it would be warranted.[42] In general, a person is not obligated to others to feel only warranted attitudes toward them or make only warranted demands on them. She might have most reason not to feel such attitudes or make such demands, and so feeling the attitudes or making the demands might be rationally deficient. But the reasons at issue do not generate an obligation.[43] Merely feeling an unwarranted attitude or making an unwarranted demand on someone would not cause her the kind of harm from which she has a right to be free. So a criminal is not warranted in demanding others to forgive her even if their forgiveness would itself be warranted.

Nevertheless, it is entirely appropriate for the criminal to request warranted

41. *Cf.* GRISWOLD, *supra* note 14, at 120 (noting that the only mode of forgiveness available to victims of dead wrongdoers is in the subjunctive).

42. *See* GRISWOLD, *supra* note 14, at 67-69; Sussman, *supra* note 24, at 87.

43. Although someone is not obligated to abstain from making unwarranted demands on others, she is obligated not to force them to fulfill such demands since such force would cause them the kind of harm from which they have a right to be free.

forgiveness for at least two reasons.[44] First, as we have discussed, warranted forgiveness would promote several good consequences for both the criminal and others. Second, the criminal's requesting warranted forgiveness would itself have good effects. It would indicate that she desires to be trusted, and so it would increase the trust warranting effect of forgiving her. Conversely, her not requesting forgiveness would indicate an indifference to whether others trust her and to the fact that they will incur the costs of insecurity in response to not trusting her. So her requesting warranted forgiveness would play an important role in restoring the conditions of trust she undermined by committing her crime. Thus, a criminal is not only permitted, but also obligated to request forgiveness from others when their forgiveness would be warranted.

VI. Conclusion

When others forgive a criminal for her crime, they judge she has restored the conditions of trust she undermined by committing it. More specifically, they judge she has restored those conditions by demonstrating that she has a good will. So forgiveness expresses an attitude of moral approval. But it in no way expresses approval of the crime at issue. It does not involve a change in judgment about the seriousness of the crime itself. Rather forgiveness expresses approval of the way the criminal has responded to her crime. It involves a change in judgment about the moral quality of her dispositions. For this reason, forgiveness makes reconciling with the criminal rationally possible without condoning the crime she committed.[45]

---

44. *See* GRISWOLD, *supra* note 14, at 50 n. 8 (crediting Ken Taylor with noting that when an offender requests or invites rather than demands forgiveness, he "shows respect and sympathy for his victim").

45. *See* Kolnai, *supra* note 24, at 95-98 (noting that forgiveness is distinct from condonation).

Bibliography

Abbott v. The Queen, 1977 A.C. 755.

Alexander, Larry. "Consent, Punishment, and Proportionality." *Philosophy and Public Affairs* 15 (1986): 178-82.

---. "Insufficient Concern: A Unified Conception of Criminal Culpability." *California Law Review* 88 (2000): 931-54.

---. "The Doomsday Machine: Proportionality, Punishment and Prevention." *The Monist* 63 (1980): 199-227.

Allias, Lucy. "Wiping the Slate Clean: The Heart of Forgiveness." *Philosophy and Public Affairs* 36 (2008): 33-68.

Arenella, Peter. "Character, Choice and Moral Agency." *Social Philosophy and Policy* 7 (1990): 59-83.

Aristotle. *Nicomachean Ethics*. Trans. Terence Irwin. 2d ed. Indianapolis: Hackett Publishing Company, 1999.

Bacharach, Michael, and Diego Gambetta. "Trust in Signs." *Trust in Society*. Ed. Karen S. Cook. New York: Russell Sage Foundation, 2001. 148-84.

Baier, Annette. "Trust and Antitrust." *Ethics* 96 (1986): 231-60.

Baird, Douglas G., Robert H. Gertner, and Randal C. Picker. *Game Theory and the Law*. Cambridge: Harvard University Press, 1994.

Baldwin, Thomas. "Punishment, Communication, and Resentment." *Punishment and Political Theory*. Ed. Matt Matravers. Oxford: Hart Publishing, 1999. 124-32.

Baron, Marcia. "Excuses, excuses." *Criminal Law and Philosophy* 1 (2007): 21-39.

