INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.
An essay on the criminal law justification defense

Nuttall, Steven Richard, Ph.D.

The Ohio State University, 1991

Copyright ©1991 by Nuttall, Steven Richard. All rights reserved.
AN ESSAY ON THE CRIMINAL LAW

JUSTIFICATION DEFENSE

DISSERTATION

Presented in Partial Fulfillment of the Requirements

for the Degree Doctor of Philosophy in the

Graduate School of The Ohio State University

By

Steven Richard Nuttall, B.S., M.A., J.D.

* * * * *

The Ohio State University

1991

Dissertation Committee:  Approved by
Donald Hubin
Bernard Rosen  Adviser
James Scanlan  Department of Philosophy
Copyright by

Steven Richard Nuttall

1991
To Christine, Melissa,

and Johnathan
# VITA

<table>
<thead>
<tr>
<th>Year</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>March 12, 1950</td>
<td>Born--Olney, Illinois</td>
</tr>
<tr>
<td>1972</td>
<td>B.S., Eastern Illinois University, Charleston Illinois</td>
</tr>
<tr>
<td>1974</td>
<td>M.A., Northern Illinois University, DeKalb Illinois</td>
</tr>
<tr>
<td>1980</td>
<td>J.D., The Ohio State University, Columbus Ohio</td>
</tr>
<tr>
<td>1981-Present</td>
<td>Assistant Professor St. Bonaventure University St. Bonaventure New York</td>
</tr>
</tbody>
</table>

## PUBLICATIONS


## FIELDS OF STUDY

Major Field: Philosophy of Law
# TABLE OF CONTENTS

DEDICATION ................................................................................................................... ii  
VITA ................................................................................................................................... iii  
LIST OF FIGURES ........................................................................................................ vi

CHAPTER  | PAGE
---------|--------
I.  |  
1.1. Introduction ............................................................................................... 1  
1.2. Types of Theories of Justification ................................................................. 4  
1.3. A Taxonomy of Criminal Law Defenses ......................................................... 7

II.  |  
2.1. Introduction ................................................................................................. 30  
2.2. The Nature of an Act .................................................................................... 31  
2.3. Act vs. Agent Evaluation .............................................................................. 53

III. |  
3.1. Introduction ................................................................................................. 72  
3.2. Justification and the Harm Requirement ....................................................... 72  
3.3. Societal Harm and Criminal Attempts .......................................................... 94  
3.4. Objective Harm and Self-defense ..................................................................... 107  
3.5. Justification and Excuse ................................................................................ 115  
3.6. The Case Against an Objective Harm Requirement ....................................... 127  
3.6.1. The Privilege and Exception Arguments .................................................... 127  
3.6.2. Justification and Act Evaluation .................................................................. 130

IV.  |  
4.1. Character Traits and Excuses ......................................................................... 151  
4.2. Moral Blameworthiness and Motivation........................................................ 173
4.3. Character Traits as Motivational Dispositions ........................................185
4.4. Moral Blameworthiness and Voluntariness .........................................204
4.5. Moral Blameworthiness and Criminal Punishment .............................205

V.

5.1. A Motivational Theory of Justification .............................................226
5.2. Incrimination versus Justification ....................................................235
5.3. The Serial View of Defenses .............................................................249
5.4. Justification and Universalization ....................................................255
5.5. Justification or Excuse: Sending the Wrong Message? .......................257
5.6. Self-defense ......................................................................................258
5.7. Conclusion .......................................................................................260

BIBLIOGRAPHY ..........................................................................................268
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURES</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Simple act-tree diagrams of level-generationally related act-tokens</td>
<td>43</td>
</tr>
<tr>
<td>2. A more complex act-tree diagram</td>
<td>44</td>
</tr>
<tr>
<td>3. Act-tree diagram with co-temporal act-tokens</td>
<td>45</td>
</tr>
<tr>
<td>4. Act-tree diagram with same-level act-tokens</td>
<td>45</td>
</tr>
</tbody>
</table>
Chapter I

1.1. Introduction

Joe closes a switch on a high-voltage electrical circuit, resulting in the electrocution of a person strapped in a chair. Joe performs this gruesome task in his capacity as state executioner. Has Joe committed a homicide? Harold is an inmate in a state prison. A fire breaks out in the cell block in which Harold is housed, threatening Harold’s life. Harold bolts. Is Harold guilty of the crime of escape? While shopping in the local drug store, a man suffers a severe heart attack. Unless given a dose of nitroglycerine, the man is likely to die. The druggist, realizing that there is no time to call a doctor, administers the drug without a prescription. Has the druggist committed the crime of dispensing a prescription drug without a prescription? As a result of an automobile accident, a passenger in Monica’s car suffers serious cuts and bruises. Realizing that the injuries require immediate medical attention, Monica speeds off to the nearest hospital emergency room. A criminal statute requires that anyone in a car accident that results in physical injuries to another must either wait for the police to arrive at the scene, or if no officer is around, report the accident as soon as physically possible. Has Monica committed the crime of hit-and-run? A fire is raging out of control in a community of closely packed
dwellings. All residents have been evacuated. Arnold deliberately blows up several houses in the path of the oncoming fire in an effort to prevent its further advance into the community. Is Arnold guilty of the crime of arson? Nick, employed as a night watchman, is stopped by two youths on his way to his car early one morning. They demand that he hand over his wallet, which he promptly does. Then one of the youths says to the other, "We don't need a witness, do we?" and pulls a knife. At this point, Nick, who has a license to carry a hand gun, pulls his revolver and fatally shoots the would-be attacker. Is Nick guilty of homicide?

Most people will not hesitate to answer that the law ought not regard any of the above acts as criminal, despite the fact that a literal reading of applicable statutes would appear to make the relevant conduct criminal. The criminal law reflects this common sense moral judgment by allowing each of the above defendants successfully to plead that their conduct was, under the circumstances, justified. The types of cases in which the plea of justification may be accepted range from defense of self (Nick), others, or property, to choosing violation of an express criminal prohibition as the lesser of two evils under the circumstances (Harold, the druggist, Monica, and Arnold), to acting pursuant to a grant of public authority (Joe). Although there would be considerable scholarly agreement that each of the above defendants has a valid defense and that the appropriate label to attach to that defense is "justification," at a deeper level one would find considerable disagreement concerning precisely what is meant by saying of each case that the conduct is justified, why such conduct should not be subject to criminal liability, and how justification defenses are to be distinguished
from other important categories of criminal law defenses.

This "deep level" disagreement manifests itself in various ways. Although there is a core of cases (represented by those above) that by general consensus raise valid justification defenses, as one departs from these paradigms problematic cases arise where theoretical differences give rise to disagreements over whether a valid defense exists at all, and if so, whether it is properly classified as a justification defense. The disagreements concerning validity, given the obvious implications for criminal liability, are disturbing. One might think that the classification disagreements are of no practical consequence, since as long as a valid defense to liability is recognized, the result for the defendant is the same: freedom. This would be a mistake. First, important substantive and procedural issues affecting liability (some having constitutional significance) have been alleged to turn on the proper classification of admittedly valid defenses. Second, the issue of the defendant's freedom may turn on the category of defense that grounds the "not guilty" verdict. Some theorists maintain that a successful excuse always indicates that the defendant suffers from some kind of weakness or defect. In extreme cases, insanity for example, freedom may be lost despite a not guilty verdict. If, however, the basis of the verdict is a justification defense, the defendant's freedom is assured.

One of the primary sources of disagreement at the "deep level" is, I think, a disagreement over the basic "logic" of criminal law justifications and other closely related defenses. How precisely do justifications, excuses,
and other criminal law defenses operate to defeat criminal liability? This dissertation is an effort to clarify the logic of criminal law justification defenses. Since it is illuminating to contrast how justification defenses operate with the operation of other closely related defenses, the logic of other defenses will also be explored. In addition, to the extent that there are parallels with how moral defenses operate, considerable light is shed on the logic of moral defenses as well.

Describing the investigation as one concerned with the "logic" of justification defenses suggests that, to the extent possible, the account strives to be independent of any particular normative moral theory. This is indeed the goal. But achieving this goal does not preclude appeals to normative argument, counter-examples that rely upon common sense moral judgment, etc. For example, one of the theories of justification that we shall examine at length holds that justification defenses function to negate "criminal harm." A perfectly reasonable strategy to evaluate this account of the logic of justification would be to construct paradigm cases in which natural language speakers (and members of our moral community) would judge an otherwise criminal act to be "justified," even though criminal harm is not negated. With these preliminary comments in mind, we turn to a simple schema for classifying theories of the logic of justification defenses.

1.2. Types of Theories of Justification

Theories of justification may be usefully divided into objective, subjective, and hybrid theories, depending on the nature of the analysis offered of the validity conditions of a justification defense. By the "validity
conditions" of a defense I mean the conditions necessary and sufficient for the defense to defeat criminal liability. An *objective* theory is one that excludes mental states of the accused from the validity conditions of a justification defense. A *subjective* theory is one that takes the validity conditions of a justification defense to include only mental states of the accused (typically, beliefs or motives). A *hybrid* theory, as the name suggests, includes among the validity conditions of a justification defense some mental states of the accused, but some non-mental state conditions as well.

