Notre Dame Law Review

Volume 77 | Issue 5 Article 4

10-1-2002

Unraveling Unknowing Justification

Anthony M. Dillof

UNRAVELING UNKNOWING JUSTIFICATION

*Anthony M. Dillof**

Faust: . . . So, once again, who art thou?
Mephistopheles: A part of that power, which is ever willing evil and ever producing good.[1]

INTRODUCTION

This Article concerns a narrow but significant topic in the theory of criminal law—the problem of unknowing justification. Briefly stated, the problem of unknowing justification is this: Should a defendant be exonerated of an offense due to circumstances he was unaware of, where he would have been entitled to a justification defense had he known of the circumstances? Examples quickly bring this convoluted question into vivid relief:

> *D1*, a treacherous gold prospector, shoots his partner so he will not have to share their newly discovered gold with him just before, unknown to *D1*, his partner was to shoot him for the same reason.
> *D2*, a vengeful farmer, torches a neighbor's wheat field ignorant that the destruction of the field will create a firebreak which will save a nearby town from an approaching forest fire.
> *D3*, a racist police officer, arbitrarily arrests and restrains an African-American woman unaware that she is the subject of an outstanding arrest warrant.

Undoubtedly, had *D1*, *D2*, and *D3* acted based on the surrounding circumstances (the partner's imminent attack, the approaching forest fire, the outstanding warrant), they would have been entitled to the justification defenses of self-defense, necessity, and law enforcement, respectively. Because they did not know of these justifying circum-

* Associate Professor of Law, Texas Wesleyan University School of Law. A.B. 1981, Harvard University; J.D. 1985, Columbia University School of Law; LL.M. 1996, Columbia University School of Law. Howard Denmark, Stuart P. Green, George Fletcher, and Kenneth W. Simons favored me with their thoughtful comments on drafts of this Article.

1 JOHANN WOLFGANG VON GOETHE, FAUST 44 (Abraham Hayward trans., Oxford Univ. Press 1993) (1833).

stances, we might call *D1*, *D2*, and *D3* "unknowingly justified." Should we allow these actors to claim these justification defenses despite their ignorance of the justifying circumstances, or should we deny them these defenses due to their ignorance? Although there are sub-issues,[2] this is the principal question addressed herein.

The problem of unknowing justification has significant practical and theoretical implications for the criminal law. Practically, of course, we must decide what punishment, if any, to impose on the unknowingly justified actor. Our criminal justice system is far from ideal: criminal behavior is partially caused by social and economic inequities that the criminal law systematically discounts; limited resources for factfinding inevitably produce erroneous verdicts; inhumane and brutal conditions of confinement dehumanize inmates and condition them to further brutality. For various likely familiar reasons, these shortcomings are not easily remedied. In contrast, establishing just schedules of punishment is a matter directly within our control. If a person, as a moral matter, deserves to have her sentence decreased or nullified due to circumstances she was unaware of, the criminal law should reflect this fact. We owe it to those defendants who would be made to suffer unjustly otherwise.[3]

---

2 The problem of unknowing justification has two main sub-issues: (1) If an unknowingly justified defendant is entitled to assert a justification defense against offense *O*, is she entitled to assert the defense against the charge of attempting *O* as well, or will she be liable for attempting *O*?; and (2) if an unknowingly justified defendant is not entitled to assert a justification defense, are there any additional requirements besides knowledge, for example, justificatory purpose, that must be present in order for her to be entitled to assert the defense? In other words, if knowledge of the justifying circumstances is a necessary condition for asserting a justification defense, is it a sufficient one as well? *See infra* notes 47–51, 113–19 and accompanying text.

3 Cases of unknowing justification are reported periodically. The most discussed is *Queen v. Dadson*, 169 Eng. Rep. 407 (Cr. Cas. Res. 1850) (constable shoots fleeing suspect unaware that criminal record of suspect licensed his action; justification defense not permitted); *see also* Regina v. Thain, [1985] N. Ir. 457 (C.A.) (soldier shot fleeing teen plead defense that he was unaware of circumstances that may have allowed to him to use force to accomplish an arrest; justification defense permitted but rejected on the facts). The most recently publicized occurred in 1997 when a petty thief stole an innocent-looking backpack left in public. The thief discovered a terrorist bomb in the backpack, reported the bomb to the police, and saved many lives. *See* Paul H. Robinson, *The Bomb Thief and the Theory of Justification Defenses*, 8 CRIM. L.F. 387, 387–89, 408 (1997) (thief not prosecuted). Other cases discussing issue of unknowing justification are collected in 2 PAUL H. ROBINSON, CRIMINAL LAW DEFENSES 14 n.4 (1984). The frequency of cases of unknowing justification, of course, is irrelevant to the question of what justice requires. The strength of an individual's right to a just punishment does not vary with the frequency that the state has occasion to violate it.

On the theoretical level, the problem of unknowing justification presents one of the fundamental questions of criminal law theory. The concerns of the criminal law most naturally divide along the lines of objective conditions relevant to offenses and defenses and mental conditions relevant to inculpation and exculpation. Thus, a person is generally liable if there are conditions relevant to an offense (such as the fact that the goods being purchased are stolen) and she knows of them; generally not liable if there are such conditions and she is unaware of them; generally not liable if there are conditions relevant to a defense (such as the putative victim's consent) and she is aware of them; but what if there are circumstances relevant to a defense which she is unaware of? The situation may be represented diagrammatically as follows:

	Conditions of the Offense	Conditions of the Defense
Known	Liability	No Liability
Unknown	No Liability	?

No theory of criminal law is complete unless it can provide an answer to the problem of unknowing justification and so fill in the fourth square.

As shall become apparent in the course of this Article, the question mark in the fourth square implies profound questions for various fields of inquiry. Expressed in moral terms, the question becomes, what are we punishing for? Are we punishing people for harms they have caused, the unjustified harms they have caused, the unexcused harms they have caused, or something else? Expressed in terms of criminal law theory, the question becomes, what is the structure of a defense? Do defenses embody free-standing norms with objective and subjective components, like offenses, or do they act instead as exceptions to offense norms, which are legally and logically prior? Expressed in terms of political theory, the question becomes whether it is legitimate for the state to employ force against its citizens in the absence of the breach of its rules of conduct. Finally, we are led to think more deeply about why we exonerate in the *typical* case of justification where the defendant knows of the justificatory circumstances. Is it because of (1) the objective circumstances exclusively, (2) the defendant's mental state exclusively, (3) both the objective circumstances and the defendant's mental state, each a sufficient, independent and nonexclusive reason, or (4) the combination of the objective circumstances and the defendant's mental state, neither alone being a sufficient reason?

These questions have incited scholarly inquiry. Currently, however, there is no consensus concerning the correct resolution of the problem of unknowing justification. One camp steadfastly holds that the unknowingly justified actor should be completely exonerated;[4] an opposing camp disagrees and would impose full liability;[5] a third camp takes the intermediate position that the unknowingly justified actor should be liable merely for attempting the offense.[6] In the face of such scholarly dispute, one might be tempted to defer to the status quo. Yet recourse to the wisdom of settled law to resolve the issue is unavailing for the law's position on the issue is itself disputed.[7] Resolution of the problem of unknowing justification thus would unify theorists and provide guidance to legislatures and courts.

---

4 *See* MICHAEL S. MOORE, ACT AND CRIME: THE PHILOSOPHY OF ACTION AND ITS IMPLICATION FOR CRIMINAL LAW 180–81 (1993) (arguing that the unknowingly justified defendant "has done no wrong"); GLANVILLE WILLIAMS, CRIMINAL LAW: THE GENERAL PART 23–27 (2d ed. 1961); B. Sharon Byrd, *Wrongdoing and Attribution: Implications Beyond the Justification-Excuse Distinction*, 33 WAYNE L. REV. 1289, 1314–16 (1987); Paul H. Robinson, *Competing Theories of Justification: Deeds v. Reasons*, *in* HARM AND CULPABILITY 45, 47 (A.P. Simester et al. eds., 1996) ("[W]here a person's conduct in fact avoids a greater societal harm but the person is unaware of this, the conduct is justified despite the actor's ignorance."); Robert F. Schopp, *Justification Defenses and Just Convictions*, 24 PAC. L.J. 1233, 1267–82 (1993) ("[J]ustification defenses should not require internal justification based on justificatory knowledge.").

5 *See* GEORGE P. FLETCHER, RETHINKING CRIMINAL LAW § 7.4 (1978); J.C. SMITH, JUSTIFICATION AND EXCUSE IN THE CRIMINAL LAW 41–44 (1989); Russell L. Christopher, *Unknowing Justification and the Logical Necessity of the* Dadson *Principle in Self-Defence*, 15 OXFORD J. LEGAL STUD. 229, 251 (1995) (recommending the subjective approach); Michael Corrado, *Notes on the Structure of a Theory of Excuses*, 82 J. CRIM. L. & CRIMINOLOGY 465, 487–91 (1991) ("[T]here is a strange sense that it would be wrong to exonerate someone who did not have the proper state of mind."); Kent Greenawalt, *The Perplexing Borders of Justification and Excuse*, 84 COLUM. L. REV. 1897, 1916 n.55 (1984) ("[J]ustification in law should correlate more closely with moral appraisal."); Brian Hogan, *The* Dadson *Principle*, 1989 CRIM. L. REV. 679, 686 (exculpating a defendant based on an unknown justification seems "wholly bizarre"); A.P. Simester, *Mistakes in Defence*, 12 OXFORD J. LEGAL STUD. 295, 303–04 (1992) (suggesting that the objective approach may have the effect of "encouraging the commission of prima-facie offences in the hope that they might turn out to be fortuitously defensible").

6 *See* Andrew Ashworth, *Belief, Intent and Criminal Liability*, *in* OXFORD ESSAYS IN JURISPRUDENCE 1, 28–29 (John Eekelaar et al. eds., 1987).

7 Based on the Model Penal Code and German authorities, Fletcher asserts that "[t]he consensus of Western legal systems is that actors may avail themselves of justifications only if they act with a justificatory intent." FLETCHER, *supra* note 5, § 7.4, at 557. Robinson disagrees with Fletcher's interpretation of the Model Penal Code, *see* 2 ROBINSON, *supra* note 3, at 21 & n.18, and argues the American and English case law is equally divided. *See* 2 *id.* at 23. He further claims that English and Welsh criminal statutes would allow the defense. *See* Robinson, *supra* note 4, at 49.

This Article will argue that, in general, the unknowingly justified actor should *not* qualify for a justification defense. Thus it will reject the so-called objective theories of justification, as well as the class of punishment theories in which wrongdoing is the central organizing principle. While wrongdoing is the critical concept for a theory of criminal conduct, i.e., what acts should be prohibited or required, responsibility for causing harm, it will be argued, should be the basis of criminal punishment, i.e., under what conditions sanction should be imposed. Furthermore, between the remaining theories of justification that would hold the unknowingly justified actor liable, this Article will contend that the "subjective theory of justification" offers the best account of justification in general and the problem of unknowing justification in particular. Where a harm has been caused, punishment is always due unless the actor can show that he can be excused for causing the harm. Finally, although this Article supports the subjective theory of justification, it supports a version that differs from the Model Penal Code and other subjective justification theories. In particular, it advances a new mental state called "regarding" as the requisite for asserting a justification defense. Thus this Article's position, by virtue of either its result or reasoning, is novel.

This Article begins in Part I with a brief discussion of its methodology, in which the moral nature of the inquiry is explicated. In Part II, the principal theories of justification are briefly reviewed. In Part III, a range of arguments on both sides of the unknowing justification debate are examined and found wanting. Part IV presents an alternative view of unknowing justification. This position rests on a theory of punishment called the "harm theory." Part IV elaborates the harm theory, shows how it implies a rejection of the unknowing justification defense, and demonstrates how the theory favorably compares with its competitors.

I. A Few Words About Methodology

Three points about this Article's methodology are worth noting at this juncture. First, this Article will discuss the problem of unknowing justification as a question of morality (or ethics) set within the context of the criminal law. In particular, this Article generally proceeds within the framework of objective retributivism, the theory that considerations of desert are central to determining what sanctions should be imposed on people, and that objective factors, such as an act's consequences, are relevant to desert.[8] Accordingly, a variety of

8 Objective retributivism is elaborated and defended by a variety of contemporary scholars. *See, e.g.*, Fletcher, *supra* note 5, at § 6.6, 466–73; Lawrence Crocker,

contingent and pragmatic considerations not relevant to desert, such as public response to acquittals based on the unknowing justification defense and practical implications for code drafting, although interesting in their own right, are not pursued.[9] The answer to the purely moral question whether, all things equal, an unknowingly justified actor should be exonerated is of some intrinsic interest and surely bears powerfully upon what some may view as the "ultimate" issue: whether particular jurisdictions should revise their penal codes to add, drop or maintain an unknowing justification defense. Clearly the proper moral resolution of the issue must be taken into account with empirical and practical issues at some point in a legislature's deliberations.

Second, as a moral question, the problem of unknowing justification must at bottom be settled by appeal to moral intuitions. Sadly, moral intuitions are not distinct, clearly labeled psychological entities with special claims to truth or authority.[10] They are just your/my/our considered opinions about what is sound, decent, sensible, and just. Accordingly, the problem of unknowing justification in theory could be solved simply by focusing narrowly on an instance of unknowing justification and asking, "Is it sound, decent, sensible, and just to decrease or nullify the defendant's sentence here?" This Article's length may therefore seem suspect. Moral intuitions, however, are not so easily ascertained, nor so readily compartmentalized. True, there is no logical connection between the problem of unknowing justification and, for example, theories of self-defense, utilitarianism, causation, and attempts—all topics discussed herein. The latter theories do not entail a position on unknowing justification. But moral deliberation is more a matter of analogy than logic. Analogies may be distant but nonetheless useful in answering moral questions. Furthermore, the integration of moral intuitions into a coherent comprehensive theory is an independent goal recommending the interplay of disparate theo-

Justice in Criminal Liability: Decriminalizing Harmless Attempts, 53 OHIO ST. L.J. 1057 (1992); Michael Davis, *Why Attempts Deserve Less Punishment Than Complete Crimes*, 5 LAW & PHIL. 1 (1986); Michael S. Moore, *The Independent Moral Significance of Wrongdoing*, 5 J. CONTEMP. LEGAL ISSUES 237, 263–70 (1994). For a discussion of retributive theories of punishment in general, see R.A. Duff, *Penal Communications: Recent Work in the Philosophy of Punishment*, 20 CRIME & JUST. 1, 25–32 (1996).

9 For a thorough discussion of these issues, the interested reader should consult Robinson, *supra* note 4.

10 *But cf.* IMMANUEL KANT, CRITIQUE OF PRACTICAL REASON (Lewis White Beck ed. & trans., Macmillan Pub'g, 3d ed. 1993) (1788) (arguing that moral truth is accessible through reason's commitment to the categorical imperative); G.E. MOORE, PRINCIPIA ETHICA (Thomas Baldwin ed., 1959) (1903) (arguing that moral truth is a nonnatural quality that must be perceived via a special human faculty).

ries.[11] A plausible picture of our moral intuitions is that they are like nodes of a large web, each connecting to, supporting, and influencing some set of nearby intuitions, which in turn connect to, support and influence a more distant set.[12] Because the state of every node has a more or less attenuated effect on every other, the range of starting places for cogent analogy and argument on a particular moral question cannot be limited in advance. Thus, discussions of the problem of unknowing justification, like morality in general, may be as far ranging as the participants' ingenuity and patience. At a minimum, there is the project of reviewing, tracing, and debunking the alleged connections that others have drawn between the problem of unknowing justification and other issues of moral, political, and criminal theory.