---. "Justifications and Excuses." *Ohio State Journal of Criminal Law* 2 (2005): 387-406.

Bayles, Michael D. "Character, Purpose, and Criminal Responsibility." *Law and Philosophy* 1 (1982): 5-20.

---. "Hume on Blame and Excuse." *Hume Studies* 2 (1976): 17-35.

Bazelon, David L. "The Morality of the Criminal Law." *Southern California Law Review* 49 (1976): 385-405.

Becker, Gary S. "Crime and Punishment: An Economic Approach." *Journal of Political Economy* 76 (1968): 169-217.

Benn, S.I. "An Approach to the Problems of Punishment." *Philosophy* 33 (1958): 325-41.

Bentham, Jeremy. *An Introduction to the Principles of Morals and Legislation*. Eds. J.H. Burns and H.L.A. Hart. New York: Oxford University Press, 1970.

Berman, Mitchell N. "Justifications and Excuses, Law and Morality." *Duke Law Journal* 53 (2003): 1-77.

Binmore, Ken, and Alex Voorhoeve. "Defending Transitivity against Zeno's Paradox." *Philosophy and Public Affairs* 31 (2003): 272-79.

Blum, Lawrence A. *Friendship, Altruism and Morality*. London: Routledge & Kegan Paul, 1980.

Brandt, R.B. "A Motivational Theory of Excuses in the Criminal Law." *Criminal Justice: Nomos XXVII*. Eds. J. Roland Pennock & John W. Chapman. New York: New York University Press, 1985. 165-200.

---. "A Utilitarian Theory of Excuses." *Philosophical Review* 78 (1969): 337-61.

---. "Blameworthiness and Obligation." *Essays in Moral Philosophy*. Ed. A.I. Melden. Seattle: University of Washington Press, 1958. 3-39.

---. "Traits of Character: A Conceptual Analysis." *American Philosophical Quarterly* 7 (1970): 23-37.

Bratman, Michael E. *Intention, Plans, and Practical Reason*. Cambridge: Harvard University Press, 1987.

Britt, Chester L. "Versatility." *The Generality of Deviance*. Eds. Travis Hirschi and Michael R. Gottfredson. New Brunswick: Transaction Publishers, 1994. 173-92.

Broome, John. "Reason and Motivation." *Supplement to the proceedings of The Aristotelian Society* 71 (2003): 131-47.

Burgh, Richard. "Do the Guilty Deserve Punishment?" *Journal of Philosophy* 79 (1982): 193-210.

Butler, Joseph. "Sermon VIII: Upon Resentment" and "Sermon IX: Upon Forgiveness of Injuries." *The Works of Joseph Butler*. Vol. 2. Ed. W.E. Gladstone. Oxford: Clarendon Press, 1896. 136-67.

Cartwright, Rosalind. "Sleepwalking Violence: A Sleep Disorder, a Legal Dilemma, and a Psychological Challenge." *American Journal of Psychiatry* 161 (2004): 1149-58.

Conee, Earl, and Richard Feldman. *Evidentialism: Essays in Epistemology*. New York: Oxford University Press, 2004.

Cooper, David. "Hegel's Theory of Punishment." *Hegel's Political Philosophy: Problems and Perspectives*. Ed. Z.A. Pelczynski. Cambridge: Cambridge University Press, 1971. 151-67.

Cooter, Robert, and Thomas Ulen. *Law and Economics*. 3d ed. Boston: Addison Wesley Publishing Company, 2000.

Copp, David. "'Ought' Implies 'Can', Blameworthiness, and the Principle of Alternate Possibilities." *Moral Responsibility and Alternative Possibilities*. Eds. David Widerker and Michael McKenna. Burlington: Ashgate, 2003. 265-300.

Corrado, Michael. "Notes on the Structure of a Theory of Excuses." *Journal of Criminal Law and Criminology* 82 (1992): 465-97.

Dagger, Richard. "Playing Fair with Punishment." *Ethics* 103 (1993): 473-88.

D'Arms, Justin, and Daniel Jacobson. "Sentiment and Value." *Ethics* 110 (2000): 722-48.