Below are some formulas that express the validity conditions of a justification defense for four sample theories of justification:

(1.1) $J(da) \leftrightarrow O(c) \& \neg H(ac)$

(1.2) $J(da) \leftrightarrow B(d)[O(c) \& \neg H(ac)]$

(1.3) $J(da) \leftrightarrow B(d)[O(c) \& \neg H(ac)] \& O(c) \& \neg H(ac)$

(1.4) $J(da) \leftrightarrow RB(d)[O(c) \& \neg H(ac)]$

According to (1.1), defendant $d$ is justified in performing act $a$ if and only if circumstances $c$ obtain such that the performance of $a$ in $c$ brings about no net harm. If we assume an objective account of harm, a theory that adopts (1.1) as an analysis of the validity conditions of a justification defense constitutes an objective theory of justification.\(^{11}\) (1.2), by contrast, represents a subjective theory. According to (1.2), $d$ is justified in performing $a$ if and only if $d$ believes that circumstances $c$ obtain such that the performance of $a$ in $c$ will bring about no net harm.\(^{12}\) (1.3), a hybrid analysis, simply fuses (1.1) and (1.2); so that not only must $d$ believe that $c$
obtain such that the performance of \( a \) in \( c \) will bring about no net harm, but \( c \) must obtain and it must be true that the performance of \( a \) in \( c \) will bring about no net harm.\(^{13}\) A theory adopting (1.4) would also be a hybrid theory. (1.4) states that \( d \) is justified in performing \( a \) if and only if \( d \) reasonably believes that \( c \) obtain such that the performance of \( a \) in \( c \) will bring about no net harm.\(^{14}\) This is a hybrid analysis, since it requires not only that \( d \) believe something, but that the belief be reasonable. An assessment of the reasonableness of a belief requires appeal to a normative standard (e.g., what a hypothetical "reasonable person" would believe in \( d \)'s situation). Since on this view whether an accused is justified is not a function solely of the mental states of the accused, the analysis is not a "subjective" one (as we have defined that term).

A theory of justification may also be classified as either strong or weak, depending on whether it accepts the following "incompatibility thesis":

\[
(IT) \text{ If two acts are incompatible, then if the performance of one is justified, the performance of the other cannot be justified.}
\]

We may say that two acts are "incompatible" if the successful performance of one precludes the successful performance of the other. Strong theories of justification accept (IT); weak theories do not.

The theory of justification to be defended in this dissertation is a weak hybrid theory. In brief, the proposal is that in a properly constructed criminal justice system justification defenses defeat a normally warranted inference from criminally proscribed conduct to substandard moral
motivation (i.e., a motivational fault). In thus construing justifications, the proposal departs from the recent trend among contemporary legal theorists of regarding justifications as essentially involving an "objective" evaluation of conduct. Extreme objectivists view an agent's motivation as irrelevant to the validity of justification defenses. Other theorists add a motivational, or "justificatory intent," requirement to what is essentially an objective theory of justification, in an effort to avoid potential counter-examples to a purely objective account. Unfortunately, absent a theory of criminal responsibility that provides a supporting rationale for its inclusion, the additional psychological requirement often appears ad hoc. The approach adopted here is to provide the necessary supporting rationale by regarding the criminal law as a motivational supplement to moral motivation. One implication of the motivational theory is that the logic of the criminal law defenses of justification and excuse is much closer than typically acknowledged.

1.3. A Taxonomy of Criminal Law Defenses

In this section I present a preliminary taxonomy for classifying criminal law defenses. Such a taxonomy is important for at least three reasons. Liability consequences, especially for intervening third parties, have been alleged to turn on whether a given defense is classified as a justification or some other defense. Constitutional implications, such as the mandated allocation of the burden of proof (risk of non-persuasion), have also been alleged to follow from such a classification decision. Finally,
as with most anything, one can better understand the nature of justification defenses by contrasting them with what they are not, but might easily be taken to be. The following taxonomy is offered as an initial step in clarifying the logic of justifications and their relation to other defenses.

The taxonomy offered here is essentially that proposed by Paul Robinson. We begin by observing that criminal law defenses are responses to a criminal "charge." The charge is the vehicle through which a prosecuting authority formally attributes criminal wrongdoing to a defendant, and so it is to the charge that the defendant must successfully respond to avoid criminal liability. Let us call a response to a criminal charge that enables the defendant to avoid liability a "valid defense." Valid defenses avoid liability in a number of distinguishable ways. Our task in this section is to identify the various ways in which criminal defenses "negate" liability and to group those that negate liability in the same way under a common label. We start by surveying the landscape in order to gain some preliminary idea of where the boundaries are to be drawn and how each type of defense functions to negate criminal liability.

The content of a modern day criminal charge is typically derived from a criminal statute defining an offense. Let us suppose that we live in a society which defines "stealing" or "theft" as follows:

A person commits theft when he knowingly:
(a) Obtains or exerts unauthorized control over property of the owner; . . . , and
(1) Intends to deprive the owner permanently of the use or benefit of the property . . . .
Now consider some of the defensive responses that might be made to the charge, "Felix stole Harold's umbrella."

D1: "No, it was Felix's twin brother Oscar who did it. Felix was visiting Disney World with me at the time."

D2: "Felix took the umbrella, yes, but he did not steal it. He made a mistake. He thought the umbrella was his when he took it."

D3: "No [Yes, but], Felix was sleep-walking at the time he took the umbrella."

D4: "Well, he certainly did take it, but the umbrella was in such poor shape (the fabric was rotting; it was badly torn) that its loss was of no consequence to Harold. In fact, Felix overheard Harold say that he intended to throw it away as soon as it was convenient to do so. Surely the taking of such a worthless article ought not constitute theft."

D5: "Yes. But he did so in order to save Harold's life. Arch Enemy had placed an explosive in the handle while Harold was not looking."

D6: "Well, yes he did. But it was pouring down rain, Felix had a terrible cold, and (not anticipating the rain) he failed to bring an umbrella with him."

D7: "Well I suppose he did. But Felix was so drunk at the time that the whole episode seemed unreal. His memory of the incident is very sketchy."

D8: "Yes, but Felix is only four years old. He did not really appreciate what he was doing."

D9: "Yes, but that happened ten years ago. It's too late to do anything about it now."

Of course this is just a partial list of defensive responses that might be given to such a charge. Nevertheless, the list contains examples of five
categories of defense, which categories exhaust the domain of defenses. The five categories are: denial, offense modification, justification, excuse, and nonexculpatory public policy defense.\(^{18}\)

Perhaps the most obvious way of defending against the charge that Felix stole Harold's umbrella is simply to deny that Felix did what he is expressly charged with doing (steal the umbrella). D1 does this by providing an alibi and attributing the theft to an agent other than Felix. D2 also denies the charge by pointing to a factual mistake in order to negate the required mental element of intent. In both of these responses the accused (Felix) says, in effect, "I did not X," where X is conduct expressly attributed to him (via the formal charge), and defined in a criminal statute. For obvious reasons we will refer to this type of defense as a "denial."\(^{19}\)

D3 could be viewed as a denial as well, but this depends, among other things, upon the "phenomenology" of sleep-walking. If the sleep-walking Felix was "going through the motions" with little or no awareness of what was going on, then an appeal to his sleep-walking state might be taken to deny presence of the relevant mental state (intention to deprive . . .). On the other hand, it may be that Felix had the relevant intent, but that the intent was not connected (causally or in some other way) with Felix's bodily movements or with other psychological states so as to have its normal criminal significance. In an extreme case of this sort it may be appropriate to deny that Felix stole the umbrella on the grounds that some psychological condition necessary for human action is absent. In such a case, as in the case where Felix is merely "going through the motions" without the required intent, D3 operates as a denial.\(^{20}\)
Some defendants who have engaged in conduct that they are expressly charged with doing should not be held liable to criminal punishment. This is because criminal charges typically do not (for reasons yet to be discussed) include all of the necessary conditions of criminal liability. But even if one were to formulate a charge so that it included all of the necessary conditions of liability, there might be good reasons for not holding a defendant liable in some cases where the defendant has engaged in the charged conduct. This would be possible if satisfaction of the entire set of necessary conditions of liability need not be sufficient for liability.

We have, then, in addition to defensive responses of the form, "I did not X," defensive responses of the form, "I did X, but . . . ." Defenses of the latter type grant (for the sake of argument) the doing of X, but go on to cite some reason for not holding the defendant liable nonetheless. There are four different versions of the "I did X, but . . ." defensive response.

One is illustrated by D4. The claim in D4 seems to be that even though on the face of it Felix’s conduct meets the statutory definition of theft, it is unreasonable to hold that Felix’s conduct constitutes the kind of taking that the legislature was trying to address when it passed the theft statute. The defense offered on behalf of Felix, sometimes called a "de minimis infraction" defense, is an instance of the kind of defense that Paul Robinson calls "offense modification." According to Robinson, offense modifications function to clarify the nature of the harm or evil sought to be prevented by the statute defining the charged offense. As such they, like denials, represent "criminalization" decisions (in contrast to
the "exculpation" decisions represented by justifications and excuses, and are typically "offense-specific," rather than being general defenses that apply regardless of the specific nature of the offense (again, in contrast to justifications and excuses).