Third, although this Article engages in a moral examination of justification, it is the legal, not the moral concept of justification, that is being examined. An examination of the moral concept of justification would elaborate under what conditions a person who prima facie deserves punishment should not be punished because she is "justified." In contrast, an examination of the legal concept of justification from a moral perspective considers under what conditions a person who has satisfied the definitional elements of an offense should be able to assert a justification defense.[13] Such an account is general in two senses. First, it applies to all justification defenses without regard to whether any particular justification defense is morally correct. A theory of legal justification should permit agreement on the general structure of a justification defense, such as defense of property, even if, for example, it is disputed whether one should be permitted to use deadly force to protect one's property. Second, an account of legal justification should be general in the sense that it applies to all offenses, regardless of whether those offenses, for example drug use, are

11 *Cf.* RONALD DWORKIN, LAW'S EMPIRE 176–224 (1986) (arguing that integrity is an independent virtue for legal systems).

12 *Cf.* W.V. QUINE, WEB OF BELIEF 9–12 (1970) (describing how our interlocking set of beliefs are assessed in light of each other); JOHN RAWLS, A THEORY OF JUSTICE 48 (1971) (discussing how considered judgments and regulative principles are assessed in light of each other).

13 The definitional elements of an offense specify (1) the required mental state, (2) the prohibited conduct, and (3) the attendant circumstances of an offense—for example, (1) knowingly (2) summoning the fire department (3) in the absence of a fire or other emergency. The definitional elements do not contain the absence of those conditions that would establish an affirmative defense to the offense, however those conditions might be defined. While the absence of such conditions may formally be a material element of the offense, *see* MODEL PENAL CODE § 1.13(10)(ii) (1962), their absence is best considered an extra-definitional material element.

morally valid. In sum, a theory of legal justification is a moral inquiry into the general structure of justification defenses as they may apply to the general range of offenses and defenses.

II. THREE THEORIES OF JUSTIFICATION

This Part briefly introduces the major theories of justification and indicates how they would resolve the problem of unknowing justification. An understanding of these theories and their terminology is important for the ensuing discussion.

According to the first theory—the "objective theory of justification"—persons who commit offenses should be held legally justified and not liable where their conduct has been determined to be permissible (or preferable).[14] Whether conduct is permissible is generally not a function of the mental state of the actor, which is instead relevant to the actor's culpability for the conduct, but is purely a function of "objective" factors, such as whether the conduct was consented to or advanced a significant social interest. Thus, under the objective theory, a person's mental state will be irrelevant to whether the person is legally justified. For example, if *A* used force against *B* and so prevented *B* from injuring *C*, *A* would be legally justified, and held not liable for assault, in both the cases where *A* knew of *B*'s contemplated attack and *A* did not know of *B*'s contemplated attack. By definition, the objective theory of justification would render the

14 Theorists differ on whether justified conduct is merely conduct that should be engaged in or whether it includes conduct that may, but need not, be engaged in. *See* Greenawalt, *supra* note 5, at 1904–07 (discussing problematic cases involving acceptable but not ideal behavior); Heidi M. Hurd, *Justification and Excuse, Wrongdoing and Culpability*, 74 NOTRE DAME L. REV. 1551, 1560 n.25 (1999) (elaborating on a range of possible positions available on the question). For example, if a run-away trolley car is headed toward *A* and could be rerouted to head toward *B* creating the same danger to human life, from a purely utilitarian perspective, all other things equal, rerouting the car is neither required nor prohibited—it is simply permissible. More generous theories of justification would hold that rerouting the car is justified because the alternative of no action is not morally preferable; more restrictive ones would hold that it is unjustified because it is not morally preferable to the alternative of no action. The Model Penal Code appears to adopt the latter version by limiting the necessity defense to cases where the act at issue avoids a "greater" evil than the one caused. MODEL PENAL CODE § 3.02(1)(a) (1962). None of the arguments in this Article depend on whether a generous or a restrictive version of justification is correct. For ease of exposition, I will refer to justified conduct as conduct that is "permissible," leaving open the question whether in order for it to be permissible it also must be, in some sense, preferable.

unknowingly justified actor not liable. In the United States, Professor Paul Robinson has been a major proponent of this theory.[15]

According to the second theory of justification—the "subjective theory"—persons satisfying the definitional elements of an offense should be held legally justified and not liable based on their mental state only. Depending on the particular form of the subjective theory at issue, the mental state of the actor establishing justification will be either believing, knowing, hoping, intending, or desiring that the conduct at issue is of a type which is, in fact, socially desirable or permissible. Thus, for example, if *A* used force against *B* believing such force was needed to prevent *B* from injuring *C*, *A* would be considered legally justified, and held not liable for assault, in both the cases where *B* was about to attack *C* and where *B* was not about to attack *C*. By definition, under the subjective theory, the unknowingly justified actor, lacking the appropriate subjective state, would not qualify for a defense and would be held liable. The Model Penal Code, construed literally, adopts such a theory of justification.[16]

According to the final theory—the "mixed theory of justification"—in order to be held legally justified, and hence not liable, persons satisfying the definitional elements of an offense must both engage in socially permissible conduct *and* believe that they are so engaged (or possess the appropriate exculpatory mental state, however defined). *A* would be considered justified in using force against *B* only if *A* was correct in believing that his use of force would have a socially desirable effect. Regarding the unknowingly justified actor, the mixed theory would follow the subjective theory of justification in holding her not legally justified and therefore liable on the ground that she lacked the proper subjective state. The mixed theory and the subjective theory, however, diverge with respect to the "mistakenly justified actor." Where a person incorrectly believes that he is acting in a

15 *See* 2 ROBINSON, *supra* note 3, § 122; Paul H. Robinson, *A Theory of Justification: Societal Harm as a Prerequisite for Criminal Liability*, 23 UCLA L. REV. 266 (1975) Robinson, *supra* note 4.

16 Sections 3.02–09 of the Model Penal Code consistently employ subjective terminology, for example, "the use of force . . . is justifiable when the actor believes . . . ." MODEL PENAL CODE § 3.04(1) (1962); *see also* 1 MODEL PENAL CODE AND COMMENTARIES §§ 3.01 to 5.07 at 11, (1985) ("If a druggist who sells a drug without a prescription is unaware that the recipient requires it immediately to save his life, the actual necessity of the transaction will not exculpate the druggist."). Although the Model Penal Code justification chapter contains a definitional provision employing objective terminology, *see id.* § 3.11(1) ("force . . . not amounting to a privilege"), the defined term "unlawful force" is incorporated within the Code's generally subjective provisions. Thus, for a defendant to establish a justification, the Model Penal Code requires simply a subjective belief in an objective condition.

manner that is socially permissible, the subjective theory would find the actor justified; in contrast, the mixed theory would describe the person as either putatively justified or excused. Professor George Fletcher is a major American proponent of the mixed theory of justification.[17]

III. CURRENT ARGUMENTS ABOUT UNKNOWING JUSTIFICATION

This Part of the Article considers six leading arguments concerning the problem of unknowing justification and attempts to expose the shortcomings of each.

A. The Definitional Argument

A threshold argument that must be considered, and dismissed, is based on the definitions of such terms as "justified," "permissible," and "excused." Although this argument is primarily addressed against the subjective theory of justification, and I shall speak of it in such terms, it also has some application against the mixed theory of justification.

As discussed above, the subjective theory of justification rejects the possibility of unknowing justification. Yet it may be contended that, in this respect, the subjective theory is erroneous, conceptually confused, or at least at odds with accepted usage on three grounds. The first ground is that the subjective theory asserts oxymoronically that an actor who is justified unknowingly is not justified at all. An actor, it might be added, cannot be both justified and not justified at all. To avoid this apparent difficulty, the subjective theory should be thought of as asserting more precisely that the actor whose *conduct* is unknowingly permissible may not establish that he is thereby entitled to the so-called legal defenses of justification, such as self-defense and necessity. The subjective theory of justification thus distinguishes between the legal status of conduct (improper vs. permissible) and the criminal status of the defendant (liable vs. legally justified). Under the subjective theory, an actor may engage in legally justified conduct but not be entitled to a justification defense.

A second ground for describing the subjective theory as conceptually confused rests on a semantic stipulation. Writing in another context, Douglas Husak asserts, "*By definition,* no instance of permissi-

17 *See* GEORGE P. FLETCHER, BASIC CONCEPTS OF CRIMINAL LAW 101–06 (1998); FLETCHER, *supra* note 5, § 7.4; George P. Fletcher, *The Right Deed for the Wrong Reason: A Reply to Mr. Robinson*, 23 UCLA L. REV. 293 (1975).

ble conduct should be subject to punishment."[18] Because unknowingly justified actors engage in permissible conduct, Husak's definition implies that the subjective theory, holding that such actors should be subject to punishment, is erroneous.

Husak's definition should be set aside for the purpose of investigating the problem of unknowing justification. Definitions can never settle substantive disagreements nor should they force one party to state her position in a manner requiring substantial redefinition of her framework. Adopting Husak's definition would force the proponent of the subjective theory, who believes the unknowingly justified actor should be subject to punishment for her conduct, to characterize such conduct as impermissible based in part on the actor's subjective state. But this would be inconsistent with accepted usage according to which subjective states, such as awareness of the consequences of a person's act, bear on the actor's culpability for the act, not the wrongfulness or permissibility of the act itself.[19] The proposition that "no instance of permissible conduct should be punished" (which is denied by the subjective theory of justification) should be a matter of substantive debate, not of definition.

Finally, the subjective theory may be thought to characterize erroneously as justifications what are by definition excuses. The subjective theory would allow justification defenses to be established based on the defendant's beliefs about his conduct. According to Husak, most criminal law theorists accept as a matter of definition that "[j]ustifications are defenses that arise from properties or characteristics of acts; excuses are defenses that arise from properties or characteristics of actors."[20] Under the plausible assumption that subjective states are properties of actors, not acts, the subjective theory of justification thus would systematically mislabel justifications as excuses. Again, for the purpose of this discussion, this Article rejects Husak's definition. In general, definitions should maximize agreement and isolate differences, rather than foist an inapt vocabulary on one of the parties to the disagreement. Whether legal justifications should be a matter of objective and/or subjective facts should be left an open question at the outset.

To summarize, in order for the problem of unknowing justification to be discussed, the meaning of some terms must be fixed at the

18 Douglas N. Husak, *Justifications and the Criminal Liability of Accessories*, 80 J. CRIM. L. & CRIMINOLOGY 491, 500 (1989) (emphasis added).

19 *See* MICHAEL S. MOORE, PLACING BLAME 406 (1997) (explaining why mental states such as emotions are not relevant to wrongdoing).

20 Husak, *supra* note 18, at 496.

outset and others must be resolved through inquiry. By "justified conduct," this Article refers to conduct that possesses the necessary objective features to qualify for being legally permissible. Whether such conduct, regardless of the actor's subjective state, morally should always provide actors a legal defense and whether only defenses based on such conduct alone should be classified as justifications are open questions.

B. The Logical Consistency Argument

Russell Christopher has recently advanced a novel and powerful argument concerning the problem of unknowing justification.[21] Like the definitional argument discussed above, Christopher's argument concerns the logical consistency of a theory of justification.[22] Unlike the definitional argument, however, Christopher's argument is directed against the objective theory of justification—the theory that actors should be relieved from liability based solely on the objective features of their conduct.[23] Furthermore, unlike the definitional argument, Christopher's argument is not based on a superficial conflict of the definitional terms. Rather, Christopher's argument purports to demonstrate that the objective theory contains a deep, uncorrectable, and unnoticed structural flaw.[24] According to Christopher, this structural flaw is manifested in the form of a paradox when the objective theory is applied to certain cases of unknowing justified conduct.[25] If Christopher's argument is sound, the objective theory of justification, as well as the mixed theory perhaps,[26] would have to be discarded in favor of the subjective theory of justification. Accordingly, the unknowingly justified actor would be fully liable.

---

21 *See generally* Russell L. Christopher, *Mistake of Fact in the Objective Theory of Justification: Do Two Rights Make Two Wrongs Make Two Rights . . . ?*, 85 J. CRIM L. & CRIMINOLOGY 295 (1994) [hereinafter Christopher, *Mistake of Fact*] (arguing that objective theories of unknowing justification should be modified to reflect a reasonable standard); Russell Christopher, *Self-Defense and Defense of Others*, 27 PHIL. & PUB. AFF. 123 (1998) (applying the *Dadson* principle to the paradox of unknowing justification may resolve inherent contradictions); Christopher, *supra* note 5 (challenging the logical consistency of Fletcher's objective justification defense).

22 *See* Christopher, *Mistake of Fact*, *supra* note 21, *passim.*

23 *See id.* at 320–30.

24 *See id.*

25 *See id.* at 298.

26 Christopher expands the scope of his argument to include mixed theories. *See id.* at 295. Christopher contends mixed theories, insofar as they contain an objective component, are open to the same criticisms that undermine the objective theory. *See id.*

1. The Paradox

Christopher's argument for the inconsistency of the objective theory is based on the following hypothetical:

> Suppose that D, unaware of V's imminent attack on D, attacks V out of spite. V, unaware of D's imminent attack, attacks D out of spite. D and V each use similarly harmful force against the other at the same time.[27]

Although Christopher does not color in the body of the hypothetical, we might think of two people drinking at a bar. As they drink, an argument develops and escalates. Carried away in anger, they simultaneously swing and land blows to the side of the other's head. In initiating his punch, neither knew the other was doing the same.

In developing the paradox, Christopher's first step is to argue that under an objective theory of justification, *D* appears legally justified, and hence not criminally liable.[28] Although *D* was unaware of *V*'s imminent attack, the objective fact of *V*'s imminent attack apparently renders *D* "unknowingly" justified in using force against *V*. Certainly if *D* knew of *V*'s imminent attack and sought to defend himself, his use of force in self-defense would have been justified. The objective theory, in essence, treats this absence of knowledge as irrelevant.[29] So *D*'s use of force against *V* seems like it should be legally justified. Next, Christopher points out, *D*'s and *V*'s positions are symmetrical.[30] *D*'s and *V*'s actions toward each other and beliefs about each other are, per hypothesis, identical. What goes for one legally must go for the other logically. Under the objective theory, therefore, it would seem that *V* is also justified in his use of force, albeit unknowingly, just as *D* is.

The penultimate step of Christopher's argument is to show the incoherence of this conclusion. Christopher asks, how can *D* be justified in using force against *V*, based on *V*'s imminent attack, given that *V*'s attack is, as we have seen, itself apparently justified?[31] As a general matter, it seems inappropriate to allow the use of force against a justified use of force. For example, the explanation for why an accomplice cannot use force to prevent a police officer from subduing his partner in crime is that the police officer's use of force was justified, and the defense of third parties cannot be triggered by the justified use of force. Fletcher has advanced the "one Right thesis," which asserts that

27 Christopher, *supra* note 5, at 241.

28 *See id.*

29 *See supra* notes 14–15 and accompanying text.

30 *See* Christopher, *supra* note 5, at 242.