---. "The Moralistic Fallacy: On the 'Appropriateness' of Emotions." *Philosophy and Phenomenological Research* 61 (2000): 65-90.

Darwall, Stephen. *The Second-Person Standpoint: Morality, Respect, and Accountability*. Cambridge: Harvard University Press, 2006).

---. "Two Kinds of Respect." *Ethics* 88 (1977): 36-49.

Davidson, Donald. "Agency." *Essays on Actions and Events*. Oxford: Clarendon Press, 1980. 43-61.

Delgado, Richard. "Ascription of Criminal States of Mind: Toward a Defense Theory for the Coercively Persuaded ("Brainwashed") Defendant." *Minnesota Law Review* 63 (1978): 1-33.

---. "'Rotten Social Background': Should the Criminal Law Recognize a Defense of Severe Environmental Deprivation?." *Law and Inequality* 3 (1985): 9-90.

Dimock, Susan. "Retributivism and Trust." *Law and Philosophy* 16 (1997): 37-62.

Dixit, Avinash, and Susan Skeath. *Games of Strategy*. 2d ed. New York: W.W. Norton, 2004.

Dolinko, David. "Some Thoughts about Retributivism." *Ethics* 101 (1991): 537-59.

Doris, John. *Lack of Character: Personality and Moral Behavior*. New York: Cambridge University Press, 2002.

Dressler, Joshua. "New Thoughts About the Concept of Justifications in the Criminal Law: A Critique of Fletcher's Thinking and Rethinking." *UCLA Law Review* 32 (1984): 61-99.

---. *Understanding Criminal Law*. New York: Matthew Bender, 1995.

Duff, R.A. *Punishment, Communication, and Community*. New York: Oxford University Press, 2001.

---. "Rethinking Justifications." *Tulsa Law Review* 39 (2004): 829-50.

---. *Trials and Punishments*. New York: Cambridge University Press, 2005.

Ehrlich, Isaac. "The Optimum Enforcement of Laws and the Concept of Justice: A Positive Analysis." *International Review of Law and Economics* 2 (1982): 3-27.

Epstein, Seymour. "Aggregation and Beyond: Some Basic Issues on the Prediction of Behavior." *Journal of Personality* 51 (1983): 360-92.

Feinberg, Joel. "Duties, Rights, and Claims." *Rights, Justice, and the Bounds of Liberty: Essays in Social Philosophy*. Princeton: Princeton University Press, 1980. 130-42.

---. *Harmless Wrongdoing*. New York: Oxford University Press, 1988.

---. *Harm to Self*. New York: Oxford University Press, 1986.

---. "The Expressive Function of Punishment." *Doing and Deserving: Essays in the Theory of Responsibility*. Princeton: Princeton University Press, 1970. 95-118.

Finkelstein, Claire. "Threats and Preemptive Practices." *Legal Theory* 5 (1999): 311-38.

Fischer, John Martin, and Mark Ravizza. *Responsibility and Control: A Theory of Moral Responsibility*. New York: Cambridge University Press, 1998.

Fletcher, George P. *Rethinking Criminal Law*. Boston: Little, Brown, 1978.

---. "The Right Deed for the Wrong Reason: A Reply to Mr. Robinson." *UCLA Law Review* 23 (1975): 293-321.

Frank, Robert H. *Passions within Reason: The Strategic Role of the Emotions*. New York: W.W. Norton, 1988.

Frankfurt, Harry. "Alternate Possibilities and Moral Responsibility." *Journal of Philosophy* 66 (1969): 829-39.

Friedman, David, and William Sjostrom. "Hanged for a Sheep: The Economics of Marginal Deterrence." *Journal of Legal Studies* 22 (1993): 345-66.

Fuller, Lon L. *The Morality of Law*. Rev. ed. New Haven: Yale University Press, 1964.

Garcia, J.L.A. "Two Concepts of Desert." *Law and Philosophy* 5 (1986): 219-35.

Gardner, John. "Justifications and Reasons." *Harm and Culpability*. Eds. A.P. Simester and A.T.H. Smith. New York: Oxford University Press, 1996. 103-29.