Another example of an offense modification is "Wharton's Rule." Under a typical definition, it is sufficient for commission of the crime of conspiracy that two "guilty minds" enter into an agreement to engage in conduct that constitutes a crime (although many jurisdictions would add an "overt act" requirement). The rationale underlying conspiracy statutes is, in part, a perceived need to address the special dangers raised by group criminal activity (i.e. organized crime). Wharton's Rule bars conviction for conspiracy, despite an agreement between two individuals to engage in such criminal conduct as dueling, bigamy, adultery, or incest. Since two guilty minds are already required for commission of such offenses, it is thought that any special dangers incident to such two-person group activity is already addressed by the primary offense and its associated penalties. A conspiracy charge would, in those cases, be unnecessary. Since Wharton's Rule serves to clarify the nature of the harm or evil that conspiracy statutes are intended to prevent, it is classified as an offense modification. Other examples of offense modifications offered by Robinson are renunciation of criminal purpose (in attempt, solicitation, and conspiracy cases), impossibility (in attempt cases), and consent.

It would seem that whether a given defense is classified as a denial or offense modification depends upon how explicit the legislature is when defining the relevant offense. Theft could be defined so that any property
taken must have a certain minimum value, or must be believed by the
defendant to have a certain minimum value. In that event, D4 would be
construed as a denial of one of the definitional elements of theft. Similarly,
conspiracy could be expressly defined so as to exclude dueling, adultery,
bigamy, and incest. Responding to a conspiracy charge by pointing out that
the alleged "conspiracy" involved an agreement to duel would then be
treated as a denial.

The question of proper classification becomes a bit more complicated
when one introduces the possibility of creative judicial construction of
statutory language. It would not take a lot of ingenuity to argue that the
theft statute as it stands requires that the defendant believe that the object
stolen has some minimum value to its owner. The language "intends to
deprive the owner permanently of the use or benefit of the property" might
be so construed. Given the dubious value of the umbrella and Harold's
expressed intention to discard it, Felix may very well have lacked the belief
that in taking the umbrella he would be depriving Harold of its use or
benefit. What this example suggests is that often facts that support an
offense modification may also provide the basis for denying that an explicit
definitional element, properly interpreted, has been satisfied. Thus,
classification of a defense may turn on whether a court ties the defense to
specific language of the statute that forms the legal basis of the charge.

Both denials and offense modifications represent judgments that what
the defendant did is not within the intended scope of a given criminal
statute. In the case of denials, this judgment may be pinned on some
specific language of the statute that forms the legal basis of the charge. In the case of offense modifications, it cannot. Both types of defense, however, rest on an interpretation of legislative intent. In light of these observations, it may come as a surprise to hear that important constitutional consequences turn on whether a defense is classified as a denial or as an offense modification. Nevertheless, the argument has been made that the constitutional allocation of the burden of proof depends on just such a classification decision.32

D5 differs in an important respect from D4. In order to appreciate the difference, it is necessary to distinguish statutory harm or evil from nonstatutory harm or evil, and prima facie from on balance wrongdoing. The distinctions are intuitively straightforward. When a criminal statute is initially drafted, the legislature has in mind some specific harm or evil that it is attempting to address through the mechanism of punishment.33 For example, a homicide statute is obviously concerned with reducing the number of deaths brought about by the intentional, reckless, or negligent conduct of others. A conspiracy statute is concerned with the special harms or evils incident to organized group activity. Call the specific harm or evil that a statute is designed to address the "statutory harm or evil." Clearly, there will be other types of harm or evil not intended to be addressed by a given statute. Homicide statutes are not concerned with invasions of property interests or infringements of sexual autonomy, for example. Call these "nonstatutory harms or evils."

Another common distinction to make is that between prima facie and on balance wrongdoing. Conduct may bring about some harm or evil, and
by virtue of doing so be considered *prima facie* wrongful. But if it was necessary to avoid or eliminate some even greater harm or evil, it may not be on balance wrongful. Medical surgery to remove a cancerous breast may cause pain and disfigurement, but if it saves someone’s life it is not on balance wrongful. The same may be said of conduct that violates a criminal statute. If I intentionally kill an aggressor in order to protect my family from a life-threatening attack, there is little doubt that I have caused the kind of statutory harm or evil contemplated by a relevant homicide statute. By virtue of doing so, my conduct is *prima facie* wrongful. Nevertheless, if the intentional killing was necessary to save several innocent lives, it is not on balance wrongful.  

The difference between D4 and D5 must now be identified. D4 maintains that the statutory harm or evil contemplated by the theft statute was not produced by Felix’s taking the umbrella. D5, on the other hand, maintains that even if Felix did bring about precisely the kind of statutory harm or evil that motivated the theft statute, and did, for this reason, engage in *prima facie* wrongdoing, his doing so in this particular instance prevented a much greater harm or evil than that occasioned by the theft. We might say in such a case that the conduct was warranted or *justified* even though it did violate the express terms of the theft statute and did cause statutory harm or evil.

All three types of defense discussed so far (denials, offense modifications, and justifications) respond to a criminal charge by denying that what the defendant did may properly be described as (on balance)
"criminal wrongdoing," although they do so in different ways. Denials deny the content of the criminal charge. Under the "principle of legality" there can be no criminal wrongdoing absent a criminal statute expressly forbidding the conduct in question.\textsuperscript{35} If the defendant’s conduct does not satisfy the description contained in some such statute, no crime has even been \textit{prima facie} committed. Offense modifications also deny that the defendant has engaged in conduct that is criminally wrongful, but they do so by pointing out that, despite the fact that conduct satisfies the express description contained in a criminal statute, the legislature never intended to criminalize this particular type of conduct, for it does not involve the kind of harm or evil the statute was designed to address. Justifications do not deny that the defendant has engaged in conduct that involves statutory harm or evil (either because there is no statute or because the harm or evil produced was not contemplated by a relevant statute). Instead they raise considerations designed to show that the conduct is not \textit{on balance} wrongful by pointing out that under the circumstances the defendant’s conduct was necessary to prevent an even greater harm or evil, or was for some other reason warranted.\textsuperscript{36}

A fourth category of defense is represented by D6-D8. These defenses do not deny that the charged act was performed or that the conduct produced the statutory harm or evil, nor do they point to circumstances that arguably justify the act. In short, they do not deny that Felix engaged in (on balance) criminal wrongdoing. They seek, rather, to defeat the propriety of holding Felix \textit{accountable} or \textit{responsible} for wrongdoing.\textsuperscript{37} They point to conditions which are designed to \textit{excuse} Felix from punishment (and the
consequent condemnation) for any alleged wrongdoing. It is clear that the author of D6 is not denying that Felix stole the umbrella or that the legislature intended the taking of umbrellas (even torn and tattered ones that the owner intends to discard) to be addressed by the theft statute. The claim being made could be that the fact that Felix had a cold and had no umbrella with him justifies his theft. But that seems dubious. More likely the claim is that Felix ought not be held accountable or responsible for the theft, given his fear that getting wet might further complicate his illness. Likewise, in D7 the author is not denying that Felix stole the umbrella or that tattered-umbrella-stealing is a harm or evil intended to be addressed by the statute. And certainly the claim is not that Felix's intoxicated state justifies the theft. The claim is (rightly or wrongly) that we ought not regard Felix as accountable for the theft. Equally clearly, the citing of Felix's tender age in D8 is not intended to deny (in any of the three ways distinguished) the criminal wrongfulness of Felix's conduct. Rather, the age is cited in order to excuse Felix from punishment (and condemnation) for his allegedly wrongful act.38

D9 represents yet a fourth way of completing the response, "I did X, but . . . ." This last category of defense stands apart from all four of the categories discussed above in an important respect. One who raises the statute of limitations as a bar to conviction is not denying that the defendant is accountable for criminal wrongdoing. The author of D9 is not denying that Felix stole the umbrella or that tattered-umbrella-stealing is a target of the theft statute, nor is the claim that the passage of time somehow converts
an earlier wrongful act into a justified act, or that Felix ought not be deemed accountable for the earlier theft. Rather, the appeal is to public policy considerations that bar conviction despite the propriety of holding Felix accountable for criminal wrongdoing.\textsuperscript{39}

Here it may be valuable to introduce a distinction between criminal guilt, in the sense of accountability for criminal wrongdoing, from liability to punishment.\textsuperscript{40} Nonexculpatory public policy defenses,\textsuperscript{41} such as an appeal to the statute of limitations, defeat liability to punishment without denying criminal guilt in the above sense.\textsuperscript{42} It is precisely because such defenses do not negate criminal guilt, in the accountable-for-wrongdoing sense, that their successful employment is frequently met by public outrage and a charge that the defendant "got off on a legal technicality."\textsuperscript{43}

It is important to realize that a single defense label may in fact designate defensive responses in more than one of the above categories. Intoxication, for example, may be the basis of a denial, as where the charge is theft and the defendant was so drunk at the time of the alleged offense as to lack comprehension of the nature of the act being performed and, as a result, to lack the intent requisite for the crime. However, in a case where the defendant had the relevant intent, but lacked the ability meaningfully to control behavior as a result of intoxication, the drunken state may function as an excuse. Under the label, "insanity defense," mental illness may likewise function to deny a required mental element of a charged offense. The classic textbook example is the defendant who suffered from a delusion that she was squeezing a lemon when in reality she was squeezing her husband's neck. If the charge were murder (i.e. "purposely or knowingly
causing the death of another human being"), such a defendant could cite the delusion as grounds for a denial that the death of another was caused "purposely or knowingly." But even if the relevant mental state were present at the time of the act, so that a defensive denial is foreclosed, the defendant's mental illness may form the basis of an excuse, again referred to as an "insanity defense." Suppose the defendant realized that she was killing another human being, but erroneously believed, as a result of mental illness, that this particular human being was the reincarnation of Hitler and that she had been instructed by God to seek atonement for Hitler's past sins. We might decide that even though she unjustifiably murdered Hitler II, it would be improper to hold her accountable for her action and so permit an insanity excuse. As George Fletcher and others have observed, the label "necessity" has been used to designate both a defensive justification (sometimes called "lesser evils") and a defensive excuse (sometimes called "duress" or "coercion"). One must be careful, therefore, to look behind the labels when attempting to classify a given defense.