31 *See id.*

in any situation of zero-sum conflict between two actors, only one can be justified.[32] Intuitively, the claim is plausible because if justification requires objectively right conduct, in any conflict it is impossible that both sides be objectively right. Robinson, among others, supports the thesis.[33] Consistently with these theorists, positive law restricts the range of attacks that force can justifiably be used against to those that are either unlawful or at least believed to be unlawful.[34]

For the final step in Christopher's analysis, we must return to the initial hypothetical. If *D* and *V* cannot both be justified in their use of force, perhaps they are both unjustified. But if *V*'s use of force is unjustified, how can *D*'s use of force against *V* also be unjustified? Generally speaking, preventing unjustified force is justified. Under objective theories of justification, *D*'s and *V*'s motivations are irrelevant. Assuming *V*'s force is unjustified seems to entail that *D*'s is justified. But obviously the conclusion that one actor's use of force is unjustified but the other's is not contradicts the basic factual symmetry of the initial hypothetical.[35]

It seems then that under an objective theory of justification, both *D* and *V* can neither be justified nor unjustified. Yet this is paradoxical for it seems that those who engage in prima facie illegal conduct must be either legally justified or not. The source of this paradox, Christopher concludes, is the objective theory of justification.[36] Thus, the theory must be rejected (along with the possibility of unknowing justification which only it permits).[37]

An analogy may help to illustrate the paradox that Christopher asserts plagues the objective theory of justification. Christopher's hypothetical generates a paradox by combining three essential elements. First, the hypothetical involves two actors who appear to be placed

32 *See* George P. Fletcher, *The Psychotic Aggressor—A Generation Later*, 27 ISRAEL L. REV. 227, 236–37 (1993); *see also* George P. Fletcher, *The Right and the Reasonable*, 98 HARV. L. REV. 949, 978 (1985) (stating that incompatible justifications are "logically impossible" under an objective theory).

33 *See* 1 ROBINSON, *supra* note 3, at 165 ("Where an aggressor has a justification defense, the proper rule is clear: justified aggression should never be lawfully subject to resistance or interference." (footnote omitted)).

34 Self-defense, as the doctrine is traditionally formulated, is limited to the use of force to prevent "unlawful bodily harm." 1 WAYNE R. LAFAVE & AUSTIN W. SCOTT, JR., SUBSTANTIVE CRIMINAL LAW 649 (1986); *see* MODEL PENAL CODE §§ 3.04, 3.11 (1962) (permitting the use of force against that believed to be threatening and unlawful, and defining unlawful force, in part, as not privileged).

35 *See* Christopher, *supra* note 5, at 243–44.

36 *See id.* at 245–47.

37 *See id.* at 250–51; *see also* Christopher, *Mistake of Fact*, *supra* note 21, at 310–12, 320–25.

symmetrically with respect to all features relevant to the objective theory of justification. Thus, both must be either justified or unjustified according to the objective theory of justification. Second, the situation is one of mutual dependency, that is, the status of each use of force as justified or not depends on the status of the other use of force. Third, the dependency of each use of force is inverse; that is, if one use of force is justified, the other must be unjustified, and if one use of force is unjustified, the other must be justified. The situation may be likened to that of the two sentences below:

The sentence to the right is false. The sentence to the left is false.

First, by symmetry either they are both true or both false. Second, the sentences are mutually dependent. Because the sentences are about each other's truth, the truth of one sentence is based on the truth of the other. Third, they depend inversely on one another in the sense that if one statement is true, the other, asserting the first's falseness, must be false, and vice-versa. Thus, if they are both true, they are accordingly both false, or if they are both false, they must, contrary to hypothesis, both be true—a paradox. If the objective theory of justification applied to situations of unknowing justification results in a characterization of the situation structurally similar to the sentences above, the theory must be defective. A minimum requirement for a theory of justification is that it is able to determine whether an actor is justified or not. If the objective theory of justification is defective, the unknowingly justified actor cannot claim a legal justification and must be punished. In this manner, Christopher claims the ambitious goal of solving the apparently moral dilemma of unknowing justification through logical analysis alone.[38]

38 Christopher's paradox is based on an example involving potential claims of self-defense. The paradox, however, arguably infects other defenses as well. For a fee, Dennis will take persons to hunt eagles, a protected species. One day, Dennis takes Otto hunting. Dennis is unaware that Otto is an undercover police officer. In order to maintain his cover and advance the investigation, Otto shoots and kills an eagle. Dennis is charged with the killing of the eagle based on a theory of accomplice liability which imputes the act to him. Otto of course would not be liable for illegally killing the eagle based on the defense of law-enforcement. May Dennis take advantage of the law-enforcement defense that shields Otto? From a purely objective approach to justification, it may seem that there is no wrongful act to impute to Dennis because the killing was an instance of justified law-enforcement conduct. But if Dennis cannot be convicted, then it seems that the killing was not objectively justified because it, in fact, did not lead to Dennis's conviction. But if it was not objectively justified (and since Dennis, unlike Otto, cannot assert an excuse or subjective justification), Otto should be convicted—a paradox!

2. Paradox Lost

Through careful distinctions, the pinch of Christopher's paradox may be avoided. If *D* and *V* use similar force in attacking each other, there are two possibilities: the force each uses can either (1) have no effect on the force used by the other and so not at all prevent the force used by the other, or (2) partially prevent the force used by the other. Obviously, the force each uses cannot wholly prevent the other force because, per hypothesis, the two similar forces actually occur.

Let us examine both cases of no-effect and partial prevention. If the forces have no effect on each other, neither force is justified. Imagine that the attack that *D* and *V* plan to launch on each other consists of a punch to the head followed by a kick to the shins a moment later. *D* and *V* punch each other simultaneously and, notwithstanding the punch they have received, deliver the kick exactly as they had planned. Whether *D*'s and *V*'s use of force is justified, of course, depends on what it is for force to be justified. Without canvassing all substantive theories of justification in general, or self-defense in particular,[39] it can generally be said that justified action either has some overriding social utility or, at a minimum, violates nobody's rights. *D*'s and *V*'s attacks clearly lacked social utility. They were not effective in preventing *D* or *V* from being injured. From a utilitarian perspective, they merely caused gratuitous pain. Furthermore, the attacks violated *D*'s and *V*'s general rights to bodily integrity. A person engaging in an unjustified attack arguably forfeits certain moral rights to bodily integrity. That forfeiture, however, is not absolute but limited to the right not to be subjected to force that would prevent an unjustified attack. The essentially preventative nature of justified force is demonstrated by the general requirement that the force employed in self-defense not exceed that which is necessary.[40] Under any consistently objective theory of justification, force is necessary only if it in fact contributes to the goal of self-defense. The force used by *D* and *V* did not. Having no useful effect and violating the rights of the persons it

39 For a survey of the theories explaining why self-defense is justified, see generally SUZANNE UNIACKE, PERMISSIBLE KILLING (1994); Stuart P. Green, *Castles and Carjackers: Proportionality and the Use of Deadly Force in Defense of Dwellings and Vehicles*, 1999 U. ILL. L. REV. 1; and Nancy M. Omichinski, Comment, *Applying the Theories of Justifiable Homicide to Conflicts in the Doctrine of Self-Defense*, 33 WAYNE L. REV. 1447, 1447–53 (1987).

40 *See* 1 LAFAVE & SCOTT, *supra* note 34, at 649 (stating that permissible force is that which is "necessary to avoid th[e] danger"); *see also* UNIACKE, *supra* note 39, at 213 ("[I]f you use force beyond what is necessary to resist, repel, or ward me off . . . , then you wrong me.").

was used against, the force could not be justified in the objective sense of the term (although the force might well have been excused).[41]

Consider now the case where the forces at issue partially prevent each other. All cases of partially prevented force can be analyzed as cases involving a force with two parts: an actual part that was not prevented and a hypothetical part that was prevented. Imagine that *D* and *V* again plan to launch limited attacks on each other consisting of a punch to the head followed by a kick to the shins a moment later. *D* and *V* simultaneously punch each other and slightly stunned, halt their partially completed attacks. Here the punches were justified. They prevented the other person from kicking. Furthermore, the kicks, if they had occurred, would clearly have been unjustified. They could not have done any good because there was no further attack for the kicks to prevent. The kicks, if they had occurred, would have been analogous to the ineffective punches considered earlier. Paradoxes only appear to arise when it is insisted that the contemplated attack, defined as a combined punch and kick, must be wholly justified or unjustified. No act or series of acts can be both wholly justified and unjustified. As the analysis shows, however, simultaneous partially successful attacks have conceptually severable aspects: that which occurs and is justified and that which is prevented and which, if it had occurred, would have been unjustified. There is no reason not to engage in this conceptual severance as a device to avoid paradox. Furthermore, because the only part of the attack that actually occurred was the punch, that is the only part of the attack the status of which as justified or unjustified is at issue.[42] It is justified, and so, if

41 With respect to cases where force has no preventative effect, another argument is open to Christopher. Besides forfeiting a right not to have defensive force used, an unjustified attacker also forfeits his right not to suffer punishment for his unjustified use of force. Thus, it could be argued that *D*'s and *V*'s uses of force were potentially justified as punitive measures against the other's use of force (unless of course, the other's use of force was itself a justified punitive measure). In most cases, however, private citizens are not legitimate agents of punishment. Private punitive force, at least in a society having criminal laws, is improper vigilantism. Furthermore, if *D*'s and *V*'s use of force is understood as potentially justified punitive acts, their status as justified or not becomes moot. If they were justified, the state would have no interest in punishment, and if unjustified, the state would also have no interest because punishment was already imposed. Finally, if force is not imposed for the right reason, it arguably cannot qualify as punitive. A fleeing criminal struck by lightning cannot avoid further punishment on the ground that being hit by lightning was a punitive measure.

42 Some cases are more conceptually difficult to divide into prevented and not prevented parts. *A* and *B* each plan to strangle the other to death. They simultaneous put their hands around the other's neck and squeeze, strangling each other into unconsciousness. The analysis is the same as the text example. Both acts of strangula-

there are no remaining requirements that are not met, the actor would be entitled to a justification defense.

At bottom, the reason that Christopher's paradox fails is that the requirement of mutual dependency identified in the previous section is not met. The status of an actual use of force as justified or not depends on the status of the force that it prevented. The status of this prevented force, however, does not depend on the status of the force that was actually used to prevent it. Rather, the status of the prevented force depends on the status of the hypothetical force, if any, that it would have prevented if it (the prevented force) counterfactually had occurred. (For example, *D*'s punch to *V*'s head was justified because the force it prevented, *V*'s kick to *D*'s shin was unjustified. *V*'s kick to *D*'s shin, if it had occurred, would have been unjustified because the only act it would have prevented, say, *D*'s clutching his head in pain, would have been justified.) The status of this hypothetical force, likewise, is not a function of the force actually used to prevent it, but a function of the sur-hypothetical force, if any, that it would have prevented. In this way, in order to determine whether force is justified or not, a series of counterfactual scenarios must be *unraveled* until an initial unjustified use of force is identified.

3. The Recursive Structure of Justification

Understood in this light, "an unjustified attack" pursuant to the objective theory may be coherently defined as "an attack that does not prevent an unjustified attack." Notice that this definition incorporates the term "unjustified attack"—the very term to be defined. Nevertheless, the definition is noncircular and unambiguously divides all attacks into unjustified and justified ones, thus avoiding Christopher's objections. This point may be demonstrated by analogy. The definition of "an unjustified attack" as "an attack that does not prevent an unjustified attack" is strictly analogous to the definition of "an odd positive number" as "a positive number that does not immediately follow an odd positive number." This definition noncircularly differentiates odd positive numbers from even positive numbers despite employing an "odd positive number" in the definition. Observe:

> 1 is an odd positive number because it is a positive number and the number it immediately follows, 0, being nonpositive, is not an odd positive number;

---

tion were justified. They each prevented further strangulation causing death. Strangulation to death would have been unjustified because strangulation into unconsciousness was sufficient to prevent the unjustified strangulation to death.

> 2 is not an odd positive number because it immediately follows 1, an odd positive number (by prior step), rather than a number that is not an odd positive number;
> 3 is an odd positive number because it is a positive number, and the number it immediately follows, 2, is not an odd positive number (by prior step); etc.

The definition is noncircular and stable because it includes a concept—"immediately follows"—that has a fixed endpoint (1 does not immediately follow any positive number). In mathematics, such a definition employing a fixed endpoint and an iterative rule for successive terms is called a "recursive" definition.[43] As discussed above, two attacks cannot prevent each other. Furthermore, an infinite regress of preventative attacks (*A* prevented *B*, which would have prevented *C*, which would have prevented *D*, *ad infinitum*) is impossible. Therefore, any chain of preventative attacks must have a fixed endpoint—an attack which would be unproblematically unjustified because it would not have prevented any other attacks. All other potential attacks by that person would, through the operation of the definition, also be characterized—determinately and unambiguously—as unjustified. Justification in the criminal law is thus saved from fatal circularity by its recursive structure.

In sum, despite Christopher's ingenious arguments, the objective theory is at least logically coherent enough to provide a basis for exonerating the unknowingly justified actor.

B. Special Treatment Argument

In the course of his writings, Fletcher advances a number of arguments in favor of his mixed theory of justification, the theory that for an act to be justified it must be both objectively justified and so understood by the actor. In *Rethinking Criminal Law*, the argument for requiring some subjective state on which Fletcher seems to place greatest weight concerns justifications' status as exceptions. According to Fletcher, justifications "represent exceptions to prohibitory norms."[44] "As exceptions," Fletcher writes, "these claims should be available only to those who merit special treatment."[45] Because merit

43 *See* ENCYCLOPEDIA OF COMPUTER SCIENCE AND ENGINEERING 1273–74 (Anthony Ralston ed., 2d ed. 1983); *see also* NIGEL CUTLAND, COMPUTABILITY: AN INTRODUCTION TO RECURSIVE FUNCTION THEORY 32 (1980) (discussing noncircularity of recursive definitions).

44 FLETCHER, *supra* note 5, § 7.4, at 565.

45 *Id.*; *see also* FLETCHER, *supra* note 17, at 106.

implies a personal worthiness, Fletcher concludes that the requirement of merit makes the actor's subjective state relevant.[46]

1. Unclear Mental State

One critique of Fletcher's position is presented by Professor Robinson. According to Robinson, Fletcher's position must be rejected because it is hopelessly unclear what type of subjective state would be required to "merit" a justification defense.[47] In order to invoke the defense, Robinson presses, must the defendant merely know of the existence of the justifying circumstances, or must he act with justificatory intent (the intent that the act produce the result underlying the justification)?[48] For example, is it sufficient that the defendant merely knew he would save a third party from an aggressor, or must he in acting have intended to save the third party? If knowledge is sufficient, one might further ask, should acting with a somewhat lesser degree of confidence, such as "strong suspicion," also suffice? If the actor must have intended the result, is intent a matter of ultimate motive or mere purpose?[49] If motive, is a but-for motive or merely a contributing motive needed? These questions may stymie the proponent of a subjective or mixed view of justification. The lack of an argument specifying the precise mental state or states required would seem to undercut the claim that a mental state is required in the first place.

There appears to be a ready response to Robinson's questions about mental state. In order to establish a claim of justification under a mixed theory of justification, the proponent of the mixed or subjective theory should require whatever mental state would suffice to establish a defense in cases of putative justification. (A putative justification is generally defined as an exculpatory mistake as to the existence of justificatory circumstances.) *A* shoots *B*, who clearly appears to be attacking him. *B*, however, is not attacking *A*. What mental state does *A* need to establish self-defense as a putative justification? Does *A* need to shoot for the purpose of saving his own life?

46 *See* FLETCHER, *supra* note 5, § 7.4, at 565–66.

47 2 ROBINSON, *supra* note 3, at 18 ("Ultimately, there seems no clear answer to what is to be required by a justificatory purpose requirement.").

48 2 *id.* at 17.