---. "The Gist of Excuses." *Buffalo Criminal Law Review* 1 (1997): 575-98.

Garrard, Eve, and David McNaughton. "In Defence of Unconditional Forgiveness." *Proceedings of the Aristotelian Society* 104 (2003): 39-60.

Garrison v. Louisiana, 379 U.S. 64 (1964).

Gibbard, Allan. *Wise Choices, Apt Feelings*. Cambridge: Harvard University Press, 1990.

Goffman, Erving. *The Presentation of Self in Everyday Life*. Rev. ed. Garden City, New York: Doubleday, 1959.

Golash, Deidre. "The Retributive Paradox." *Analysis* 54 (1994): 72-78.

Goldman, Alvin I. "What is Justified Belief?" *Justification and Knowledge*. Ed. George S. Pappas. Dordrecht: Reidel, 1979. 1-23.

Govier, Trudy. "Forgiveness and the Unforgivable." *American Philosophical Quarterly* 36 (1999): 59-75.

Gottfredson, Michael R., and Travis Hirschi. *A General Theory of Crime*. Palo Alto: Stanford University Press, 1990.

Greenawalt, Kent. "The Perplexing Borders of Justification and Excuse." *Columbia Law Review* 84 (1984): 1897-1927.

Griswold, Charles L. *Forgiveness: A Philosophical Exploration*. New York: Cambridge University Press, 2007.

Hampton, Jean. "Correcting Harms Versus Righting Wrongs: The Goal of Retribution." *UCLA Law Review* 39 (1991-92): 1659-702.

---. "Forgiveness, resentment and hatred." *Forgiveness and Mercy*. Eds. Jeffrie G. Murphy and Jean Hampton. New York: Cambridge University Press, 1988. 35-87.

---. "The Moral Education Theory of Punishment." *Philosophy and Public Affairs* 13 (1984): 208-38.

Harman, Gilbert H. "The Inference to the Best Explanation." *Philosophical Review* 74 (1965): 88-95.

Hart, H.L.A. "Are There any Natural Rights?" *Philosophical Review* 64 (1955): 175-91.

---. "Legal Responsibility and Excuses." *Punishment and Responsibility: Essays in the Philosophy of Law*. Oxford: Clarendon Press, 1968. 28-53.

---. "Prolegomenon to the Principles of Punishment." *Punishment and Responsibility: Essays in the Philosophy of Law*. Oxford: Clarendon Press, 1968. 1-27.

Hegel's Philosophy of Right. Trans. T.M. Knox. New York: Oxford University Press, 1942.

Hieronymi, Pamela. "Articulating an Uncompromising Forgiveness." *Philosophy and Phenomenological Research* 62 (2001): 529-55.

Hobbes, Thomas. *Leviathan*. Ed. Richard Tuck. New York: Cambridge University Press, 1996.

Hoekema, David A. "Trust and Obey: Toward a New Theory of Punishment." *Israel Law Review* 25 (1991): 332-50.

Holmes, O.W. *The Common Law*. Boston: Little, Brown, 1881.

Hume, David. *A Treatise of Human Nature*. 2d ed. Eds. L.A. Selby-Bigge & P.H. Nidditch. New York: Oxford University Press, 1978.

Husak, Douglas N. "Conflicts of Justifications." *Law and Philosophy* 18 (1999): 41-68.

---. "Why Punish the Deserving?" NOUS 26 (1992): 447-64.

Horder, Jeremy. *Excusing Crime*. New York: Oxford University Press, 2004.

Kadish, Sanford H. "Excusing Crime." *California Law Review* 75 (1987): 257-89.

Kahneman, Daniel, Jack L. Knetsch, and Richard H. Thaler. "Experimental Tests of the Endowment Effect and the Coase Theorem." *Journal of Political Economy* 98 (1990): 1325-48.

Kant, Immanuel. *Groundwork of the Metaphysics of Morals*. Ed. Mary Gregor. New York: Cambridge University Press, 1997.

---. *The Metaphysics of Morals*. Ed. Mary Gregor. New York: Cambridge University Press, 1996.