Having laid some necessary groundwork in this chapter, we turn next to a presentation of a theory of human action that will be relied upon in subsequent portions of the dissertation. It is commonly assumed that criminal liability requires performance of a criminal act. Although this is not an issue of primary concern in this dissertation, it would seem that any reasonable assessment of an "act requirement" would require an understanding of the nature of an act. Of more immediate relevance for our purposes is the claim that the two primary categories of exculpating
defenses, justification and excuse, are to be distinguished on the grounds that the former involves a kind of act evaluation, while the latter involves an evaluation not of an act, but of an agent who has performed an act. Typically conjoined with this claim are statements that act evaluations (and hence justifications) are universalizable, while agent evaluations (and hence excuses) are not. In order to assess these claims we need a theory of human action.
ENDNOTES

1"(1) A person is guilty of criminal homicide if he purposely, knowingly, recklessly or negligently causes the death of another human being." American Law Institute, Model Penal Code § 210.1 (1985) [hereinafter cited as MPC].

2"(1) Escape. A person commits an offense if he unlawfully removes himself from official detention or fails to return to official detention following temporary leave granted for a specific purpose or limited period." MPC § 242.6.

3"1. No drug for which a prescription is required . . . shall be distributed or dispensed to any person except upon a prescription written by a person legally authorized to issue such prescription." NY Educ § 6810. Violation of this section is a class A misdemeanor. NY Educ § 6811.19.

4"2.a. Any person operating a motor vehicle who, knowing or having cause to know that personal injury has been caused to another person, due to an incident involving the motor vehicle operated by such person shall, before leaving the place where the said personal injury occurred, stop, exhibit his license and insurance identification card . . . to a police officer, or in the event that no police officer is in the vicinity of the place of said injury, then, he shall report said incident as soon as physically able to the nearest police station or judicial officer.

The first violation of the provisions of paragraph a of this subdivision shall constitute a Class B misdemeanor . . . ." NY Veh. & Tr. § 600.

5"(1) Arson. A person is guilty of arson, a felony of the second degree, if he starts a fire or causes an explosion with the purpose of:

(a) destroying a building or occupied structure of another . . . ." MPC § 220.1.
For a typical homicide statute, see supra note 1.

P. Robinson, Criminal Law Defenses: A Systematic Analysis, 82 Colum. L. Rev. 199, 234, 257 (1982) (constitutionally mandated allocation of the burden of persuasion may depend on classification of defense as "failure of proof" or "offense modification") [hereinafter cited as Analysis]; P. Robinson, A Theory of Justification: Societal Harm as a Prerequisite for Criminal Liability, 23 UCLA L. Rev. 266, 279 (1975) (liability of accomplice depends on classification of defense of principal as "justification" or "excuse") [hereinafter cited as Justification].

See, for example, K. Greenawalt, Distinguishing Justifications from Excuses 49 Law and Contemporary Problems 89 (1986).

If the verdict is not guilty by reason of insanity (or some roughly equivalent special verdict), the defendant will be subject to automatic commitment (in most states) for psychiatric evaluation, with an eye toward more permanent commitment (typically employing civil commitment standards).

States include events and processes. As will be made clear later, I do not subscribe to the view that mental states must be "private" states (events or processes) to which only the person whose states they are has privileged access. In this regard the use of "objective" and "subjective" has the potential to mislead. But I do think that many legal scholars assume that mental states are private, and that to some extent this motivates attempts to offer "objective" accounts of criminal law justification defenses. Whatever one's conception of the nature of mental states may be, an objective theory denies the relevance of such states to an account of the logic of criminal law justification.

See, e.g., P. Robinson, Justification.

The Model Penal Code is sometimes cited as an example of such a subjective theory. See L. Katz, Bad Acts and Guilty Minds 39 (1987). But the accuracy of this attribution may be questioned, since the Code denies a justification defense to an actor who "was reckless or negligent in bringing about the situation requiring a choice of harms or evils or in appraising the necessity for his conduct," where the offense charged is one "for which recklessness or negligence, as the case may be, suffices to establish culpability." MPC § 3.02(2).
This may be George Fletcher's theory, although he would prefer the term 'wrongdoing' to 'harm'. See G. Fletcher, Rethinking Criminal Law 564-65 (1978) [hereinafter cited as Rethinking]; G. Fletcher, The Right Deed for the Wrong Reason: A Reply to Mr. Robinson, 23 UCLA L. Rev. 293, 298, 320-21 (1975).

See W. LaFave and A. Scott, Jr., Handbook on Criminal Law 386 (1972) [hereinafter cited as LaFave and Scott].

Supra note 7.


This is part of the definition of the consolidated theft crime, as defined by the Ill. Crim. Code of 1961, cited in Kadish and Paulsen, Criminal Law and Its Processes (3d ed. 1975) at 725.


Robinson uses the expression "failure of proof defense" to refer to instances of this first type of defense. Defenses 72. Unfortunately, this expression misleadingly suggests that the defense essentially involves an evidentiary challenge. And this suggests that any successful challenge to the sufficiency of the evidence on some issue with respect to which the state bears the burden of proof would count as a failure of proof defense. This is not, however, the intended interpretation. The intent, rather, is to define a class of defenses that deny that a statutory definitional element of the charged offense was satisfied by the defendant's conduct, regardless of who bears the burden of proof on that element. So understood, failure of proof defenses raise a substantive issue, not an evidentiary one. There is reason to think that states are constitutionally required to bear the burden of proof on all statutory definitional elements. But they may choose to bear the burden on other issues as well. Should a state fail to satisfy the burden of proof on one of these nondefinitional issues, this failure does not provide the grounds for a Robinsonian failure of proof defense. Since the language, "failure of proof," suggests that it might, I prefer to avoid the language, and with it the potential for confusion. Robinson is aware of the potential for confusion embodied in the "failure of proof" label. He suggests the less confusing, but cumbersome, "absence of an element" as an alternative. Defenses 72 n.1.
Another possibility is that cases of somnambulism should be regarded as involving defects of perception akin to that accompanying certain instances of intoxication, and so should be classified as excuses, rather than as denials. See P. Robinson, *Analysis* 223. (Actually, Robinson would regard the "no act" version of somnambulism as an excuse as well, *Analysis* 222, 242.)

See the discussion of the distinction between incrimination and justification, *infra* 235.

Some theorists hold that liability ascriptions are "defeasible" in the sense that they may be "defeated" even though all conditions necessary and normally sufficient for liability have been met. See H.L.A. Hart, *The Ascription of Responsibility and Rights*, 49 Proceedings of the Aristotelian Society 171, 175 (1948/49); J. Feinberg, *Doing and Deserving* 122 (1970) [hereinafter cited as Deserving].

See MPC § 2.12.

*Analysis* 208-9, 242.

Robinson's use of the expression 'harm or evil' might be thought to be a concession to those who hold that the criminal law may legitimately proscribe some nonharmful wrongs. But in fact Robinson thinks that the only relevant criminal evil is harm, or harmful conduct. Robinson, *Analysis* 209, 220. As I shall use 'harm or evil,' the expression may designate not only harm, or harmful acts, but also nonharmful wrongs. At least I do not wish to rule out the possibility that the criminal law may be properly concerned with some nonharmful forms of conduct.

Robinson, *Analysis* 209, 220. The suggestion that one can meaningfully distinguish criminalization from exculpation is an intriguing one. We shall have more to say about it *infra* at 235.


See MPC § 5.03(1) and (5); Lafave and Scott 476.

Robinson, *Analysis* 242. Although Robinson classifies consent as an offense modification, there is reason to think that it may not always operate in this way. Sometimes it may operate as a justification. See G. Fletcher, *Rethinking* 557, 566, 769.
One might even reach the Wharton’s Rule results via a statutory construction route. This could be achieved by interpreting the "two guilty minds" requirement as tantamount to requiring at least one more guilty mind than is required by the substantive offense that is the subject of the agreement. Since dueling requires for its commission an agreement between two minds to duel, a conspiracy to duel would, on the proposed construction, require an agreement among a minimum of three minds.

The issue for classification purposes is whether the defense is tied to language of the statute that forms the legal basis of the charge. A defense may be expressly recognized by some statutory section other than the one that forms the basis of the charge. Such a defense might be classified as an offense modification, but it would not be a denial. An example of this would be a statutory scheme that proscribed conspiracy in one section and then, in a separate section, provided for a defense in cases of conspiracy to duel, etc., where only the minimum number of persons required to duel, etc., were involved in the conspiracy.

See Robinson, Analysis 234, 257.