49 The distinction between motivation and purpose is a familiar one in criminal law. Motivations underlie purposes, although they sometimes are identical in content. For example, in purposely saving another's life, one may be motivated either by the saving of another's life or by the reward that is expected to follow from the saving of another's life. *See* Anthony M. Dillof, *Punishing Bias: An Examination of the Theoretical Foundations of Bias Crime Statutes*, 91 Nw. U. L. REV. 1015, 1065–67 (1997).

Must this be *A*'s sole or primary motivation? If *A*'s belief that *B* is attacking will suffice, with what degree of confidence must this belief be held? Are the answers to these questions affected if *A*'s purpose is not to shoot *B* but *A*'s shooting of *B* is a side effect of *A*'s true goal? These questions turn upon subtle and interesting issues of culpability.[50] Yet even those who hold an objective theory of justification, such as Robinson, need an answer to these questions, for nobody wishes to deny the validity of putative justification. The required mental state for putative justification will likely be identical to the mental state required for actual justification under a mixed or subjective theory. Clearly, the mental state needed for a mixed theory of justification should be no more restrictive than that needed for putative justification; that would unfairly penalize a defendant who had the mental state for putative justification, but not actual justification under the mixed theory, on the basis of the fortunate occurrence of justificatory circumstances. The other half of the biconditional is a little less clear. Might a mental state unable to support a claim for putative justification suffice to establish the moral merit needed for a justification defense? While no compelling argument can be given in answer to this question, Ockham's razor[51] suggests that there should not be two distinct types of merit: that sufficient for a putative justification and that sufficient for true justification. What will do for putative justification seems adequate for true justification. The lack of an independent theory of merit for true justification is not a strong objection to Fletcher's mixed theory of justification.

2. Further Argument Required

A second argument advanced by Robinson is more convincing. Robinson objects to Fletcher's requirement of merit on the ground that it is conclusory. Robinson asks, "[w]hy cannot one as properly say

50 *See, e.g.*, R.A. DUFF, INTENTION, AGENCY & CRIMINAL LIABILITY: PHILOSOPHY OF ACTION AND THE CRIMINAL LAW 109–10 (1990) (reviewing the "consequentialist" view of criminal agency); Gerald Dworkin, *Intention, Foreseeability, and Responsibility*, *in* RESPONSIBILITY, CHARACTER, AND THE EMOTIONS 338, 339–47 (Ferdinand Schoeman ed., 1987) (arguing there is a moral distinction between direct intent and indirect intent); H.L.A. Hart, *Intention and Punishment*, *in* PUNISHMENT AND RESPONSIBILITY: ESSAYS IN THE PHILOSOPHY OF LAW 113, 122–25 (1968) (criticizing doctrine of double effect as a basis for making moral judgments); Warren S. Quinn, *Actions, Intentions, and Consequences: The Doctrine of Double Effect*, 18 PHIL. & PUB. AFF. 334, 335–41 (1989) (reviewing the doctrine of double effect).

51 Named for the philosopher William of Ockham, Ockham's razor refers to the principle that "entities should not be multiplied unnecessarily." THE CAMBRIDGE DICTIONARY OF PHILOSOPHY 545 (1995).

that the privilege to violate a norm is available to all those who by doing so in fact avoid a greater harm or evil than that sought to be prevented by the prohibitory norm?"[52] Just as royalty may acquire privilege through the fortunate accident of being born to noble parents, Robinson might argue, so would-be criminals may acquire the privilege to engage in prima facie criminal conduct through the fortunate accident of acting in justificatory circumstances. Whatever special treatment is claimed by the unknowingly justified actor can be claimed based on the special circumstances of the actor's conduct. Furthermore, to the extent that lack of personal merit is a morally relevant factor, the criminal law can respond to this lack of merit by convicting the unknowingly justified actor for attempting the offense he believed he was committing. At least under subjective theories of attempt liability, an actor's lack of personal merit manifested by his intent to engage in a wrongful act is a ground for some punishment. At a minimum, Fletcher should provide additional argument to support his requirement of merit.

C. Conflicting Norm Argument

Fletcher offers a second, more interesting argument in support of his view that an appropriate subjective state is needed to supplement justified conduct in order for an actor to qualify for a justification defense. The special treatment argument characterized justifications as mere exceptions to prohibitory norms.[53] Justifications seemed merely to express the limits of prohibitory norms. Prohibitory norms were thus implicitly ontologically prior and primary in the way that there can be rules without exceptions, but not exceptions without rules. In contrast, the conflicting norm argument characterizes justificatory norms as ontologically equal and substantively superior to prohibitory norms. Fletcher writes that

> the logic of justification is more complicated than simply tagging on negative elements to the definition of the offense. Claims of justification represent conflicting norms that collide with the prohibition of the offense and under circumstances prevail over the prohibition. For example, the commandment to observe the Sabbath conflicts with the imperative to protect human life, and the latter will typically prevail.[54]

Fletcher thus perceives a symmetry between prohibitory norms, such as do not kill, which are codified as criminal offenses, and justifi-

52 2 ROBINSON, *supra* note 3, at 26.

53 *See supra* note 44 and accompanying text.

54 FLETCHER, *supra* note 17, at 104.

catory norms, such as prevent unlawful killings, which are codified as justification defenses.[55] This symmetry, Fletcher contends, extends to justificatory norms having objective and subjective components as prohibitory norms do.[56] Just as a prosecutor may not charge a criminal offense unless the actor has some mens rea with respect to prohibited conduct, so too a defendant may not raise a justification defense unless she has the appropriate subjective state with respect to the justificatory circumstances. Of course, unlike prohibitory norms, justificatory norms are permissive, not mandatory, in form. The criminal law generally does not require, but only permits, that force be used to prevent unlawful aggression.[57] Nevertheless, the permissive nature of justificatory norms has never been thought to be key in understanding whether a subjective component is necessary to invoke them as defenses. If justificatory norms were mandatory, the same issue whether they could be fulfilled unknowingly would arise.

Robinson has criticized Fletcher's mixed theory of justification on the ground that it is an unprincipled amalgam of objective and subjective theories of justification that does not rest on a general theory of liability.[58] This criticism may be elaborated in the following way: it is unclear what the moral significance of knowingly engaging in justified action could have for a criminal justice system. In cases of putative justification, an actor mistakenly believes that he is acting in a justified manner. Where this belief is reasonable, all agree that the actor should not be liable. The reason is that the actor is not culpable for his wrongful conduct by virtue of his commendable subjective state. Where the actor reasonably *and correctly* believes that his conduct is justified, again, all agree that the actor should not be liable. But why? Proponents of the subjective theory of justification believe the reason is the same as for the putatively justified actor, to wit, no moral culpability by virtue of his commendable moral state. Proponents of the objective theory believe the primary reason is that no wrongful conduct has occurred and a sufficient, albeit secondary, reason is that the

55 *See id.* at 105.

56 *See id.* at 105–06.

57 Although justifications are permissive in form, there are limited factual settings where engaging in conduct falling within an offense definition may be mandatory. If *A* sets fire to a warehouse and *B*, a security guard suffering burns, runs out, *A* may be required to steal a car to take *B* to the hospital, assuming no other means of helping *B* are available.

58 *See* Robinson, *supra* note 4, at 68 ("[The mixed theory of justification] seems internally inconsistent in its view on the significance of resulting harm and on the sufficiency of culpability for full liability.").

actor is not morally culpable.[59] If asked the content of this claim that there are two sufficient reasons for exoneration, proponents of the objective theory could coherently respond with the counterfactual assertion that even if the actor had acted for a malicious purpose, she should still be exonerated. In the case of the knowingly justified actor, proponents of the mixed theory reject as an independent reason for exoneration that the conduct is not wrongful (the objective theory position). To distinguish themselves from subjective theorists, however, they must find that the fact that the conduct was objectively justified is somehow morally relevant. But it is unclear what the content of this claim could be. What possible contribution to justification could the fact that the conduct was not wrongful be making without recourse to the counterfactual claim (unavailable to mixed theorists) that "in the absence of knowledge, exoneration would still be appropriate?" That the actor's conduct was actually justified seems a makeweight, a third wheel, a fact not doing any moral work because the actor's subjective state is sufficient to avoid liability.[60]

The virtue of the conflicting norms argument is that it suggests a picture of how justifications work that responds to the criticism above. Consider the case of a person who burns the field of another in order to create a firebreak which will save a town from an oncoming forest fire. On the conflicting norm picture, the person has violated the prohibitory norms not to destroy the property of another rather than fallen into an exception to it. Because he did so intentionally, he should be blamed for violating the prohibition. But the person has also prevented the town from being destroyed. Having complied with the justificatory norm of preventing greater evil, he, by symmetry, deserves praise or credit. Indeed, the praise or credit he is due is greater than the blame he is due for the destruction of the field because the town's survival is morally more important than the field's. Under this picture, knowingly justified conduct precludes criminal liability because the credit it entails outweighs the blame entailed by the breach of the prohibitory norm. In this light, it can clearly be seen why actors who are unknowingly justified may not escape criminal liability. While blameworthy for engaging in prohibited conduct, they cannot claim credit for the justificatory circumstances or results. The notion of attribution, central to desert-based theories of criminal liability, under-

59 *See generally* Douglas N. Husak, *The Serial View of Criminal Law Defenses*, 3 CRIM. L.F. 369 (1992) (discussing traditional prioritizing of objectively based defenses over subjectively based ones).

60 A similar objection to Fletcher's theory may be raised by Robert Schopp. *See* Schopp, *supra* note 4, at 1277 (describing the requirement of knowledge as "redundant").

writes the mixed theory of justification. Just as persons should not be punished without some culpability-establishing mental state, so Fletcher might argue, persons should not receive credit without justificatory knowledge or purpose. This claim is confirmed by our moral practices. An "accidental hero" will not receive the same type of praise as one who has also acted from heroic ideals. Imagine a drunken reveler on a yacht who throws a life preserver overboard in jest; the life preserver lands in the hands of a drowning child who had fallen into the water unobserved and who is thereby saved. The type of praise due the accidental hero, if any, is a shallow or mock type of praise, analogous to the disapproval due a person who has caused great harm, but clearly through no fault of her own. Just as the criminal law does not recognize this disapproval through sanctions, so, it might be argued, it should not register the weak praise due the unknowingly justified actor.

There is a difficulty, however, with the conflicting norm's conceptualization of justification (as I have elaborated it above).[61] The difficulty is that it is not thoroughly consistent with objective retributivism. According to objective retributivism, persons deserve punishment based on their culpability for wrongful conduct.[62] Attempts to commit offenses are punished because attempting an offense is wrongful, though not as wrongful as successfully completing the offense.[63] As a corollary, credit is deserved for attempting to cause a desirable result, although not as much credit as for successfully causing a desirable result. This corollary squares with our moral practices. A person may be praised for swimming out to save a drowning child but will be treated as a hero only if she succeeds. Pursuant to this view of desert, a person who committed an offense and successfully achieved a socially beneficial result should be treated differently than a person who committed the same offense but failed to achieve a socially beneficial result. For example, if *A1* should not be punished for causing the death of ninety-nine people where *A1* did so in a reasonable and successful effort to save one hundred lives (the credit being just sufficient to offset the blame), then it would seem that *A2* should be punished for causing the death of ninety-nine people where so doing was a reasonable, but wholly unsuccessful, effort to save one hundred lives. The blame generated by killing ninety-nine would not be outweighed

61 *See supra* notes 54–57 and accompanying text.

62 *See* Moore, *supra* note 8, at 237 (distinguishing culpability and wrongdoing as independent bases of desert).

63 *See* Crocker, *supra* note 8, at 1063, 1096 (arguing that the punishing of unsuccessful attempts is based on their wrongful imposition of harm on their intended targets).

by the credit generated by merely attempting to save one hundred, which would be far less than the credit for actually saving the lives. But this conclusion is flawed. We feel that *A2*, just as *A1*, should not be liable. The conflicting norm conceptualization of the mixed theory of justification thus cannot be correct. Justifications are not simply norms like prohibitory norms, generating credit instead of blame. The moral significance of justifications, unlike prohibitions, does not vary with whether the intended results occurred. Asymmetries in our intuitions about the significance of unsuccessful attempts and putative justifications block any effort at conceptualizing justifications as full-fledged norms comprised of objective and subjective elements.

D. The No-Harm Argument

Let us now shift to the first of two arguments advanced by Professor Robinson in support of the objective theory of justification and the availability of justification defenses to unknowing actors. In an early statement of his views, Robinson argues that liability is inappropriate in cases of unknowing justification because (1) harm is a prerequisite for criminal liability, and (2) in cases of unknowing justification, there has been no harm.[64] This argument seems incomplete at best. Robinson relies on the example of a maliciously set fire to a field which saves a town from destruction. Robinson argues, "since no harm has occurred, the incident should be of no concern to the criminal law."[65] Nevertheless, it seems difficult to deny that the field owner has been harmed by the loss of his field. Perhaps Robinson means that there has been no societal harm in the sense that society, as a whole, has not been harmed but has benefited. This manner of characterizing the events seems to imply an artificial and somewhat totalitarian personification of "society" as the true concern of the criminal law. A more traditionally liberal, and arguably superior, understanding of the criminal law is that it is concerned with citizens as individuals and only derivatively with society, conceived of as an aggregate of individuals. Any alternative characterization which slights the harm to individual citizens is flawed. Alternatively, Robinson may simply mean that where the act is approved of, there should be no punishment. But why? It appears that Fletcher accurately characterizes Robinson's discussion as a work of scholarship internal to the criminal law.[66] Robinson seems primarily concerned with arguing that punishment of the unknowingly justified actor would be inconsistent with what he terms

64 *See* Robinson, *supra* note 15, at 288–91.

65 *Id.* at 290.

66 Fletcher, *supra* note 17, at 294–95.

the "accepted role"[67] of the criminal law and less concerned with defending that accepted role as a matter of justice.

E. Failed-Attempt Analogy Argument

In a more recent writing, *Competing Theories of Justification*, Robinson explicitly makes the claim that the objective theory of justification is more just than the mixed or subjective theories of justification.[68] His argument is straightforward. Robinson's premise is objective retributivism, the view that resulting harm is at least sometimes relevant to determining the amount of punishment that should be imposed.[69] Robinson argues by analogy. In Case 1, *D1* maliciously attempts to ignite the field of his neighbor but does not succeed. In Case 2, *D2* unknowingly saves a town by maliciously burning the field and thereby creating a firebreak that prevents a forest fire from reaching the town. According to Robinson, just as the objective occurrence of the failure to ignite the field in Case 1 decreases *D1*'s penalty to, at most, attempted arson, so the objective occurrence of the saving of the town in Case 2 should allow *D2* to invoke the necessity justification and decrease his penalty to, at most, attempted arson.[70] For Robinson, what is critical about each case is the absence of net social harm.[71] In a reformulation of his earlier position of societal harm as a prerequisite for criminal liability, Robinson now contends that *net* societal harm is what triggers potential criminal liability.[72] In Robinson's view, Case 1—a case of failed attempt—and Case 2—a case of unknowing justification—simply represent different but morally equivalent ways that a person intent on an act that he believes will cause net social harm could fail to produce net social harm.[73]

The flaw in Robinson's argument is its explicit reliance on the notion of net social harm. Although Case 1 and Case 2 share the absence of net social harm, Robinson has not established that net social harm is the relevant concept for determining what punishment a person deserves. Indeed, the concept of net social harm seems irrelevant to the inquiry. Consider the doctrine of transferred intent. Pursuant

67 Robinson, *supra* note 15, at 266.

68 *See* Robinson, *supra* note 4 at 61–67. In *Competing Theories of Justification*, Robinson makes a variety of other arguments. *See id.* These arguments do not speak to the justness of the objective theory of justification, the principal concern of this Article.