Kenny, Anthony. "Duress Per Minas as a Defence to Crime: II." *Law, Morality and Rights*. Ed. M.A. Stewart. Dordrecht: Reidel, 1983. 345-53.

Kolnai, Aurel. "Forgiveness." *Proceedings of the Aristotelian Society* 74 (1973-74): 91-106.

Kornblith, Hilary. "Beyond Foundationalism and the Coherence Theory," *Journal of Philosophy* 77 (1980): 597-612.

LaFave, Wayne R., Jerold H. Israel, and Nancy J. King. *Criminal Procedure: Hornbook Series*. 4th ed. St. Paul: West, 2004.

Langan, Patrick A., and David J. Levin. *Recidivism of Prisoners Released in 1994, Bureau of Justice Statistics Special Report*.. Http://www.ojp.usdoj.gov/bjs/pub/pdf/rpr94.pdf.

Lewis, David. "A Subjectivist's Guide to Objective Chance." *Philosophical Papers: Volume II*. New York: Oxford University Press, 1986. 83-113.

---. "Do We Believe in Penal Substitution?" *Papers in Ethics and Social Philosophy*. New York: Cambridge University Press, 2000. 128-35.

Lifton, Robert Jay. *Thought Reform and the Psychology of Totalism: A Study of "Brainwashing" in China*. New York: Norton, 1961.

Locke, John. *Second Treatise of Government*. Ed. C.B. Macpherson. Indianapolis: Hackett 1980.

Mackie, J.L. "Duress and Necessity as Defences to Crime: A Postscript." *Law, Morality and Rights*. Ed. M.A. Stewart. Dordrecht: Reidel, 1983. 365-69.

Matravers, Matt. *Justice and Punishment, The Rationale of Coercion*. New York: Oxford University Press, 2000.

McCord, Joan. "Understanding Motivations: Considering Altruism and Aggression." *Facts, Frameworks, and Forecasts*. Ed. Joan McCord. New Brunswick: Transaction Publishers, 1992. 115-36.

McMahan, Jeff. "The Basis of Moral Liability to Defensive Killing." *Philosophical Issues* 15 (2005): 386-405.

Michael, Jerome, and Herbert Wechsler. "A Rationale of the Law of Homicide II." *Columbia Law Review* 37 (1937): 1261-1325.

Milgram, Stanley. *Obedience to Authority*, New York: Harper & Row, 1974.

Mill, J.S. *On Liberty*. Ed. Elizabeth Rapaport. Indianapolis: Hackett, 1978.

---. *Utilitarianism*. Ed. Roger Crisp. New York: Oxford University Press, 1998.

Miller, Joshua D., and Donald Lynam. "Structural Models of Personality and their Relation to Antisocial Behavior: A Meta-Analytic Review." *Criminology* 39 (2001): 765-98.

Mischel, Walter. *Personality and Assessment*. New York: Wiley, 1968.

Model Penal Code.

Moore, G.E. *Principia Ethica*. Rev. ed. Ed. Thomas Baldwin. New York: Cambridge University Press, 1993.

Moore, Michael S. "Justifying Retributivism." *Israel Law Review* 27 (1993): 15-49.

Moral Responsibility and Alternative Possibilities: Essays on the Importance of Alternative Possibilities. Eds. David Widerker & Michael McKenna. Burlington: Ashgate, 2003.

Morris, Herbert. "A Paternalistic Theory of Punishment." *American Philosophical Quarterly* 18 (1981): 263-71.

---. "Persons and Punishment." *Theories of Punishment*. Ed. Stanley E. Grupp. Bloomington: Indiana University Press, 1971. 76-101.

Morse, Stephen J. "The Twilight of Welfare Criminology: A Reply to Judge Bazelon." *Southern California Law Review* 49 (1976): 1247-68.

Murphy, Jeffrie G., and Jules L. Coleman. *Philosophy of Law: An Introduction to Jurisprudence*. Rev. ed. Boulder: Westview, 1990.

Murphy, Jeffrie. "Forgiveness and resentment." *Forgiveness and Mercy*. Eds. Jeffrie G. Murphy and Jean Hampton. New York: Cambridge University Press, 1988. 14-34.