Sometimes it is difficult to say precisely what this evil is. Even the legislators themselves may disagree. Often the legislation that finally results is a compromise among competing views of the nature of the harm or evil and how it should be addressed. See the discussion of the Mann Act in E. Levi, An Introduction to Legal Reasoning 30-40 (1949). This uncertainty does not, however, threaten the distinction between the statutory harm or evil (however precisely defined) and other types of harm or evil which clearly were not the subject of a given statute.

It is important to require that the conduct be necessary to avoid the greater evil. Wrongdoing is not a simple function of the overall goodness or badness of the consequences of an individual act. Suppose, for example, that it were fairly easy to disable the aggressor without killing him. In that case killing him may still result in a net savings of lives, but there was another act that would have been even more life-preserving, and which, presumably, should have been performed. Despite its overall good consequences, we might declare killing the aggressor wrongful.

See LaFave and Scott 177.

The reason for the qualification is to allow for the possibility of deontological justifications, as well as teleological justifications based upon
the production of some *positive good* sufficient to outweigh the harm caused.

I do not mean to suggest that these alternative ways of describing the connection between wrongdoing and agent, if the agent is to be properly held criminally liable for wrongdoing, are necessarily equivalent. At this stage of the inquiry I wish to leave that issue open. There is, however, a certain ambiguity that infects the expressions, ‘holding accountable’ and ‘holding responsible,’ that I think should be made explicit at this juncture. It is useful to invoke an analogy first introduced by Joel Feinberg. J. Feinberg, *Deserving* 124-25. Suppose we imagine a kind of moral accounting system that makes use of a personal ledger upon which entries are made, based (perhaps among other things) upon the moral qualities of an agent’s conduct. We might then distinguish what kind of entry should be made (debit, credit, zero) from whether any entry should be made at all. Suppose an agent intentionally caused great harm (e.g. death) to another. Normally this act would warrant a rather sizable debit entry on the agent’s moral ledger. But now suppose we are told that the agent was in a severe psychotic state at the time and that he sincerely (but mistakenly) believed that the victim was about to murder his entire family. Given this additional information, we might decide that it would be inappropriate to make any entry at all on the moral ledger. We could describe the situation where no entry is appropriate as one in which the agent is "not accountable," or ought not be "held accountable" for his conduct. This situation should be contrasted with that in which a non-psychotic person kills an aggressor in self-defense. In the latter case an entry is appropriate. It’s just that the appropriate entry to make is not the one typically made when one person intentionally kills another. Depending upon one’s moral assessment of the act, either a credit or a zero should be entered instead of a debit.

The questions of whether an entry should be made, i.e., whether one may properly hold an agent accountable in the first sense for the corresponding act, as well as what kind of entry to make, should both be distinguished from the question of to what use the ledger or an entry therein may be put. Should we, for example, use the ledger as a basis for praising, blaming, or punishing the agent? A second sense of ‘hold accountable’ is that which refers to one of these uses to which the ledger or an entry may be put. Obviously, there may be reasons for not holding someone accountable in the second sense, even when they may properly be held accountable in the first. Suppose you discover that last week Jane lost your two thousand dollar laptop computer, and that her doing so was the result of gross negligence. She is accountable for the loss, and may be held accountable in
the sense of having a debit entered on her moral ledger. But suppose further that since losing your computer Jane has suffered one catastrophe after another and is in such a fragile state that one more negative happening (e.g. criticism from you) will almost certainly put her "over the edge." Under the circumstances, even though she may in the first sense be held accountable (an entry in her ledger is proper), you should not hold her accountable in the sense of overtly blaming her for losing the computer.

In order to avoid confusion I propose to restrict the use of 'hold accountable' to the first sense. When I wish to refer to some use to which a moral ledger or entry may be put, I will use the expression 'liable to X,' where X designates the use in question (e.g. praise, blame, punishment). Thus, in Jane's case I would say that Jane is accountable for losing your computer, may properly be held accountable for losing it (an entry may be made), but is not (or ought not be) liable to blame (or other overt moral criticism) for losing it.

Given the potential for confusion, one might legitimately wonder why I do not simply jettison the use of 'hold accountable' altogether in favor of 'is accountable' and 'is liable.' The reason for not following this course is that the decision of whether to make an entry on an agent's moral ledger is itself a normative one. The normative character of the decision is transparent when one asks whether one ought to hold an agent accountable, as opposed to simply asking whether the agent is accountable. Similar remarks apply to 'is liable' and 'ought to be held liable.' For convenience I will often use the shorter expressions, but one should not forget that the decision whether an agent is accountable or liable is a normative one.

An additional complication should be mentioned here. Some theorists would regard infancy as either a radically different kind of excuse from the "normal" excuse, or would refuse to classify it as an excuse at all, regarding it instead as an "exemption" or jurisdictional challenge. See M. Zimmerman, An Essay on Moral Responsibility 61-62, 66, 76, 108 (1988); Fletcher, Rethinking 836.

Some of the policy considerations that have been offered to support a statute of limitations defense are that prosecutions should be based upon reasonably fresh evidence, a person ought not have to live indefinitely with a fear of prosecution, the community's retributive impulse diminishes over time, while the likelihood that the offender has reformed increases. Robinson, Analysis 230.
I realize that it is common to use the expression "criminal guilt" to include all conditions required for punishability. This usage coincides with the general verdict of "not guilty" in criminal trials. "Not guilty" here simply means "not convicted" or "not punishable." But this usage (as well as the general verdict upon which it is based) glosses over several important distinctions. When a defendant is found "not guilty" in the broad, or inclusive, sense we do not know whether the verdict is based upon a successful denial, offense modification, justification, excuse, or nonexculpatory public policy defense. The end result typically is the same regardless of the basis of the verdict: the defendant walks. (There are exceptions, e.g. where a special verdict is employed--the "not guilty by reason of insanity" verdict being one such example. Here we know--roughly at least--the basis of the nonconviction. Moreover, in the case of the insanity verdict, the defendant ordinarily will not simply walk, but will be committed for a period of observation and evaluation with an eye towards a more prolonged commitment to a secure psychiatric facility.)

I agree with Robinson's suggestion that a greater reliance on special verdicts may be desirable. See Robinson, Analysis 247, 290. I realize that many view the general verdict as an important device for checking the abuse of government power and tempering the rigidity of legal rules. But when a defendant has engaged in criminal wrongdoing and yet is deemed not punishable because some nonexculpating public policy is implicated, I think it is important for the public to be aware of the grounds for nonconviction. Moreover, it may be reasonable to attach to this kind of nonconviction some of the collateral consequences that frequently accompany conviction. See Robinson, Analysis 285-291. At any rate, the distinction between defenses that defeat accountability for criminal wrongdoing and those that do not is sufficiently important to be marked in some clear way. The proposed narrow definition of "criminal guilt" is one way of doing that.

The term is Robinson's. Analysis 229-232.

Other examples of nonexculpatory public policy defenses are: double jeopardy, diplomatic immunity, testimonial immunity, plea-bargained immunity, judicial, legislative, and executive immunities, incompetency, amnesia, dismissals based upon exclusionary rule or prosecutorial misconduct, and entrapment. Robinson, Analysis 243. As these examples illustrate, a defendant who pleads one of these defenses is not alleging non-accountability for criminal wrongdoing, but rather that for public policy reasons punishment ought not be imposed despite accountability.
Douglas Husak refers to nonexculpatory public policy defenses as "procedural defenses," in contrast with all other defenses, which he calls "substantive defenses." D. Husak, Philosophy of Criminal Law 187-88 (1987) [hereinafter cited as Philosophy].

This is substantially the Model Penal Code definition. MPC §§ 210.1 and 210.2. In addition, under the MPC definition a person may be guilty of murder if the person recklessly causes death "under circumstances manifesting extreme indifference to the value of human life." MPC § 210.2(1)(b). I have omitted this element from the above definition in order to simplify discussion.

G. Fletcher, The Individualization of Excusing Conditions, 47 S. Calif. L. Rev. 1269, 1273-1280 (1974); G. Fletcher, Rethinking 818.
Chapter II

2.1. Introduction

The main task of this chapter will be to articulate a theory of human action that is not only plausible, but also useful for facilitating the discussion of many issues of moral and criminal theory. The model to be presented is thoroughly deterministic, or is at least compatible with a deterministic view of human nature. It is far from uncontroversial, and I will be able to provide only a brief sketch of some of the major features. Its utility for such investigations as that before us certainly must be weighed in any overall attempt to assess the merits of the model.¹

As was mentioned at the end of the previous chapter, a common way of distinguishing criminal law justifications from criminal law excuses is to maintain that the former involve evaluations of conduct, while the latter involve evaluations of persons or agents. We shall, therefore, take a brief look at the end of this chapter at the utility and tenability of the distinction between act evaluation and agent evaluation. The theory of human action sketched in the first part of the chapter will help guide this discussion. We turn first, then, to a theory of human action.
2.2. The Nature of an Act

Following Alvin Goldman's lead, I propose to adopt a property criterion for individuating human actions, or to be more precise, for individuating human act-tokens. According to this view, an act-token is the exemplification of an act-type (property) by a person at a time (or during a time interval). John's act of turning on the light is to be analyzed as the exemplification by John of the act type, turning on the light, at time \( t \).\(^2\)

It follows from this property criterion of act individuation that John's act of turning on the light at \( t(1) \) is a different act from his turning on the light at \( t(2) \). So also is John's act of turning on the light at \( t(1) \) different from Harold's act of turning on the light at \( t(1) \). These two consequences of the analysis seem relatively uncontroversial. A more controversial consequence is that John's act of (A) flipping the light switch at \( t(1) \) is a different act (token) from that of (A') John's turning on the light at \( t(1) \)--and this even if John's flipping the switch at \( t(1) \) is the cause of the light's going on at \( t(1) \). The latter consequence follows from the analysis in virtue of the fact that the act-type, flipping the light switch, is not identical with the act-type, turning on the light. Other act-tokens that are not identical in virtue of being exemplifications of different act properties are (B) John's extending his arm out the car window at \( t(1) \) and (B') John's signaling a left turn at \( t(1) \); (C) John's jumping 6 feet 3 inches at \( t(1) \) and (C') John's outjumping George at \( t(1) \) (assuming that George has just jumped six feet); (D) John's saying "hello" at \( t(1) \) and (D') John's saying "hello" loudly at \( t(1) \); (E) John's hitting the tallest man in the room at \( t(1) \) and (E') John's hitting the wealthiest man in the room at \( t(1) \)--even if the tallest man in the room is
the wealthiest man in the room.