69 *Id.* at 56.

70 *Id.*

71 *Id.* at 45.

72 *Id.*

73 *Id.* at 57–58.

to this doctrine, if *A* shoots at *B* intending to kill her, but instead hits and kills *C*, an unforeseeable bystander, *A* will be held liable for murder. Although the doctrine's place in criminal law is well established,[74] whether the doctrine is just is controversial.[75] The core objection is that *A* did not intend, and could not foresee, *C*'s death and so should not be held responsible for it. There is a perceived lack of concurrence between the harm intended and the harm caused which occurs in paradigm cases of personal responsibility. *C*'s death seems an unforeseeable, albeit a tragic, accident, and the criminal law generally does not hold persons liable for accidents, even when the persons are of bad character.[76] Furthermore, although *A* intended to cause a net social harm (*B*'s death) and did cause a net social harm (*C*'s death), this concurrence of intent and result is not considered sufficient to bring the accidental death within the responsibility-conferring paradigm of concurring intended and actual harm. The reason is that in order to establish this concurrence of intent and result, both *A*'s intent and *C*'s death must be described in artificial and overly abstract terms, terms that do not jibe with our natural understanding of the events. *A* did not think of himself as trying to cause "a net social harm"; *C*'s family and friends would not have thought of *C*'s death as a "net social harm." Moreover, positive law uncontroversially limits the criminal doctrine of transferred intent to cases where the harm caused is of the same type as the harm intended.[77] Imagine that *A* intends to set off a bomb to explode a building, and the bomb

74 *See* 1 LAFAVE & SCOTT, *supra* note 34, at 399.

75 *See* DON STUART, CANADIAN CRIMINAL LAW 246 (1982) ("The doctrine of transferred malice should be rejected . . . ."); WILLIAMS, *supra* note 4, at 134–36; Anthony M. Dillof, *Transferred Intent: An Inquiry into the Nature of Criminal Culpability*, 1 BUFF. CRIM. L. REV. 501, 510–36 (1998) (arguing against the moral soundness of transferred intent); Martin R. Gardner, *The* Mens Rea *Enigma: Observations on the Role of Motive in the Criminal Law Past and Present*, 1993 UTAH L. REV. 635, 712 (arguing that the doctrine of transferred intent is "theoretically incoherent"); David J. Karp, Note, *Causation in the Model Penal Code*, 78 COLUM. L. REV. 1249, 1268–72 (1978) (arguing that in such cases, transferred intent should be replaced by a rebuttable presumption of aggravated recklessness).

76 The felony murder rule is an exception to this principle. The felony murder rule, however, is generally condemned by commentators and confined by courts. *See* James J. Tomkovicz, *The Endurance of the Felony-Murder Rule: A Study of the Forces That Shape Our Criminal Law*, 51 WASH. & LEE L. REV. 1429, 1431 n.9 (1994) (citing extensive literature critical of the felony murder rule).

77 *See* MODEL PENAL CODE § 2.03(2)(a) (1962) (stating that the actual result may diverge from intended "on the respect that a different person or different property is injured or affected"). *See generally* Queen v. Faulkner, 11 Ir. R.-C.L. 8 (Cr. Cas. Res. 1877) (overturning the conviction of a sailor who accidentally set fire to his ship, destroying it, while attempting to steal rum); Queen v. Pembliton, 2 L.R.-C.C.R. 119

misfires and breaks *B*'s leg. Even if breaking a person's leg constitutes the same amount of net social harm as destroying a building, *A*, liable for attempted arson and perhaps reckless assault, will not be punished as the person who intends to break another's leg and does so. The two cases are not analogous despite their involving the same net social harm; only in the latter can the actor be said to be responsible for the result because he intended it. Although it might be said that *A* intended to cause a net social harm based on the destruction of the building, and did cause such a harm based on the breaking of *B*'s leg, this attempt to bring *A* within the concurrence paradigm is artificial and unpersuasive. These examples imply that (1) for results to be relevant for criminal liability, they must be describable at a level of concreteness such that it can fairly be said that they were intended (or risked) by the actor, and (2) that "net social harm" is a description that generally operates at too high a level of abstraction to be relevant for the criminal law.

This lengthy analysis has a quick payoff for the problem of unknowing justification. Although the unknowingly justified actor caused no net social harm, this is not the relevant description of the results for determining liability. Rather, for purposes of determining desert, we should say that the unknowingly justified actor caused a result of a particular type (the destruction of a field, the killing of a person) that was foreseen or intended and, additionally, caused a result of another type (the saving of a town, the preventing of an attack) that was unforseen and unintended. Because the latter result was unforseen and unintended, it should be considered a mere accident that happened to accompany an otherwise criminal act. Therefore the result—like the breaking of *B*'s leg and the death of the unintended victim—should be ignored by the criminal law when assessing liability.

F. The No-Wrongdoing Argument

Michael Moore presents a somewhat more sophisticated version of the failed attempt analogy argument. As discussed below, Moore's argument, rather than relying on the problematic concept of net social harm, employs the basic, ineliminable, moral concept of wrongdoing. Based on this argument, Moore asserts the validity of the objective theory of justification, under which the unknowingly justified actor is not liable.[78] Indeed, Moore seems to take the hard-line position that the availability of a justification defense to the unknow-

(Cr. Cas. Res. 1874) (overturning the conviction of a defendant who threw a stone at a person and broke a window when he missed).

78 *See* MOORE, *supra* note 4, at 177–83.

ingly justified actor rests not merely on a plausible analogy but is virtually entailed by objective retributivism.[79]

Moore's argument rests on a premise that is not challenged in this Article. According to Moore, aside from attempts, the criminal law properly prohibits only conduct that is objectively morally wrong.[80] Additional prohibitions of conduct unduly limit a person's liberty and cannot be justified. This requirement makes sense. If the criminal law is to have a basis in morality, what it prohibits must be conduct that, in some moral sense, should not be performed.[81] The second step of Moore's argument relies on a particular view of justification defenses. According to Moore, the legal category of justifications are just qualifications of the broadly and loosely stated rules regarding morally wrong conduct contained in offense definitions.[82] There is, Moore argues, no substantive distinction between those elements of the ultimate moral norm that appear in the offense definition and those that appear in the justification definition.[83] In other words, the absence of justifying circumstances is an integral, undistinguished part of the immoral conduct. Thus, if the justifying circumstances occurred, there would be no wrongdoing to trigger an objective retributivist response by the criminal law, regardless whether the actor was aware of the justifying circumstances.

An example clarifies Moore's position. Although killing another is, as a general rule, wrongful, it is not wrongful in all cases. An im-

79 *See id.* at 181 (noting that the contrary position "makes no sense to me").

80 *See id.* at 179, 243.

81 Conduct that is not immoral may properly be prohibited in limited circumstances. The best examples of such prohibitions are those employed to solve coordination problems. Coordination problems arise where the appropriate conduct for each actor is indeterminate because such conduct depends on the appropriate conduct of others. To take a well-known example, it does not matter which side of the road people drive on, as long as everyone drives on that side. Accordingly, the state may properly prohibit driving on the left side of the road in order to establish a conventional side for people to drive on. This prohibition would be proper despite the fact that driving on the left side is not immoral in the absence of an existing convention to the contrary. Discussions of coordination problems in law may be found in JOHN FINNIS, NATURAL LAW AND NATURAL RIGHTS 231–59 (1980); JOSEPH RAZ, THE MORALITY OF FREEDOM 49–50 (1986); Chiam Gans, *The Normativity of Law and Its Co-ordination Function*, 16 ISRAEL L. REV. 333 (1981); Leslie Green, *Law, Co-ordination and the Common Good*, 3 OXFORD J. LEGAL STUD. 293 (1983). Enforcement of prohibitions employed to solve coordination problems, however, need not fall to the criminal law nor be considered instances of punishment. Thus, they are not necessarily counter-examples to the claim that the criminal law should only punish immoral conduct.

82 *See* MOORE, *supra* note 4, at 179.

83 *See id.*

portant qualification of the general rule concerns the killing of an unjustified attacker wielding lethal force. It is not wrongful to kill such a person. Most penal codes reflect this moral fact by establishing a general prohibition on killing and a justification defense permitting the killing of such attackers. A penal code, however, could alternatively simply prohibit the killing of all "unjustified attackers threatening lethal force." The manner in which a penal code is drafted is arbitrary from the perspective of morality. Under either formulation, if an unjustified deadly attacker is killed, there simply has been no objective wrongful conduct for the criminal law to address or redress, regardless whether the defendant knew that the victim was an attacker.

This argument may be generalized. Because the essential prohibition of the criminal law is the prohibition of immoral conduct, and justification defenses are no more than an optional tool of drafting, we might imagine a unified idealized code that contained a single provision:

Unified Code

(1) It is a felony to engage culpably in wrongful conduct.

If a standard instance of unknowingly justified conduct occurred under this code, it would appear that there should be no liability because courts would only be authorized to impose a sanction, pursuant to (1), where there has been wrongful conduct. Unknowingly justified conduct, such as killing an attacker, is not wrongful conduct because it is morally justified. Because the imagined code above captures the essence of the criminal law, Moore might argue that more elaborate and detailed criminal codes likewise should not criminalize unknowingly justified conduct.

1. Rules of Conduct, Evaluation, and Legality

The force of Moore's position flows from the generally close connection in law and morality between rules of conduct and rules of evaluation (also known as rules of decision, rules of adjudication, or rules of culpability). Rules of conduct include such prohibitions as "do not lie" or "do not take what is not yours," which are addressed and apply to persons generally and establish legal obligations that bind law-abiding citizens irrespective of the existence of enforcement sanctions.[84] Rules of evaluation are rules such as "condemn liars" or "imprison thieves" that apply either to persons or, more relevantly, to

84 *See* MOORE, *supra* note 19, at 405.

state organs or officials such as courts or judges.[85] What triggers an evaluation and what a person is evaluated in light of, it may be thought, is a breach of the rules of conduct. Put in moral terms, what retribution is exacted for, under this view, is exactly a breach of the rules of moral conduct. Criminal codes are uniformly drafted to foster this appearance of a close analytic connection between rules of conduct and rules of evaluation. A penal statute that states, "*X* is a felony" simultaneously directs citizens not to engage in *X* and authorizes a court to punish a person as a felon for engaging in *X*.[86] This tie between rules of conduct and rules of evaluation is bolstered by the principle of legality.[87] According to this principle, an individual should only be punished for engaging in conduct previously prohibited by statute.[88] Thus, the proponent of punishing the unknowingly justified actor would appear to be committed to either (a) expanding the criminal law to prohibit nonwrongful conduct, such as unknowingly justified conduct, or (b) to rejecting the principle of legality in order to punish nonprohibited conduct, such as justified conduct. Neither course seems acceptable.

Nevertheless, there is a way out. The dilemma created by the principle of legality may be avoided by restructuring the criminal code to break the connection between rules of conduct and rules of evaluation. Imagine an alternative code that explicitly distinguished such rules.[89] The simplest such bifurcated penal code might have two provisions:

Bifurcated Code

(1) Do not engage in morally wrongful conduct.

(2) Courts may penalize persons for engaging in morally wrongful conduct.

---

85 *See id.* (comparing norms of wrongdoing and culpability); *see also* Meir Dan-Cohen, *Decision Rules and Conduct Rules: On Acoustic Separation in Criminal Law*, 97 HARV. L. REV. 625, 625–30 (1984) (distinguishing conduct rules from decision rules).

86 For an argument rejecting the (intuitively obvious) thesis that laws direct citizens to act in conformity with them, see, generally Gregory M. Silverman, *Imperatives, Normativity, and the Law*, 31 CONN. L. REV. 601 (1999).

87 *See* Bernard E. Gegan, *More Cases of Depraved Mind Murder: The Problem of* Mens Rea, 64 ST. JOHN'S L. REV. 429, 460 n.109 (1990) ("A growing number of scholars have criticized [the requirement of knowledge of the justifying circumstances] as violative of the basic principle of legality.").

88 *See* John Calvin Jeffries, Jr., *Legality, Vagueness and the Construction of Penal Statutes*, 71 VA. L. REV. 189, 190 (1985).

89 The idea of a penal code's explicitly distinguishing rules of conduct from rules of evaluation is not as unusual as it may seem. *See generally* Paul H. Robinson et al., *Making Criminal Codes Functional: A Code of Conduct and a Code of Adjudication*, 86 J. CRIM. L. & CRIMINOLOGY 304 (1996) (advocating such a penal code).

It would not be necessary, however, that the conduct and evaluation provisions of a bifurcated code parallel each other as do the two provisions above so that what is prohibited is what is sanctioned. Indeed, an advantage of a bifurcated code, besides its theoretically greater clarity, would be its potential to permit divergent conduct and evaluation provisions. A bifurcated penal code might contain the following divergent provisions:

Divergent Bifurcated Code

(1) Do not engage in morally wrongful conduct.

(2) Courts may penalize persons for engaging in morally wrongful conduct *and* for engaging in conduct that would be morally wrongful but for any justifying circumstances of which the actor is unaware.

In this divergent bifurcated code, the scope of (2), the penalty authorization provision, is broader than (1), the conduct prohibition provision. Such a code initially may strike the reader as conceptually confused. In what sense can conduct be penalizable, yet not prohibited (or vice-versa)? Is such conduct "permissible" or not? But the novel should not be mistaken for the nonsensical. There is no necessary connection between the prohibited and the punishable. It is possible to be subject to a legal obligation that will not be enforced, just as it is possible to be obligated by a promise even if breaching the promise will result in no adverse consequences to the promisor.[90] Likewise, as relevant here, one may have no legal obligation to act in a particular way but merely a prudential incentive to not engage in acts that would trigger a penal response.[91]

Even if logically coherent, would the divergent bifurcated code set forth above be objectionable on the ground that it would unduly infringe a citizen's liberty? Here questions of political legitimacy arise, for it may be thought that the coercive power of the state can only be validly triggered by the breach of the rules it has imposed on

90 Indeed, not only is it possible that a bifurcated penal code might punish what was not explicitly prohibited, it is to be expected. An elegantly drafted bifurcated code, while of course providing in the evaluation section that attempts should be punished, would not contain rules of conduct that explicitly prohibited attempts. Rules of conduct explicitly prohibiting attempts would be unneeded because in directing *P* that conduct *C* not be engaged in, *P* is directed not to attempt to engage in *C*. The commands "Do not kill" and "Do not attempt to kill" are equivalent commands despite the fact that "Joe killed" and "Joe attempted to kill" are not equivalent statements. Attempts therefore represent a natural case of divergence of conduct and evaluation rules. *See* Heidi M. Hurd, *What in the World Is Wrong?*, 5 J. CONTEMP. LEGAL ISSUES 157, 187–93 (1994).