---. "Marxism and Retribution." *Philosophy and Public Affairs* 2 (1973): 217-43.

Nagel, Thomas. *The View from Nowhere*. New York: Oxford Universitiy Press, 1986.

Nino, C.S. "A Consensual Theory of Punishment." *Philosophy and Public Affairs* 12 (1983): 289-306.

Nozick, Robert. *Anarchy, State, and Utopia*. New York: Basic Books, 1974.

Olweus, Dan. "Stability of Aggressive Reaction Patterns in Males: A Review." *Psychological Bulletin* 86 (1979): 852-75.

Quinn, Warren. "Actions, Intentions, and Consequences: The Doctrine of Doing and Allowing." *Philosophical Review* 98 (1989): 287-312.

---. "The Right to Threaten and the Right to Punish." *Philosophy and Public Affairs* 14 (1985): 327-73.

Parfit, Derek. *Climbing the Mountain*. Forthcoming.

---. "Equality or Priority?" The Lindley Lecture, University of Kansas 3-4 (Nov. 21, 1991).

---. "Personal Identity." *Philosophical Review* 80 (1971): 3-27.

---. "Rationality and Reasons." *Exploring Practical Philosophy: From Action to Values*. Eds. Dan Egonsson, Jonas Josefsson, Bjorn Petersson, and Toni Ronnow-Rasmussen. Burlington: Ashgate, 2001. 17-41.

---. *Reasons and Persons*. New York: Clarendon Press, 1984.

Pettit, Philip. "The Cunning of Trust." *Philosophy and Public Affairs* 24 (1995): 202-25.

Piquero, Alex, Raymond Paternoster, Paul Mazerolle, Robert Brame, and Charles W. Dean. "Onset Age and Offense Specialization." *Journal of Research in Crime and Delinquency* 36 (1999): 275-99.

Polinsky, Mitchell A. *An Introduction to Law and Economics*. 2d ed. Boston: Little, Brown, 1989.

Posner, Richard A. *Economic Analysis of Law*. 5th ed. New York: Aspen, 1998.

Primoratz, Igor. "Punishment as Language." *Philosophy* 64 (1989): 187-205.

Rabinowicz, Wlodek, and Toni Ronnow-Rasmussen. "The Strike of the Demon: On Fitting Pro-Attitudes and Value." *Ethics* 114 (2004): 397-423.

Rawls, John. *A Theory of Justice*. Cambridge: Harvard University Press, 1971.

---. "Legal Obligation and the Duty of Fair Play." *Collected Papers*. Ed. Samuel Freeman. Cambridge: Harvard University Press, 1999). 117-29.

Reid, Thomas. "Of Mr. Locke's Account of Our Personal Identity." *Personal Identity*. Ed. John Perry. Berkeley: University of California Press, 1975. 113-118.

Riggs v. Palmer, 22 N.E. 188, 190 (N.Y. 1889).

Robinson, Paul H. "Criminal Law Defenses: A Systematic Analysis." *Columbia Law Review* 82 (1982): 199-291.

Robinson, Paul H., and John M. Darley. "Does Criminal Law Deter? A Behavioral Science Investigation." *Oxford Journal of Legal Studies* 24 (2004): 173-205.

Robinson, Paul H., and John M. Darley. "The Role of Deterrence in the Formulation of Criminal Law Rules: At Its Worst When Doing Its Best." *Georgetown Law Journal* 91 (2003): 949-1002.

Ross, Lee, and Richard E. Nisbett. *The Person and the Situation*. New York: Mcgraw-Hill, 1991.

Sabini, John, and Maury Silver. "Lack of Character? Situationism Critiqued." *Ethics* 115 (2005): 535-62.

Scanlon, T.M. "Punishment and the Rule of Law." *The Difficulty of Tolerance: Essays in Political Philosophy*. New York: Cambridge University Press, 2003. 219-33.

---. *What We Owe to Each Other*. Cambridge: Harvard University Press, 1998.

Scheffler, Samuel. "Doing and Allowing." *Ethics* 114 (2004): 215-39.

---. *The Rejection of Consequentialism*. Rev. ed. New York: Oxford University Press, 1994.