Some theorists prefer a less "fine-grained" analysis of acts, according to which the two acts in some or all of the above act-pairs A-A', B-B', C-C', D-D', and E-E' are identical. Goldman does give some rather impressive arguments for preferring the fine-grained analysis. But equally impressive responses have been given to those arguments. I do not intend to adjudicate the merits of the debate here. Nevertheless, it will be useful, in order to introduce the important concept of level-generation, to restate one of the arguments Goldman presents against the thesis that the relation holding between the acts in all of the above act-pairs is that of identity (the "identity thesis").

Note that in the case of each of the first three pairs of acts one can correctly say that the second act is performed "by" performing the first. John's turning on the light is performed by his flipping the switch. His signaling a left turn is performed by his extending his arm out the car window. And, given the circumstance that George has just jumped six feet, John's outjumping George is performed by his jumping 6 feet 3 inches. This "by" locution denotes an asymmetric, irreflexive, and transitive relation. John's turning on the light is performed by his flipping the switch, but his flipping the switch is not performed by his turning on the light. John does not outjump George by outjumping George. Finally, if John signals a left turn by extending his arm out the car window, and he alerts another motorist of his intent to turn left by signaling a left turn, then he also alerts another motorist of his intent by extending his arm out the car window. The
fact that the constituent acts of each of these act-pairs are related in a manner expressible via the "by" locution constitutes one good argument against their being identical, for if two objects are identical, there can be no relation holding between them which is asymmetric or irreflexive. Of course it is incumbent upon one who rejects the identity thesis to indicate how the constituent acts of each of the above act-pairs (and other similar act-pairs) are related. For Goldman, the appropriate relation in the case of the first four pairs (A-A', B-B', C-C', and D-D') is that of "level-generation"; in the case of the last pair (E-E') it is that of "same-levelness." A brief explanation of each of these concepts is, therefore, in order.  

In order to explicate the concept of level-generation, Goldman employs a set of temporal conditions on generation sans phrase, together with a characterization of four types of dependency relation holding between "generational acts." These dependency relations distinguish one species of level-generation from another. The first temporal condition is that "neither one of a pair of generational acts is subsequent to the other." The test for nonsubsequence is the inappropriateness of the "and then" locution. It would not be appropriate to say that John flipped the switch and then turned on the light; nor would it be correct to say that John extended his arm out the car window and then signaled a left turn; and so on for the other generational pairs above.  

Given this temporal condition, one might say (and Goldman does) that in a sense pairs of generational acts are always done at the same time (i.e. during the same interval of time). Although this is a convenient way of speaking, it is potentially misleading. For if "at the same time" is taken
to imply that both members of a generational act-pair span precisely the same interval of time, I think that there is reason to doubt the claim that generational acts always occur at the same time.

Consider the act-tokens of "John’s moving his index finger," "John’s squeezing a trigger," "John’s firing a gun," and "John’s killing Harold." I have intentionally omitted the temporal references in order to make a point. Clearly, these four acts are just the sorts of acts that Goldman would say are generationally related.10 They are four act-tokens of the same agent. The "by" locution holds. John performed the act of killing Harold by firing a gun, and he fired a gun by pulling a trigger, and he pulled a trigger by moving his index finger. None of the acts is subsequent to any of the others. John did not move his index finger and then pull the trigger and then fire the gun and then kill Harold. But there does seem to be at least a prima facie problem with saying that all of the acts were performed during precisely the same interval of time. John’s moving his index finger does seem to span exactly the same time interval as his pulling a trigger, and although it is not quite so obvious (there may be a "split second" between the squeezing of the trigger and the gun’s going off), perhaps it is reasonable to say that John’s firing a gun spans the same time interval as well. But what about his killing Harold? Suppose the bullet fired from John’s gun does indeed find a resting place in Harold’s abdomen. But further suppose that Harold is stubborn and refuses to die for several days. Unhappily (for Harold), on the fifth day he does die. When (during what interval of time) did the act of John’s killing Harold occur? Should we say,
as Goldman seems to suggest, that this act spans exactly the same interval as John’s moving his finger? But how could it? After all, Harold was still living well after John moved his finger.11

By way of response, one might propose that we distinguish the ontological question of when (during what interval of time) an act occurs from the epistemic question of when we come to know that the act has occurred. In the case of John’s killing Harold, we may know right away that John has moved his finger, has squeezed a trigger, and has fired a gun. But it may be some time before we come to realize that during the same interval of time John also performed the act of killing (i.e. causing the death of) Harold. We won’t know whether during this interval John performed the act of killing Harold until we discover that Harold died some five days later. On this view, during any given interval of time there may be a great number of acts that we are performing without at that time knowing it because such knowledge must await the occurrence of future events (e.g., the death of Harold).

Unfortunately, the proposed distinction does not really address the objection. It is not simply that we do not know before Harold’s death whether John has performed the act of killing Harold, and must await later developments to find out. The act of killing Harold cannot occur prior to Harold’s death.

Perhaps the most obvious alternative answer to when (during what time interval) the act of killing Harold occurred is that it occurred continuously throughout the interval beginning with the first movement of John’s finger and ending with the death of Harold. This option would still
allow one to claim that both acts of a generational act-pair occur "during the same interval," but the interval during which both acts are performed would be that of the longer (in case there is a longer) of the two generational acts. It would no longer be implied that generational acts must span exactly the same interval; one of the acts may occur during an interval less than, but included within, that spanned by the other. This suggestion is not, however, without its problems.

There seem to be some counter-intuitive consequences of allowing an act to continue beyond a point in time when the generating bodily movement has ceased. Suppose, for example, that immediately after firing the gun John dies (from a heart attack induced by shock at having realized what he had just done). Are we to say that John’s act of killing Harold is not consummated until Harold dies, so that an act of John’s continues to exist after his own demise? A view that allows acts to continue in existence after their agent’s death may appear to some to be metaphysically suspect.

It might go some way toward mitigating the oddness of this second option to point out that some very respectable acts require no bodily movement at all. Consider the act of sunbathing, for example. The simple fact that under the second option acts may continue after the generating bodily movement has ceased is not, therefore, a good reason for rejecting that option. Moreover, it is possible, through the vehicle of writing a will, for example, for an agent to exercise control over future events in much the same manner as if he were alive. A will can be written so that the vesting of property interests is contingent upon certain future events happening or not
happening. For example, one can provide that a niece receive a substantial sum of money only if and when she receives a college diploma. One can also take away property if certain events occur or fail to occur. One can, for example, provide that one's spouse lose property left to him if he should remarry. It is not uncommon in such cases for people to speak as if the deceased's actions continue into the present, or will continue into the future. In response to the questions, "Who is preventing you from getting your money?" and "Who will take your property if you remarry?" it would not seem inappropriate to reply, "Uncle Ned," and "My ex-wife." Now of course those who would so respond may not take themselves to be speaking literally. But I think it fair to say that countenancing acts that continue long after the generating bodily movement has ceased would not do violence to our manner of speaking about acts. And given what to me are the only obvious choices for temporally locating acts such as that of killing George, I think it reasonable to prefer the second option. In any event, I think that one must be wary of taking too literally Goldman's claim that generational acts always occur "at the same time." Having said all of this, I should add that for Goldman's purpose of distinguishing level-generation from other relations among acts, he need not claim that generationally related acts span exactly the same time interval. It is enough that neither act be subsequent to the other. In general I would prefer to paraphrase "acts A and B occur at the same time" as "neither act A nor act B is subsequent to the other," so as to avoid even the potential for misrepresenting the temporal relation involved.

The second temporal condition that Goldman places upon level-generation is that "no member of a level-generational pair [may] be a
temporal part, i.e. proper part, of its level-generational mate." The example used to illustrate this requirement is the act of playing the C-scale. This act is composed of a series of smaller acts (playing the individual notes of the scale), each of which is a temporal part of the longer act of playing the C-scale. None of these shorter constituent acts is generationally related to the act of playing the C-scale.