91 *See* H.L.A. HART, THE CONCEPT OF LAW 79–88 (1961) (distinguishing having a prudential reason—being obliged—from having an obligation).

its citizens. But this objection is unfounded. First, as a formal matter, the code would not create legal obligations to refrain from conduct that are more extensive than those created by the nondivergent bifurcated penal code presented earlier in the paragraph. Provision (1) of each code establishes the legal obligations of citizens, and these provisions are identical. In particular, as relevant to the current discussion, under the divergent code a law-abiding citizen would not be, and would not understand herself to be, under a legal duty to refrain from unknowingly justified conduct. Provision (1) of the divergent code, like that of the nondivergent code, merely prohibits citizens from engaging in morally wrongful conduct. Such conduct does not include any justified conduct, either known or unknown. Second, as a practical matter, under a divergent bifurcated code, it is not the case that a citizen's liberty would be infringed relative to the standard nondivergent code because of the fear of being penalized for unknowingly justified conduct. Because unknowingly justified conduct is relatively rare, it is unlikely as a practical matter that any person would be deterred from engaging in any additional conduct that she was not already deterred from based on likelihood that it was subject to sanction as morally wrongful conduct. Stated in other words, because by definition you never know you will be engaging in unknowingly justified conduct, you will never be deterred from such conduct based on the possibility that it is unknowingly justified conduct. The divergent bifurcated code above punishes more conduct but does not deter more conduct. Furthermore, to the extent it does deter additional conduct that might have been engaged in on the off-chance it might be unknowingly justified, virtually all of the conduct so deterred will be wrongful conduct.

Even if not unduly liberty-infringing, would the penal code set forth above violate the principle of legality? Let us imagine a citizen who has engaged in unknowingly justified conduct, say burning another's field to create a town-saving firebreak, and then is informed he will be punished for his conduct despite the fact that the conduct in those circumstances, owing to its overall good consequences, would not be considered morally wrong and so is not prohibited. Arguably, to sanction a citizen under those circumstances would violate the letter of the legality principle because a court would be punishing conduct that is not specifically prohibited by the rules of conduct. Nevertheless, such a code would not be inconsistent with the principle's spirit. According to Moore, the principle of legality rests on the

values of liberty, democracy, equality, and fairness.[92] None of these values are offended by the divergent bifurcated code suggested in this Article. Clearly, such a code could be adopted democratically and would not promote unequal and arbitrary application of the law. More to the point, not giving citizens the benefit of justifying circumstances that they were unaware of does not unfairly surprise them. First, as discussed above, in any particular case of unknowing justification they could not have relied on the presence of these circumstances. Second, citizens might have affirmative prior knowledge of the provision of the penal code that permits punishing unknowingly justified conduct. Third, even if they did not have such knowledge, they would likely have no strong expectations regarding whether the legal system incorporates a mixed or objective theory of justification. Thus, being punished would not come as a surprise.[93] Finally, the content of the bifurcated code is determinate enough not to curtail liberty by causing uncertainty regarding the law's application. For these reasons, even when coupled with the requirement that only immoral conduct be prohibited, sanctioning the unknowingly justified actor does not offend the legality principle's rationales of democracy, equality and fairness.

In sum, the premise of Moore's argument[94]—only wrongful conduct should be prohibited—does not yield the conclusion of his argument—unknowingly justified conduct should not be punished.[95] The syllogism "what is justified is permissible, and what is permissible is not punishable" was earlier considered and found unpersuasive when construed as a matter of definition.[96] While Moore recasts the syllogism as a matter of morality, it is no more compelling in this form. Unknowingly justified conduct represents a narrow exception to the legality principle that only conduct made illegal by the legislature should be punished. Even though unknowingly justified conduct should not be prohibited, it can be punished. Such punishment

92 *See* MOORE, *supra* note 4, at 239–44. *See generally* Jeffries, *supra* note 88 (examining the rationales of the legality doctrine).

93 The Supreme Court recently affirmed the constitutionality of imposing criminal liability for conduct that was not prohibited when undertaken on the ground that its subsequent judicial criminalization was not an unfair surprise. *See* Rogers v. Tennessee, 532 U.S. 451, 467 (2001) (affirming a state supreme court decision abolishing the common-law "year and a day" rule and holding it was not a prohibited ex post facto law). The Court thereby implied the principle of legality was no broader than the no-surprise rationale that arguably supports it.

94 *See* MOORE, *supra* note 4, at 179, 243.

95 *See id.* at 177–83.

96 *See supra* notes 18–19 and accompanying text.

neither unduly interferes with liberty nor offends any of the values that underlie the legality principle.

2. Counter-Examples Involving Nonproximate Results

Having hopefully blunted the force of Moore's position, let me now present a series of counter-examples to Moore's thesis that punishment based on objective features of a person's conduct is inappropriate where the conduct is not objectively wrongful. The first counter-example is based on the lesser evil principle that conduct is not objectively wrongful if it produces a net social benefit. Consider a variation of the unknowing justification scenario in which *D1* maliciously burns his neighbor's field. In this variation, as a result of the fire, the neighbor cancels his plans to take his family on a vacation voyage on an ocean liner, the Titanium; the Titanium sinks and all its passengers are drowned. Or, in another variation, the neighbor, as a result of the fire, decides to move cross country where he wins a large lottery and uses the proceeds to create a medical research facility which discovers an important cancer treatment. In both variations, a net social benefit occurred as a result of the burning of the neighbor's field. The social benefits of saving his family from drowning or of extending the lives of cancer victims far offset the cost of the destroyed field. Thus, under an objective theory of justification, because *D1*'s conduct was justified under the lesser evils principle, *D1* would not be punished. This is a counter-intuitive result. The next counter-example is based on the defense-of-others principle that it is not immoral to kill a person—even a morally innocent one, such as a hallucinatory aggressor—to save the life of a wholly innocent person. Consider *D2*, who maliciously kills *V*, who, if he had not been killed, would have many years later been responsible for a car crash killing innocent person *P1*. *D2* thus saved an innocent person's life by killing *V*, who was no more morally innocent than a hallucinatory aggressor. Under the defense-of-others principle, *D2*'s conduct would be objectively justified. Again, if a person should only be punished for conduct where that conduct is objectively immoral, it seems *D2*, counter-intuitively, should not be punished for his conduct. Unknowingly, he did the right thing.

Implementing Moore's theory into the criminal law would have dramatic consequences. It is arguably impossible to assess all the future consequences of an act.[97] It is clearly impossible to prove beyond a reasonable doubt that the net consequences of an action will be

97 As the philosopher G.E. Moore has written concerning ethical reasoning based on consequential considerations:

negative or that an innocent life will not consequently be saved if those consequences extend to infinity. Indeed, it is not uncommon that criminal acts will have net positive consequences if a particular category of nonproximate results are considered. Many criminal acts have the positive outcome of preventing further similar crimes by either triggering greater precautions against similar crimes or deterring those who have learned of the perpetrator's capture.[98] If a penal code incorporates some consequentialist considerations in defining prohibited conduct and consequences include all of an act's results, it very often will be impossible to establish beyond a reasonable doubt that a defendant is criminally liable. This is a plainly unpalatable conclusion and may be considered a *reductio ad absurdum* implication of Moore's position that the unknowingly justified act should be exonerated based on the absence of objective wrongdoing.[99]

> The first difficulty in the way of establishing a probability that one course of action will give a better total result than another, lies in the fact that we have to take account of the effects of both throughout an infinite future. . . . [I]t is quite certain that our causal knowledge is utterly insufficient to tell us what different effects will probably result from two different actions, except within a comparatively short space of time. . . . Yet, if a choice guided by such considerations is to be rational, we must certainly have some reason to believe that no consequences of our action in a further future will generally be such as to reverse the balance of good that is probable in the future which we can foresee. . . . Our utter ignorance of the far future gives us no justification for saying that it is even probably right to choose the greater good within the region over which a probable forecast may extend.

MOORE, *supra* note 10, at 202.

98 A *reductio ad absurdum* application of this principle would have potentially every criminal act justified on the ground that it resulted in the conviction and incapacitation of the perpetrator. Such an argument was in fact suggested in *Vaden v. State*, 742 P.2d 784 (Alaska Ct. App. 1987), *aff'd*, 768 P.2d 1102 (Alaska 1989). There, an undercover police officer shot four foxes illegally from a plane piloted by the defendant. *Id.* at 785. The officer of course was not liable for killing the foxes because of the law enforcement justification; the court, however, would not allow the defendant to raise this defense to preclude his liability as an accomplice despite his being unknowingly justified. *Id.* at 786.

99 It will not do to shift the burden of proof to the defendant. Such a shift would be unprincipled. The defendant has no better access than the prosecution to evidence regarding the ultimate effects of his actions. Furthermore, to shift the burden of proof to the defendant with respect to some issues is contrary to the premise of Moore's position, which is that the absence of justifying conditions is an integral, undistinguished part of the underlying norm that the criminal law enforces. *See* MOORE, *supra* note 4, at 181. Nor will it do to shift the blame to the criminal law's employment of consequentialist norms of justification. As noted earlier, the validity of a theory of legal justification should not depend on the moral validity of a particular legal system's rules of justification. *See supra* note 13 and accompanying text.

One response to the above counter-examples is that they improperly relied on justifying circumstances (lives saved, cancers treated) that were not proximately caused. The criminal law, it may be argued, is only concerned with proximately caused consequences. This point is demonstrated by the fact that the criminal law will not punish a person based on the nonproximate harms of her actions. Thus, with respect to the scenarios of the previous paragraph, the relevant legal norms are "do not *proximately* cause the destruction of property unless the destruction *proximately* causes a social benefit outweighing the harm" and "do not *proximately* cause the death of another unless so doing is necessary to *proximately* prevent the death of an innocent person." Because these more precisely stated norms were breached, it may be argued, *D1* and *D2* should be liable for arson and murder, respectively. Thus, by including the proximity requirement within the rules of conduct, Moore's principle which would vindicate the unknowingly justified actor—"no punishment without wrongful conduct"—can be maintained.

The issue turns on whether considerations of proximity are part of the norms that govern conduct and define wrongfulness. A person, it is generally agreed, does not deserve to be punished based on the nonproximate results of her actions. This fact, however, does not imply that considerations of proximity are relevant to wrongfulness. Desert is a concept of two components. Many criminal law theorists analyze desert in terms of wrongdoing and attribution (or harm and responsibility).[100] Wrongdoing concerns the breach of the applicable norms of conduct; attribution concerns the actor's accountability, responsibility, or blameworthiness for the breach.[101] Mens rea is the standard example of an element of an offense that is relevant to attribution rather than wrongdoing. A person does not deserve to be punished based on the unintended, unforeseen, or unforeseeable results of his conduct, even if it is objectively wrongful. One way of putting this point is that the norms of conduct do not include matters of attribution, such as mental states. "Do not kill," rather than "do not intentionally kill" or "do not knowingly kill," is the rule of conduct.[102] Intent, knowledge, or lack thereof, is relevant only to the evaluation of the actor.

Research has disclosed no specific discussion of whether the criminal law's proximate causation requirement is relevant to the establish-

100 *See* FLETCHER, *supra* note 5, § 6.6, at 461; MOORE, *supra* note 4, at 243; ANDREW VON HIRSCH, DOING JUSTICE: THE CHOICE OF PUNISHMENTS 79–81 (1976).

101 *See* Byrd, *supra* note 4, at 1290.

102 *See* FLETCHER, *supra* note 5, § 6.6, at 457.

ment of wrongdoing or the attribution of wrongdoing.[103] Nevertheless, a strong case can be made for its relevance to attribution of wrongdoing. To the extent that the relevant norms enforced through the criminal law are consequentialist, it seems irrelevant how the morally relevant consequences come about. At the most basic level, society arguably wants to increase happiness, protect innocent lives, vindicate legitimate expectations, maximize efficiency, provide opportunities for flourishing, preserve security, and so on. Whether these ends are achieved through direct or indirect causal chains should not matter. Consequences are consequences.

Furthermore, the requirement of causal proximity naturally fits as a component of attribution. First, there is a close connection between proximity and one concept clearly relevant to attribution: foreseeability. Despite the wrongfulness of his conduct, we do not hold the actor responsible for his conduct if it were unforeseeable that it would be wrongful. According to a widely held view of proximate causation, cause and effect are not proximately related if their connection could not have been foreseen.[104] If norms of conduct do not include mens rea terms such as foreseeability, it is unclear why they should include foreseeability-derived terms such as proximate causation. Second, conclusions about proximity are often described in terms of responsibility or attribution. In determining whether an actor may fairly be held accountable for a result, the nature of the causal chain between the person's basic act and the result is clearly relevant. If the chain is too attenuated or unusual, the result will not be attributed to the actor. Terms such as "blame" and "culpable for" are often employed to express a judgment of attribution. It makes sense to say that an actor "should not be blamed" for an occurrence that is causally remote or that an actor is presumed to be "culpable for" the natural and direct consequences of her acts. In this way, the proximity of a result, like foreseeability of a result, seems to go to whether the conduct should be attributed to the actor, not to whether the conduct is wrongful.

103 Moore indicates that causation is part of the actus reus. MOORE, *supra* note 4, at 177. Fletcher refers to causation as relevant to objective attribution. FLETCHER, *supra* note 5, § 5.3, at 386–87. Neither discuss the proximity requirement specifically. Robinson defines causation with the proximity qualification in his adjudication section of his model code, but suggests that the definition merely clarifies the meaning of the rules of conduct. *See* Robinson et al., *supra* note 89, at 319, 345.

104 *See* DAN B. DOBBS, LAW OF TORTS 463–70 (2000); W. PAGE KEETON ET AL., PROSSER AND KEETON ON THE LAW OF TORTS § 43, at 297–98 (5th ed. 1984) (analyzing proximate causation in terms of foreseeability and comparing foreseeability to what "any reasonable person would really have had . . . in mind").

Assuming that causal proximity is relevant to attribution and not wrongdoing, it should no more matter for purposes of punishment that a person's conduct was not wrongful based on consequences that were unforeseen by the actor than that her conduct was not wrongful based on consequences that were not proximately caused by the actor. As discussed below, in both cases there is responsibility for some harm, which, if not excused, should be the basis for punishment. Thus, contrary to Moore's position, the fact that an unknowingly justified actor did not act wrongfully should not be a bar to punishment.

IV. AN ALTERNATIVE APPROACH

Thus far, this Article has been critical. Having surveyed the approaches to unknowing justification found in the existing literature, I now will seek to articulate a theory of punishment and justification that properly handles cases of unknowing justification.

A. The Harm Theory Explained

The theory I propose takes harm as the basic ground for punishment and so may be called the harm theory of punishment. As the name implies, the harm theory is a theory about the conditions for just punishment and should not be confused with other "harm theories" which address the conditions under which the government may legitimately limit liberty.[105] The core idea of the harm theory is that it is morally permissible or appropriate to harm a person to the extent that that person is morally responsible for causing or risking harm to another.[106] By "harm," I mean suffering, disadvantage, a set back of interests, or, more broadly, a state of affairs, such as the breaking of a promise, that provides a moral reason, even if a defeasible one, for not engaging in the act causing the state of affairs.[107] According to

105 *See, e.g.*, JOHN STUART MILL, ON LIBERTY 9 (Alburey Castell ed., Harlan Davidson 1947) (1859) ("[T]he only purpose for which power can be rightfully exercised over any member of a civilized community, against his will, is to prevent harm to others.").

106 The theory is stated using the disjunction "permissible or appropriate" to allow for negative and positive versions of the theory. A negative theory of punishment provides necessary conditions for inflicting punishment; a positive theory provides necessary and sufficient conditions. *See* Kevin Cole, *The Voodoo We Do: Harm, Impossibility, and the Reductionist Impulse*, 5 J. CONTEMP. LEGAL ISSUES 31, 31–32 (1994) (distinguishing positive and negative theories of retribution); Sanford H. Kadish, *Foreword: The Criminal Law and the Luck of the Draw*, 84 J. CRIM. L. & CRIMINOLOGY 697, 698 (1994) (same).