Schroeder, Mark. "Means-end Coherence, Stringency, and Subjective Reasons." *Philosophical Studies* (forthcoming).

Sendor, Benjamin B. "Mistakes of Fact: A Study in the Structure of Criminal Conduct." *Wake Forest Law Review* 25 (1990): 707-82.

Shaw, William. "The Consequentialist Perspective." *Contemporary Debates in Moral Theory*. Ed. James Dreier. Malden: Blackwell, 2006. 5-20.

Sher, George. *Desert*. Princeton: Princeton University Press, 1987.

Shoemaker, Sydney. "Persons and Their Pasts." *American Philosophical Quarterly* 7 (1970): 269-85.

Simmons, A. John. *Moral Principles and Poltical Obligations* Princeton: Princetong University Press, 1981.

Simon, Leonore M. J. "Do Criminal Offenders Specialize in Crime Types?" *Applied and Preventive Psychology* 6 (1997): 35-53.

Spence, A. Michael. *Market Signaling*. Cambridge: Harvard University Press, 1974.

Stephen, Sir James Fitzjames. *A History of the Criminal Law of England*. Vol. 2. London: MacMillan, 1883.

Stigler, George J. "The Optimum Enforcement of Laws." *Journal of Political Economy* 78 (1970): 526-36.

Strawson, P.F. "Freedom and Resentment." *Freedom and Resentment and Other Essays*. London: Methuen, 1974. 1-25.

Sussman, David. "Kantian Forgiveness." *Kant-Studien* 96 (2005): 85-107.

Tadros, Victor. *Criminal Responsibility*. New York: Oxford University Press, 2005.

Temkin, Larry S. "A Continuum Argument for Intransitivity." *Philosophy and Public Affairs* 25 (1996): 175-210.

---. "Inequality." *Philosophy and Public Affairs* 15 (1986): 99-121.

---. *Inequality*. New York, Oxford Univesity Press, 1993.

The Queen v. Parks [1992] 2 S.C.R. 871 (Canada).

U.S. v. Alexander, 471 F.2d 923 (D.C. Cir. 1973).

U.S. v. Batchelor, 19 C.M.R. 452 (1954)

U.S. v. Fleming, 19 C.M.R. 438 (1954)

U.S. v. Hearst, 412 F. Supp. 873 (N.D. Cal. 1976)

U.S. v. Olson, 20 C.M.R. 46 (1955).

van Inwagen, Peter. "Ability and Responsibility." *Philosophical Review* 87 (1978): 201-24.

Vitacco, Michael J., Craig S. Neumann, Angela A. Robertson, and Sarah L. Durrant. "Contributions of Impulsivity and Callousness in the Assessment of Adjudicated Male Adolescents: A Prospective Study." *Journal of Personality Assessment* 78 (2002): 87-103.

von Hirsch, Andrew. "Desert and Previous Convictions in Sentencing." *Minnesota Law Review* 65 (1981): 591-634.

Vranas, Peter B. M. "The Indeterminacy Paradox: Character Evaluations and Human Psychology." *Nous* 39 (2005): 1-42.

Vuoso, George. "Background, Responsibility, and Excuse." *Yale Law Journal* 96 (1987): 1661-86.

Watson, Gary. "Responsibility and the Limits of Evil: Variations on a Strawsonian Theme." *Perspectives on Moral Responsibility*. Eds. John Martin Fischer and Mark Ravizza. Ithaca: Cornell University Press, 1993. 119-50.

Westen, Peter, and James Mangiafico. "The Criminal Defense of Duress: A Justification, Not an Excuse - And Why It Matters." *Buffalo Criminal Law Review* 6 (2003): 833-950.

Williams, Bernard. "A Critique of Utilitarianism." *Utilitarianism: For and Against*. Eds. J.J.C. Smart and Bernard Williams. New York: Cambridge University Press, 1973. 75-150.

---. "The Self and the Future." *Problems of the Self*. New York: Cambridge University Press, 1973. 46-63.

Zahavi, Amotz and Avishag. *The Handicap Principle*. Trans. Naama Zahavi-Ely and Melvin Patrick Ely. New York: Oxford University Press, 1997.