But, to return to our example of John's killing Harold by moving his finger, etc., why isn't the act of moving his finger during the interval t(1)-t(2) a temporal part of John's act of killing Harold during the interval t(1)-t(3)? I think that the answer lies in an unstated assumption (or stipulation) on Goldman's part that every "temporal part" of an act must itself be an act. Furthermore, in order for act A to be a "proper part" of act A' the subtraction of A from A' must leave a remainder that is itself an act. The act of playing the C-scale is divisible into a number of smaller acts (=temporal parts) and if any one of these acts were omitted the remainder would be an act (the act of playing a certain sequence of notes). By way of contrast, if one were to subtract from the act of John's killing Harold the act of John's moving his index finger, the remainder would not be an act. Presumably, we are to imagine a gun firing or a bullet being propelled from the barrel without John (or anyone else) squeezing the trigger, Harold being struck by the bullet, and Harold dying some five days later. The resulting event or series of events would be a (rather surprising) happening, but it would not be an act.
Although one might question Goldman's assumption that only acts may be temporal parts of acts (Why isn't the event of Harold being struck by a bullet a temporal part of the act of killing Harold? Why isn't the event of the gun's firing a temporal part of the act of firing the gun?), I think that Goldman may justifiably ignore any putative non-act temporal parts of acts, given the task at hand, viz., that of distinguishing level-generation from other types of relation holding among acts with which level-generation may otherwise be confused.\footnote{14} Level-generation (as Goldman conceives it) is not a cause and effect relation. The nonsubsequence requirement insures that level-generation is not confused with any such relation. Level-generation is not a part-whole relation. The requirement that neither member of a generational act-pair be a "proper temporal part" of the other is designed to guarantee that level-generation is distinguished from part-whole relations. Since the task of distinguishing level-generation from part-whole relations is one of distinguishing different types of relations holding among acts, one can reasonably ignore, for the purpose of performing this task, any possible non-act constituents of acts.

In addition to requiring that neither member of a generational act-pair be subsequent to or be a proper temporal part of the other, we must require that generational acts not be "co-temporal" acts--the criterion of co-temporality being the correctness of saying that one co-temporal act is done "while also" doing the other.\footnote{15} Although two co-temporal acts may be performed by the same agent during the same interval of time, neither is "dependent" upon the other. An example of a pair of co-temporal acts would be John's rubbing his stomach \textit{while also} patting his head. It seems
clear that when one act level-generates another, the performance of the second in some sense depends on the performance of the first.

A pair of generational acts, then, is a pair of acts performed by the same person, neither of which is subsequent to the other, neither of which is a proper temporal part of the other, and one of which depends for its performance on the performance of the other. Different species of level-generation, correspond to different ways in which the performance of one act of a generational pair "depends" on the performance of the other.

In the case where John's flipping the light switch generates his turning on the light, John's first act causes an effect, E (the light's going on), and the second act consists in John's causing that effect. Since the performance of the generated act depends on the generating act's having a certain causal effect, Goldman refers to this species of generation as "causal generation."^16

A different kind of dependency relation obtains between the act of extending one's arm out the car window and the act of signaling a left turn. The generated act (signaling) does not consist in John's causing some effect, E, by extending his arm out the window. John's extending his arm out the car window may cause another driver to believe that John is signaling, and this in turn may cause that driver to believe that John is about to turn left. But John's act of signaling a left turn consists in neither his causing the other driver to believe that he is signaling nor in his causing the other driver to believe that he is about to turn left. The other driver may not see John extend his arm, and hence may not form either of the above beliefs
and yet John may signal a left turn all the same. What is characteristic of the species of generation involved in the signaling example is that there exists a set of rules, or conventions, in virtue of which one act (extending one's arm out the car window) performed under specified circumstances (e.g., while driving), is sufficient for the performance of another act (signaling a left turn). This species of level-generation Goldman calls "conventional generation."^18

The third type of level-generation is illustrated by the example of John's jumping 6 feet 3 inches generating his outjumping George. As with conventional generation, John's jumping 6 feet 3 inches does not cause some event, E, the causing of which constitutes his outjumping George. Moreover, there is in this case no need for a rule, or set of rules in order for the generation to occur. John's jumping 6 feet 3 inches, given the circumstance that George has just jumped 6 feet, by itself is sufficient for John's having outjumped George. "Simple generation," as Goldman calls this type of level-generation, "is like conventional generation minus the rules."^19

We turn now to an examination of the fourth pair of generational acts described above. This pair of acts illustrates what Goldman refers to as "augmentation generation," for in such cases the generated act is generated by "augmenting" the generating act with some relevant fact or circumstance.20 John's saying "hello" loudly is formed by augmenting John's saying "hello" by the manner in which he says it--loudly. Other examples of augmentation generation are John's extending his arm generating John's extending his arm out the car window (the augmenting
fact being that the arm went out the car window); John's running generating John's running at 8 m.p.h. (the augmenting fact being how fast John was running); and John's shooting a basketball generating John's shooting a jump shot (the augmenting circumstance being that John was jumping while shooting).\(^{21}\) The distinctive feature of augmentation generation is that the performance of the generated act (more accurately, a statement describing such performance) logically entails the performance of the generating act (or a statement describing such performance), while the performance of the generating act does not entail the performance of the generated act.\(^{22}\)

It should be acknowledged that augmentation generation does not fit the level-generation model as comfortably as one might like. For one thing, the "by" locution test seems to fail in the case of augmentation generation. It does not seem quite correct to speak of saying "hello" loudly by saying hello, or of running 8 m.p.h. by running, or of jump-shooting by jumping (or by shooting). On the other hand, acts related by augmentation generation do satisfy the temporal conditions placed upon level-generation. Neither is subsequent to the other (one does not say "hello" and then say "hello" loudly), neither is a temporal part of the other (if one subtracts the act of saying "hello" from the act of saying "hello" loudly, one does not have an act as a remainder), and the two acts are not co-temporal (one does not say "hello" while also saying "hello" loudly). Moreover, the performance of the generated act depends upon the performance of the generating act in the sense that if the generating act did not occur, the generated act would not occur. If John did not say "hello," he would not say "hello" loudly. If he
did not run, he would not run 8 m.p.h., etc. It seems reasonable, therefore, despite the failure of the "by" locution test, to regard the relation of augmentation as a species of level-generation.\textsuperscript{23}

It is useful to have some way of diagraming the structural relationships that hold among the act-tokens of a single agent. The device employed by Goldman for this purpose is the act-tree diagram. Using circles for individual act-tokens and vertical or diagonal lines connecting such circles for the relationship of level-generation, diagrams of some of the simple examples of level-generation just discussed follow. The letters "C," "CV," "S," and "A" located beside the connecting lines represent the types of generation involved: causal, conventional, simple, and augmentation. In order to simplify the diagrams, I have omitted the agent and temporal references.

![Figure 1: Simple act-tree diagrams of level-generationally related act-tokens](image)

Although I will not explore the various possibilities here, much more complex relationships can be represented using this device. Consider the following example.\textsuperscript{24}
Breaking a long-standing tradition

Declining the nomination for vice-president

Indicating refusal

Moving his head from side-to-side

Moving his head

Disappointing his followers

Upsetting his glasses

Co-temporal acts may also appear on an act-tree diagram, but there will be no line connecting them directly, for they are not generationally related. They may, however, each generate a higher level "compound act." Consider the jump shot example (where the concentric circle represents a compound act).
Having discussed both generational and co-temporal act pairs and indicated how they would be diagramed, let us address the concept of "same-levelness." Two act-tokens are on the same level if they differ only in having different individual concepts of the same object. For example, John’s act of hitting the tallest man in the room (at t) is on the same level as his act of hitting the wealthiest man in the room (at t), if the tallest man in the room is the wealthiest man in the room. Neither of these acts generates the other, since level-generation is an asymmetric relation, and any relation between these acts would appear to be symmetric. It does, however, seem reasonable to stipulate that any act that generates one same-level act also generates the other, and that any act generated by one same-level act is also generated by the other. Same-level acts will be represented on act-tree diagrams by circles connected with a horizontal line. Here is a partial diagram containing two same-level act pairs.

Figure 3: Act-tree diagram with co-temporal act-tokens

Figure 4: Act-tree diagram with same-level act-tokens
The act-tree diagraming technique suggests that any such tree will have a starting point, or set of starting points: acts that generate higher level acts, but are not themselves generated by any lower level acts. Such acts might be called "basic acts"--the analogue in action theory to sense data in theories of perception. If there are such basic acts, how might they be defined? Goldman suggests that one begin by defining the concept of a "basic act-type," for basic act-tokens would seem to be instances of basic act-types.28

The fundamental feature of a basic act-type (or property) is that it is exemplifiable at will.29 Some non-controversial examples are:

Extending one’s arm
Moving one’s finger
Bending one’s knee
Shrugging one’s shoulder
Opening one’s eyes
Turning one’s head
Puckering one’s lips
Wrinkling one’s nose

Of course one must add the proviso that there are no external forces preventing the agent from exemplifying the property in question. If John’s hand is tied down, he will not be able to raise his hand even if he wants to. And of course he will not be able to raise his hand if it is already raised. Goldman proposes to deal with such complications by stipulating that "A is a basic act-type for S at t only if: if S were in standard conditions with respect to A at t, then if S wanted to exemplify A at t, S’s exemplifying A at t would result from this want."31 He then defines ‘standard conditions’ in the obvious fashion: "S is in standard conditions with respect to property A
at t just in case (a) there are no external forces at t making it physically impossible for S to exemplify A at t, and (b) if exemplifying A involves a change into state Z, then S is not already in Z. He further stipulates that in applying the above characterization of a basic act-type, one must assume that the agent has no stronger competing wants. Otherwise, even though the agent wants to exemplify A at t, other competing wants may prevent his doing so.