107 *See generally* 1 JOEL FEINBERG, THE MORAL LIMITS OF THE CRIMINAL LAW: HARM TO OTHERS 31–64 (1984) (discussing the concept of harm). In contrast to Feinberg's

the harm theory, a person is "morally responsible" for the harm she intentionally, knowingly, recklessly, or negligently causes or risks, absent a factor negating moral responsibility for the harm. Examples of such factors are extreme youth, duress, reasonable mistake, and reasonably believing the act is justified.

A system of criminal law based on the harm theory would resemble in basic outline the Model Penal Code. Offenses would be formulated in terms of the intentional, knowing, reckless, or negligent engaging in harmful conduct, where such conduct is defined as that which causes or creates a significant risk of more than de minimis harm.[108] The "justification" defenses of a penal code, i.e., those concerned with circumstances bearing on wrongdoing, would largely be drafted in subjective terms in the manner of the Model Penal Code so that they would, when applicable, negate moral responsibility.[109] For example, "It is a defense to liability that the actor (reasonably) believed that his or her act was necessary to prevent a greater harm," might be how the justification defense of necessity would be formulated. The harm theory thus reflects the subjective theory of justification. The "excuse" defenses would be drafted in the usual manner, for example, "It is a defense to liability that the actor was under ten years of age," or "It is a defense to liability that the actor was insane." Excuses, unlike justifications, would not necessarily refer to the actor's beliefs or desires.

Under the harm theory, justification defenses construed as wrongdoing-related defenses based solely on objective states of affairs would exist in only a limited form. To the extent that offenses were overbroadly formulated either to criminalize conduct that was not harmful or to impose sanctions disproportionate to the harm caused by some instances of the prohibited conduct, purely objective justification defenses would be appropriate. For example, imagine that a statute criminalized eating in a restaurant or imposed the death penalty

---

definition of harm, *see* 1 *id.* at 34–35, my definition does not entail the proposition that all harms are wrongfully or unjustly imposed. Thus, in my way of speaking, it is not an oxymoron to say that a person was justifiably harmed. Feinberg's definition of harm seems specially crafted to make plausible his "harm principle." *See* 1 *id.*

108 Other restrictions on "harm" besides the non-de minimis restriction may be appropriate. For example, it may not be appropriate to criminalize the "harm" parents do to their children when they raise them less than ideally or the "harm" a person does to herself when she uses recreational drugs. A detailed theory of the scope of harms that the criminal law should recognize is not needed for the purpose of this argument.

109 *See, e.g.*, MODEL PENAL CODE § 3.02 (1962) (establishing a defense for "[c]onduct that *the actor believes* to be necessary to avoid a harm or evil to himself or to another" (emphasis added)).

for firing a gun. In these cases, it would be appropriate to establish specific justification defenses of "Paying for the food eaten" or "Not killing any person by firing the gun." It should not be necessary that the defendant have knowledge of the justifying circumstances set forth in these defenses to invoke them because the harm that might justify the sanction has not occurred. These defenses essentially trim or refine offenses that are overbroad with respect to the harms (stealing, killing) that underlie them. The general form of these harm-refinement justifications is, "Despite the prohibited conduct, the harm, or risk of harm, that would morally justify the imposition of the sanction did not occur." Because it is likely that some actual offenses impose penalties disproportionate to the harm caused by at least some instances of the prohibited conduct, some unknowing justifications in our current criminal scheme are likely appropriate even under the harm theory. The circumstances relevant to such purely objective justification defenses, however, could easily be formulated as negative elements or "exceptions" of the offense. With respect to these circumstances, Moore is correct in stating that it is a matter of drafting preference whether they are contained in the definition of the offense or not.[110]

Critically, under a criminal law system modeled under the harm theory, the major categories of justifications, such as self-defense and necessity, could not be characterized as harm-refinement justifications. Consider the case of burning the field and thereby saving the town. This is not a case in which there was no harm or no harm proportionate to the punishment for arson. Rather, the harm of burning another's field occurred. Likewise, in the case of force used against an attacker, the harm of being subjected to physical force or experiencing pain occurred. Rather than indicating that the harms did not occur, the circumstances of saving the town or preventing the attack would be objectively justifying (though not liability-relieving) by virtue of offsetting or overriding of the harms. This is the central distinction between punishment theories such as Moore's and Robinson's and the harm theory. Under Moore's and Robinson's theories, all justifications are refinements. For Moore, they are refinements of the conditions of wrongdoing;[111] for Robinson, they are refinements of the conditions of causing net harm.[112] Wrongdoing and net harm are bottom-line negative moral judgments concerning conduct that, once

110 *See* MOORE, *supra* note 4, at 179.

111 *See id.* ("The justifications operate as exceptions to the rules describing such acts as wrong . . . .").

112 *See* 2 ROBINSON, *supra* note 3, at 27 n.33.

established, cannot be overridden. If an act is wrongful (for Moore) or causes net social harm (for Robinson), then it must be prohibited regardless of what other circumstances obtain. In contrast, harmful acts may be either morally justified, as in cases where there are overriding or offsetting considerations, or morally impermissible, as in cases where there are no such considerations. Under the harm theory, the justification of conduct becomes a matter not of construing it in context as non-wrongful, but of overriding or offsetting the harm that has been created. In this way, when an act is justified, there is still something that persists—the harm—that may provide the basis for punishment.

The harm theory implies an answer to the problem of unknowing justification. The actor who satisfies the definitional elements of an offense and who is unaware of circumstances that provide an offsetting reason to act, such as the prevention of a greater evil or the saving of an innocent life, is not entitled to a defense. Rather, such an unknowingly justified actor should be subject to punishment just as any other actor responsible for causing harm. This conclusion follows directly from the premise of the harm theory: a person should be punished for the harms for which he is responsible. The fact that her conduct was objectively justified would be irrelevant because it would neither negate the harm caused nor the actor's responsibility for that harm.

B. The Harm Theory Compared

How does the harm theory compare with its competitors? As noted, the penal code implied by the harm theory would resemble the Model Penal Code. Despite this similarity, the harm theory is a fundamentally different theory of punishment than that of the Model Penal Code drafters. The Model Penal Code is primarily constructed on the utilitarian view that those who present dangers to society should be incapacitated for the good of society.[113] This premise easily generates the conclusion that the unknowingly justified actor should have no defense. The fortuity of justifying circumstances does not make the actor any less dangerous to society. In contrast, the theory advanced here is wholly retributive or desert-based. It is based on the nonconsequentialist idea that what harms you have imposed on others may or

113 *See* MODEL PENAL CODE § 1.02(1)(b), (2)(a) (1962); 1 MODEL PENAL CODE AND COMMENTARIES §§ 1.01 to 2.13, at 18 (1985) (stating pervasive purpose of the grading of particular offenses is the identification of "individuals whose criminal propensities are relatively large"); 1 *id.* §§ 6.01 to 7.07, at 20 (declaring that incapacitation and individual deterrence are proper guides to sentencing).

should be imposed on you. Thus, it is responsive to the intuitions informing objective retributivism that the results of a person's actions are relevant to determining his punishment for those actions. Such retributive sentiments are largely foreign to the Model Penal Code.[114] Though I will not argue for the superiority of retributivism over utilitarianism here,[115] this superiority supports the harm theory, as presented here, over the Model Penal Code.

The harm theory's differences with Moore's theory are more significant for the operation of the criminal law. Earlier, this Article criticized Moore's theory on the ground that it implied that nonproximate results of conduct could render an actor immune from punishment.[116] A more fundamental criticism, however, is that under Moore's theory, the unknowingly justified actor unjustly escapes punishment. As discussed previously, the criminal law has two branches: a branch directed to citizens concerning what acts are prohibited, and a branch directed to officials regarding how they should treat citizens.[117] The first branch rests on a theory of immoral conduct; the

114 Although the drafters of the Model Penal Code eschew the view that the criminal law has a single view, *see* 1 *id.* §§ 6.01 to 7.07, at 22 n.18, they concede that the Model Penal Code's approach to sentencing is "basically utilitarian or consequentialist." 1 *id.* §§ 6.01 to 7.07, at 2. Retributivism, in its limited negative form, is offered as only an alternative to "enlightened utilitarian grounds" for limiting punishment, 1 *id.* §§ 6.01 to 7.07, at 4, and even then the constraints of negative retributivism are not considered absolute. 1 *id.* §§ 6.01 to 7.07, at 17–18. Positive retributivism, according to which desert alone is a sufficient basis for punishment, is rejected wholesale as "inhumane." 1 *id.* §§ 6.01 to 7.07, at 16.

115 The principle criticism of utilitarianism is that it is an impoverished moral theory, incompatible with our commitment to personal liberty or the notion of moral or legal rights. *See* Jules L. Coleman, *Efficiency, Utility and Wealth Maximization*, 8 HOFSTRA L. REV. 509, 511 (1980). Utilitarianism also is viewed as suffering from the doctrinal problems of determining which preference should count and whether total or average utility should be maximized, and the theoretical problem of interpersonal utility comparisons. *Id.* For further criticism see H.J. McCloskey, *A Non-Utilitarian Approach to Punishment*, *in* CONTEMPORARY UTILITARIANISM 239, 248–54 (Michael D. Bayles ed., 1968). *See generally* Louis Michael Seidman, *Soldiers, Martyrs, and Criminals: Utilitarian Theory and the Problem of Crime Control*, 94 YALE L.J. 315 (1994) (arguing that because of inefficient distribution of the cost of crime, it is impossible to achieve a low level of crime under a utilitarian analysis); Andrew Von Hirsch, *Desert*, *in* PHILOSOPHY OF LAW 573 (Joel Feinberg & Hyman Gross eds., 2d ed. 1980) (arguing that desert, instead of utilitarianism, should determine punishment). The most sustained and successful attack on utilitarianism as a general principle for ordering society is found in RAWLS, *supra* note 12. While Rawls appears to endorse at least negative retributivism, *see id.* at 315 ("[I]n a just society legal punishments will only fall upon those who display [bad character]."), he does not address the topic at length.

116 *See supra* notes 78–83 and accompanying text.

117 *See supra* notes 85–86 and accompanying text.

second rests on a theory of punishment. The harm theory is a theory of punishment. With respect to prohibited conduct, wrongdoing is the central concept (or net harm causing, if net harm causing constitutes wrongdoing) because wrongdoing should be prohibited. Harm, however, is the more appropriate basis for punishment because the harm caused manifests the actor's culpability. That your action will cause a death of a person is the affirmative reason for not engaging in the act. Absent an excuse, you are culpable because in acting you failed to grasp, regard, and appreciate this reason. The death makes your immorality concrete and so triggers a retributivist response to the manifestation of immorality in the world. Punishment in these circumstances is a way of communicating to the actor: "This harm is what happens when you act on or in spite of such a reason. You are responsible for it." Punishing based on harm is a way for society to symbolically recognize and respond to the harm.[118]

A possible response to this view is to deny the moral significance of harm causing. It may be argued that where an attacker has been killed or where a town has been saved at the cost of a field, there has been no manifestation of culpability, nothing to regret, no loss to be responsible for and to suffer on account of. There has only been permissible conduct and acceptable results. But the existence of justifying circumstances does not make harm evaporate. The clearest examples of harm arise in cases of necessity. The burning of the field and the saving of the town are temporally and spatially distinct events. The saving of the town is a result that offsets the regrettable destruction of the field. Not to see that the destruction of the field is regrettable because it was necessary is morally obtuse.[119] Although it was not such a significant loss that it was wrong to burn the field, it still carried moral weight. Likewise, the saving of the potential victim is a result that offsets the regrettable taking of a life. There is something valuable about the life of a person, even if that person is about to attack another. It is not as if the attacker about to engage in his attack is suddenly transformed into a nonperson so that killing him is equivalent to killing a fly. Under most moral theories, a fly can be killed even if its death serves no purpose. In contrast, as the criminal

118 For a discussion of the symbolic and expressive function of punishment, see generally Joel Feinberg, *The Expressive Function of Punishment*, 49 MONIST 397 (1965), *reprinted in* JOEL FEINBERG, DOING AND DESERVING 95 (1970).

119 *Cf.* Simester, *supra* note 5, at 303 (arguing that the owner of a diamond ring stolen under duress "suffers a significant loss, one demanding the attention of our criminal law").

law recognizes,[120] a would-be attacker cannot be killed unless his attack is thereby prevented. Killing a deadly human attacker may be no more wrongful than killing a fly, but in taking a human life, it is more harmful. A person should be treated in light of the morally significant states of affairs she is responsible for causing. Harm is such a state of affairs. This recognition undermines Moore's theory, which focuses on wrongdoing to the neglect of harm causing.

Although Fletcher's theory agrees with the harm theory on the critical point that the unknowingly justified actor should be punished, it differs from the harm theory in two respects. First, most obviously, it incorporates a subjective theory of justification rather than a mixed one.[121] Thus, the theory provides a different explanation of why persons are exonerated in situations where both (1) they believe that their conduct is objectively justified, and (2) it is objectively justified. Fletcher would say that they would be exonerated because of both of these facts, while the subjective theory would say that only (1) is relevant.[122] As argued earlier,[123] it is unclear how to give content to Fletcher's view.

Second, and more relevant here, the harm theory and Fletcher's theory differ regarding the nature of offenses. According to the harm theory, a properly drafted penal code will establish offenses such that, absent an excuse or justification (subjectively defined), individuals satisfying the definitional elements of an offense will deserve punishment. The definitional elements of the offense therefore must be themselves morally significant. This theoretical requirement is met by drafting offenses in terms of the culpable causing or risking of harms and setting appropriate penalties. Harm is an inherently moral concept. It is conceptually prior to wrongdoing in the sense that only through considering the harmful implications of a potential act can one conclude whether it would be wrongful.

Fletcher's theory seems to lack a concept like harmful that is conceptually prior to wrongful. Fletcher believes that the conduct defined by offenses is conduct that is "typically"[124] wrongful, where "typically" is a "statistical"[125] matter or a matter of contingent social

120 *See supra* note 34 (discussing the traditional formulation of self-defense doctrine).

121 FLETCHER, *supra* note 5, § 7.4, at 555–56.

122 *See id.*

123 *See supra* notes 47–51 and accompanying text.

124 FLETCHER, *supra* note 17, at 325; *see also* FLETCHER, *supra* note 5, § 7.4, at 562 ("[T]he violation is typically sufficient to regard the act as wrongful.").

125 FLETCHER, *supra* note 17, at 329.

norms.[126] For Fletcher, offenses seem to be merely useful rules of thumb, or "paradigms,"[127] for approximating wrongdoing, but they have no independent moral significance. Like Moore, Fletcher believes that justifying circumstances serve to mark out the exceptional case where that which is typically wrongful is not actually wrongful.[128] It follows from this perspective that where conduct is not typically wrongful, it will be irrelevant what the actor knows of the circumstances of the act—there is not even prima facie liability. Marrying a second time is an example Fletcher uses of such typically benign conduct.[129] According to Fletcher, an actor should not be liable for bigamy based on marrying a second time regardless whether the actor is aware or not that his earlier marriage has been legally terminated.

Different theories of offense generate different cases of conduct for which an actor will be liable even if unknowingly justified. All conduct that is typically wrongful, hence establishing prima facie liability for Fletcher,[130] will be conduct that risks or causes harm. In doing what is typically wrongful, an actor significantly risks wrongdoing and, in risking wrongdoing (where wrongdoing is the causing of unjustified harm), risks harm. But not all conduct that is harmful will significantly risk being wrongful, a requirement for typicality. In particular, harmful conduct that is normally justified will not be typically wrongful. Consider the moderate disciplining of a child or the removal of a diseased tooth. Both acts are harmful because they cause the subject psychological or physical pain, and causing pain is a reason not to engage in the conduct. But they are not typically wrongful because the administering of moderate discipline and the extraction of a diseased tooth (let us assume) are virtually always for the good of the subject. Consider the cases of an angry parent who administers moderate discipline believing it to be purely detrimental or a malicious dentist who removes a diseased tooth believing it to be healthy. Each is aware of the harm involved but unaware of the circumstances that in fact justify the act. Under Fletcher's theory, such an actor is in the position of the person who marries mistakenly believing he is engaging in bigamy.[131] No punishment is appropriate because there is nothing typically wrongful in the acts. In contrast, the harm theory would punish because the parent and dentist are responsible for caus-

126 FLETCHER, *supra* note 5, § 7.4, at 566 ("The question is whether in our society, in the here and now, these forms of conduct are identifiable as typically wrongful.").

127 *Id.*

128 *See id.* § 7.4, at 562.

129 *See id.* § 7.4, at 557.

130 *See id.* § 7.4, at 553.

131 *See id.* § 7.3, at 537.

ing significant pain. The pain is the external manifestation of the parent's and the dentist's moral culpability. Punishing the parent and the dentist like other unknowingly justified actors is the more principled, and hence preferable, result.

Another offense-related difference between Fletcher's theory and the harm theory also concerns the place of consent. Fletcher asks whether nonconsent should be an element of larceny, such that if it does not occur there would be no liability, or whether consent should be a justification, such that if it does unknowingly occur there could be liability.[132] Fletcher indicates that this methodology of looking to what is typically wrongful and inconsistent with social mores does not yield a clear answer.[133] Such an answer, however, can be generated by the harm theory. Under the harm theory, the question would be whether the circumstance of consent should be described as negating harm or as supplying a reason for acting in spite of the harm that occurs. If the former, then consent would be a harm-refinement justification and, under the harm theory, would be a defense even if the consent was not known by the actor. If the latter, consent would be a harm-offsetting justification and would be a defense only if the actor was aware of it. Although there is no universal answer to the question whether consent negates or offsets harm, distinctions between cases of the former and the latter can systematically be made. Sometimes consent is given because one desires what is consented to. For example, a person might consent to sexual intercourse because of physical attraction to the other party. Sometimes, however, consent is given despite one's antipathy to what is consented to. For example, a person might consent to sexual intercourse because of the payment of a significant sum. In these latter cases, what is prima facie undesirable becomes acceptable based on overriding factors. Likewise, inflicting the unwanted, but consented-to, state of affairs on another is a reason not to act, but a reason that is overridden by the existence of consent. Thus, if a man has sex with a woman who wished to have sex with him, he should not be convicted of rape even if he was unaware of her desire. Because no harm was done, he should be liable for attempted rape at most. In contrast, consider the case of a man who has sex with a woman believing her to have submitted to his threats. In fact, she has engaged in sex, despite her abhorrence of prostitution, because a third party has offered her drugs to which she is severely addicted. The man should have no defense to rape. He has caused the prostitute to undergo a significantly distasteful (even if ultimately accept-

132 *Id.* § 7.4, at 568.

133 *Id.*

able) experience, and he should be punished in light of its occurrence. Thus, a merit of the harm theory is that it presents systematic methodology for resolving hard cases concerning when a knowing justification is needed.

C. *Knowledge vs. Justificatory Intent vs. Regard*

Under the harm theory, as well as any other subjective or mixed theory of legal justification, the question arises: what mental state must be possessed by an actor whose conduct is justified in order for that actor to qualify for a justification defense? The Model Penal Code would allow an objectively justified actor to qualify for a justification defense where the actor merely knows of the justifying circumstances.[134] Fletcher would insist on justificatory intent, that is, the purpose to bring about those conditions (e.g., the prevention of the assault or the saving of a life) that offset the harm associated with the offense.[135] My version of harm theory proposes an alternative to these traditional mental states: regard. Regard, defined more fully below, requires more than mere knowledge but less than justificatory intent. It strikes the right balance in determining which actors engaging in justified conduct should be exonerated.

On the one hand, mere knowledge of the justificatory circumstances should not suffice to avoid liability. If Jim sets fire to his neighbor's field with malicious intent, clearly it should not matter that Jim, just after setting the field aflame, happens to glance over his shoulder, sees an approaching forest fire in the distance, and realizes that the fire he has just set will create a town-saving firebreak. His possessing this bit of knowledge has no bearing on what punishment he deserves because he only acquired the knowledge after the critical event from the perspective of moral culpability: the choice to set the field on fire.[136] Imagine now that Jim happens to see the forest fire just before he acts but after he has made his choice to act. Jim possesses information that is morally relevant, but so what? Again, the information did not, in fact, enter into his decisionmaking process. For the purpose of assessing culpability, it is as if Jim had an unread newspaper in his back pocket reporting that such a fire was approaching. In both cases, the knowledge, though "in Jim's possession," did not enter into his

134 *See* MODEL PENAL CODE §§ 3.02–.07 (1962). The Model Penal Code formulates its requirement in terms of what the actor "believes" rather than knows in order to reach those cases where the actor's belief is false. *Id.*

135 FLETCHER, *supra* note 5, § 7.4, at 564–66.

136 *See generally* Michael S. Moore, *Choice, Character, and Excuse*, SOC., PHIL. & POL'Y, Spring 1990, at 29 (discussing centrality of choice for moral culpability).

decisionmaking process and so seems unconnected to the morally critical moment of Jim's choice. Finally, imagine that Jim knows of the approaching fire before he makes his decision to act, but Jim, in making his decision, gives no significance or relevance to this fact. That is, Jim disregards his awareness of the approaching fire and employs precisely the same reasoning and weighing of considerations employed in the previous hypotheticals. There should be no difference in outcome. The knowledge of the approaching fire cannot be exculpatory for Jim because it made no difference to him. Acting merely with knowledge of the justifying circumstances is not enough.

The conclusion that an actor should not be exculpated merely because she acted with knowledge of the justifying circumstances is consistent with general principles of fault that inform the criminal law. Where a person is held liable based on mere knowledge that a harm will occur, her fault lies in failing to properly take into account a morally relevant piece of information. For example, when a store owner sets fire to her store for the insurance and thereby causes the deaths of the employees inside, she can be criticized for not taking the employees' presence into account when deciding whether to burn down the store. Likewise, Jim, in setting the fire in the last scenario, knew he would be creating a firebreak, but this fact failed to play any role in his practical reasoning in deciding to set the fire. His fault is of a kind with the storekeeper's. Jim's knowledge of the justifying circumstances did not make a difference to him, and so it should not make a difference to our evaluation of him.[137]

On the other hand, full justificatory intent should not be required to avoid liability. The most cogent reason for this is not the one discussed by Robinson.[138] Robinson argues that if justificatory intent were required, a person motivated by the desire to do harm and aware of the existence of justifying circumstances would be deterred from acting, thus resulting in a loss to society.[139] For example, Tim, a malicious person who would like to destroy his neighbor's field, would not do so where so doing would save the town because he would know that he would be still liable for arson under a theory of justification requiring justificatory intent.[140] Robinson regards this conclusion as a strike against an intent requirement. That a particular version of re-

137 *See* Kenneth C. Simons, *Culpability and Retributive Theory: The Problem of Criminal Negligence*, 5 J. CONTEMP. LEGAL ISSUES 365, 390–92 (1994) (discussing the moral significance of a belief's role in an actor's practical reasoning and reaching similar conclusions to those in this Article).

138 *See* 2 ROBINSON, *supra* note 3, at 18.

139 *See* 2 *id.*

140 *See* 2 *id.* at 14–15, 18.

tributivism would be inconsistent with maximizing social welfare, however, is not a criticism of that version.[141] It is commonplace that many forms of retributivism place limits on punishment that bar socially optimal levels of deterrence from being achieved.[142] Where this happens, hardcore retributivists bite the bullet and argue that less-than-maximum social welfare is the cost of justice. Less dogmatic retributivists may limit the scope of their retributivism to situations where the cost is not too high. In neither case, however, is the form of retributivism itself rejected. At most, Robinson's hypothetical presents yet another case where the divergence of retributivism and utilitarianism would have to be acknowledged. Without such a divergence, no retributive theory would have any practical interest when compared to contending utilitarian theories. A limited divergence from utilitarianism is a point for, not against, a justificatory intent theory of justification.

A better argument against the requirement of justificatory intent may be formulated within the framework of retributivism. An actor who engages in justified conduct should not be required to act for the purpose of achieving the results underlying the justification defense. A lower standard should suffice. No common English term, however, quite captures the mental state meeting this standard. Let us say then that an actor "regards" the justificatory circumstances surrounding his conduct where (1) he believes justifying circumstances exist, and (2) based on his belief, he decides to engage in the conduct.[143] "Regard" may be thought of as shorthand for the more bulky phrase "takes into account in his reasoning process when deciding what to do" or "func-

141 *See* 2 *id.* at 18.

142 Prime examples are crimes such as property crimes where the gain to the perpetrator is comparable to the loss of the victim. Assuming a simple form of retributivism, according to which the harm caused places a ceiling on the penalty, if the harm deserved is equal to the harm caused and that some percentage of such crimes will not result in convictions, a rational non-risk-adverse person would not be deterred from committing the crime.

143 Regard may be compared to Alan Michaels's concept of acceptance. Michaels believes that persons who "accept" that their acts may constitute the actus reus of an offense should be treated like those who know that they do. With respect to conduct, Michaels defines "acceptance" as the state where (1) a person acts recklessly with respect to whether his act constituted such conduct, and (2) the person would have so acted had the person been aware that the conduct was of that nature. *See* Alan C. Michaels, *Acceptance: The Missing Mental State*, 71 S. CAL. L. REV. 953, 961 (1998). Acceptance is like regard in that it has (1) an epistemic component, and (2) a counterfactual component. The second component of regard, "'based on his belief, he decides to engage in the conduct'" is roughly equivalent to "if not for his belief, he would not have decided to engage in the conduct."

tions as a reason (even if not the sole reason) relevant to the actor's choice." Although regard includes purpose, purpose does not entail intent. Suppose, for example, that Mike dislikes Vincent, but restrains himself from attacking Vincent because he believes people are generally entitled to physically security. One day Mike sees that Vincent is about to assault Mary, whose fate Mike is completely indifferent to. Seizing the opportunity, Mike punches Vincent and prevents the assault. Here Mike has regarded the justifying circumstances—Vincent's imminent assault of Mary—for the circumstances made a difference to him. Mike's belief that Vincent was acting impermissibly eliminated the consideration (respect for Vincent's physically security) that had prevented Mike from acting on other occasions. Mike, however, did not act with the purpose to protect Mary. Protecting Mary was but a foreseeable side effect of his conduct. The regarding of justificatory circumstances, in the absence of justificatory intent, may thus involve the defeating of a reason that had constrained action and need not provide an affirmative reason for action.

If an actor, such as Mike, regards the justificatory circumstances surrounding his contemplated conduct, he should not be blamed or punished for choosing to engage in it. Critically, the actor has not chosen to engage in conduct understood to be morally impermissible. In deciding how to act, persons, in the first instance, must abide by the distinction between the permissible and the wrongful, the primary distinction demarcated by the criminal law's rules of conduct. If they have honored this distinction, the criminal law should not consider them blameworthy.[144]

Consider the case of Rich, who opens a business for the sole purpose of taking customers from Penny, a person he dislikes. Let us assume, as is generally thought, that this conduct is a morally permissible exercise of liberty despite the adverse effects on Penny. In deciding to open the business, Rich did not choose to cross the line between permissible and wrongful action; Penny, he realized, had no

144 An exception should apply where the actor's belief in the justifying circumstances is negligent or reckless and the offense he is charged with permits liability based on negligence or recklessness. *Cf.* MODEL PENAL CODE § 3.09(2) (1962) (reckless or negligent use of otherwise justifiable force). Thus, if Bernhard Goetz had unreasonably, but correctly, believed he was going to be attacked with deadly force and killed one of his attackers and had taken this fact into account in deciding to shoot, he should have been liable for negligent homicide. He intentionally killed a person and so should bear prima facie responsibility for that harm. Although he believed he was being attacked and acted based on this belief, the belief was unreasonable and so only partially excuses his conduct. *See* People v. Goetz, 497 N.E.2d 41, 43–44 (N.Y. 1986) (providing facts of the case).

right to be free of competition. Admittedly, by maliciously acting to cause economic harm to Penny, Rich deserves our disapproval. He, however, has not disrespected the rights of anyone and so does not deserve the qualitatively higher degree of approbation appropriate to a person who has chosen to violate the rights of others.[145] An actor, such as Mike, who takes into account the justifying circumstances when engaging in justified conduct is in the same situation as Rich. Although harm was intended, his action was only taken in light of the legitimate opportunity to cause that harm. Like Rich, he does not deserve to be punished.

Extending justification defenses to actors who merely regard the justifying circumstances is fully consistent with acknowledging the significant difference between merely regarding the justifying circumstances and acting with justificatory intent. Only in the latter case may an actor be a candidate for praise. The criminal law sets minimum standards for conduct which even those of the worst character are supposed to regard and respect. In contrast, acting with justificatory intent may reflect an above-average character deserving of praise. The doling out or withholding of praise is the responsibility of a wide range of social institutions, many of which are nongovernmental, informal, or discretionary in nature. In contrast, the criminal law is limited to the business of imposing punishment when punishment is just. The boundary of just punishment stops exactly at the point where a person, in causing harm, acts with regard for the justifying circumstances.

CONCLUSION

This Article has sought to examine a variety of approaches to the problem of unknowing justification and to advance one as superior. The approach advocated, the harm theory, in general, would not grant a defense to the actor who has engaged in harmful conduct that, unknown to him, was objectively justified. The harm theory is not logically contradictory. Defining "justified" to imply "permissible," and "permissible" to imply "not punishable," cannot solve the moral issue presented by the problem. By the same token, the harm theory is not logically compelled. The alternative position that the unknowingly justified actor is not liable does not lead to contradiction. The recursive structure of justification in criminal law enables cases of conflicting forces to be unproblematically analyzed as com-

145 Of course, as argued earlier, if the actor is not aware of facts that render him within his rights, he is due full approbation, and his punishment is to be measured by the harm he has caused. *See supra* notes 100–02 and accompanying text.

prised of justified and unjustified actual and potential forces. The harm theory, however, is morally compelling. When reasons not to act are ignored and these reasons are objectively manifested as harms, a punitive response is appropriate absent any excusing conditions.

The harm theory has three salient characteristics. First, it sharply distinguishes between rules of conduct (norms) on one hand and rules of evaluation (punishment conditions) on the other in order to allow a substantive divergence between them. Second, it rests on a concept, harm, that is a full-blooded moral concept on par with wrongdoing. Third, by focusing on harm, it admits the possibility of an unknowing justification defense in those limited cases where sanction for the offense is disproportionate to the harm actually caused by virtue of unknown harm-mitigating circumstances. By virtue of these characteristics, this Article argued that the harm theory yields a system of punishment that properly reflects the culpability of an actor for his conduct. One cannot expect to convince all of one's readers of all of one's conclusions all of the time. A more modest goal would be to establish that any coherent theory of justification that maintains liability for the unknowingly justified actor must incorporate the characteristics and paths of argument found in the harm theory and presented herein. It is hoped that this more modest goal has been fully achieved.