Notice that the concept of a basic act-type is relativized to individual agents and that the definition of 'standard conditions' is such that conditions "internal" to an agent are not included. Thus, a person paralyzed from the waist up may not have as a basic act-type "turning one's head." Of course the person may later acquire that basic act-type if the paralysis is successfully treated.

Notice also that the above characterization of a basic act-type is formulated as a necessary condition only. There is good reason for this. It seems that there are properties that some agents can exemplify at will that are not appropriately classified as basic act-types. With practice, some people can imitate famous personalities at will. Others can make themselves see double. Still others can cause themselves to vomit by placing a finger down their throat. Goldman attempts to deal with these three cases (and other similar ones) by noting that in each case the exemplification of the relevant property requires either generational or causal knowledge (or belief). Exemplification of the relevant property does not result directly (or simply) from wanting to exemplify the property. The
want is mediated by some intervening generational or causal knowledge or belief. The ability to imitate famous personalities at will requires knowledge of how to talk (what kinds of noises and gestures are required) in order to level-generate the act of imitating a given personality. Seeing double requires knowledge of how to press one’s eyeball in order to causally generate the act of making oneself see double. Vomiting at will requires that one know that placing one’s finger down one’s throat will cause one to vomit.34

Contrast these cases with examples from our earlier list of basic act-types. No generational or causal knowledge is required (except possibly the knowledge that the act will result from the want) in order to exemplify the property of extending one’s hand. One need only want to do so and, assuming one is in standard conditions, one does so. Similarly, no generational or causal knowledge is required in order for one to turn one’s head or shrug one’s shoulders. Simply wanting to do so (in standard conditions) will guarantee that the act occur. With the above considerations in mind, Goldman offers the following as a definition of ‘basic act-type’:

Property A is a basic act-type for S at t if and only if:

(a) If S were in standard conditions with respect to A at t, then if S wanted to exemplify A at t [and S had no stronger competing wants at t], S’s exemplifying A at t would result from this want; and

(b) the fact expressed by (a) does not depend on S’s level-generational knowledge nor on S’s cause-and-effect knowledge, except possibly the knowledge that his exemplifying A would be caused by his want.35
Having defined 'basic act-type,' it is tempting to define 'basic act-token' as any instance of a basic act-type. However, such a definition would be too broad. Although every act-token is an instance of a basic act-type, the converse is not true. The property, "coughing," for example, may be exemplifiable at will, but it is sometimes exemplified regardless of one's will— as when one coughs spontaneously in response to a certain tickling sensation in the throat. Even a basic act-type such as "raising one's right leg" may have instances that are not basic act-tokens. Suppose I raise my right leg by raising it with my right hand. In that case, the act-token of raising my right leg is level-generated by the basic act-token of raising my right hand. In order to narrow the definition of 'basic act-token' sufficiently to rule out these and similar property-exemplifications, Goldman adds two additional necessary conditions to that of being an instance of a basic act-type. The resulting definition is:

S's exemplifying property A (at t) is a basic act-token if and only if

(a) property A is a basic act-type for S (at t);
(b) S's exemplifying A (at t) is caused, in the characteristic way, by an action-plan of S; and
(c) S's exemplifying A (at t) is not level-generated by anything else satisfying clauses (a) and (b), except possibly by augmentation generation.

Some explanatory comments are in order. First, an action-plan is a combination of an agent's wants and a projected act-tree, where a projected act-tree is a set of beliefs of an agent regarding how various potential actions are generationally (or otherwise) related, which beliefs may be
represented via an act-tree diagram. The requirement that such an action-plan cause the exemplification of A "in the characteristic way" is included to preclude from the category of basic act-tokens exemplifications caused indirectly by action-plans. Suppose, for example, that while at a dinner engagement I decide to offend my host by grimacing in response to tasting the dinner wine. I have an action-plan that includes my desire to offend my host and a belief that I can do so by grimacing at the appropriate time. Further suppose that my practical joker friend, Oscar, learns of my plan and decides to prevent me from intentionally offending my host. Oscar, without my knowledge, places some foul-tasting substance in my wine glass. When I sip the wine I do in fact grimace in response to the foul taste and do in fact thereby offend my host. Nevertheless, my grimacing is not a basic act-token in this case, for although my action-plan (through the agency of my friend) did cause the grimacing, it did not do so directly (i.e. in the characteristic way).\textsuperscript{38}

The caveat contained in (c) regarding augmentation generation is included because otherwise the definition would exclude such act-tokens as "saying 'hello' loudly" and "walking quickly." However, such exemplifications do seem to belong in the category of basic act-tokens.

Having briefly discussed the nature of level-generation and same-levelness, as well as what it is to be a basic act-token, we are in a position to provide a comprehensive definition of 'act-token'. The definition is a recursive one that builds upon the notion of a basic act-token:

(1) If A is a basic act-token, then A is an act-token.
(2) If A is on the same level as an act-token, then A is an act-token.
(3) If A is level-generated by an act-token, then A is an act-token.
(4) If A is a temporal part of an act-token, then A is an act-token.
(5) If A is a temporal sequence of act-tokens, then A is an act-token.
(6) Nothing else is an act-token.

Since the concept of intentional action is of such great importance to the criminal law, we should not leave this discussion of Goldman's theory without an indication of how he defines an intentional act. Oversimplifying a bit, the intuitive idea is that all and only acts caused in the characteristic way by an action-plan and that occur in the manner conceived in that plan are intentional acts. A more formal analysis might go as follows:

Suppose S has an action-plan which includes acts A(1), A(2), A(3), ..., A(n), where A(1) is a basic act and n > 1. S wants to do A(n), and S believes (to some degree) of each of the acts A(1), A(2), A(3), ..., A(n) firstly, that it will either be generated by A(1) or be on the same level as A(1), and secondly, that it will either generate A(n) or be on the same level as A(n). If this action-plan, in a certain characteristic way, causes S's doing A(1), then A(1) is intentional. And if some of the other acts A(2), A(3), ..., A(n) are performed in the way conceived in the action-plan, then these acts are also intentional. All other acts on the (actual) act-tree are non-intentional.

A consequence of this definition, in conjunction with the definition of a basic act-token, is that every basic act-token is an intentional act. To see this, recall that a necessary condition for a property exemplification to be a basic act-token is that it be caused in the characteristic way by an action-plan. But under the above analysis of intentional action, being
caused in the characteristic way by an action-plan of an agent is a sufficient condition for a basic act-token’s being intentional. Hence, all basic act-tokens are intentional acts.

An interesting feature of the Goldman model of human action is that it regards all actions as ultimately caused by wants and beliefs of an agent. This follows from the recursive definition of ‘act-token’ and an intuitively plausible causal principle. The principle is that a cause of act A is (at least a partial) cause of every act generated by A, of every temporal part of A, and of every act of which A is a temporal part. Suppose, for example, that Golda wants to impress her music instructor and believes that flawlessly playing the C-scale on the piano will do so. She has an action-plan consisting of the predominate desire to flawlessly play the C-scale and a projected act-tree representable by an act-tree diagram. At the bottom of the diagram are a series of movements of various fingers of Golda’s right hand. Each of these movements level-generates the playing of an individual note of the scale. The sequence of these notes constitutes the act of playing the C-scale. Now let us suppose that this action-plan causes Golda, in the characteristic way, to flawlessly play the C-scale on the piano. Furthermore, let us suppose that in playing the C-scale Golda awakens her pet cat Murdock—a completely unforeseen consequence—and that Murdock (without Golda’s knowledge) had earlier destroyed Golda’s favorite stuffed elephant.

Now each basic act on the actual act-tree is (by definition) caused by a set of wants and beliefs of Golda (viz., an action-plan). Furthermore, each
level-generated act of playing a note of the scale is caused by the same action-plan that causes each generating basic act-token. The sequence of act-tokens which constitutes the act of playing the C-scale is also caused by the action-plan that in turn causes the playing of the individual notes. Even level-generated unintentional acts are caused by the action-plan, as in the act of awakening Murdock, and so are caused by wants and beliefs that cause other intentional acts. And finally, the act of awakening the cat responsible for destroying Golda’s favorite stuffed elephant, an act on the same level as that of awakening Murdock, is caused by the same action-plan that causes the act of awakening Murdock. This is, of course, just one example. But it does seem reasonable to think that a similar causal story could be told in connection with the execution of any action-plan that one chooses.

This concludes our presentation of the Goldman model of human action. The claim was made at the outset that this model is useful for facilitating the discussion of issues of moral and criminal theory. With that in mind, we turn next to a distinction commonly made in moral theory, and frequently invoked in theoretical discussions of the criminal defenses of justification and excuse: the distinction between act evaluation and agent evaluation.

### 2.3. Act vs. Agent Evaluation

It is common in moral theory to distinguish act evaluations from agent evaluations. Our language, to some extent, reflects this distinction. Certain evaluative expressions seem to apply exclusively to conduct (e.g.,
"right," "wrong," "permissible," "obligatory"); while others apply only to agents (e.g., "trustworthy," "unreliable," "long suffering"). Still others seem equally at home in either evaluative context. We commonly classify persons as good, bad, courageous, cowardly, kind, selfish, and blameworthy. But we also speak of good, bad, courageous, cowardly, kind, selfish, and blameworthy conduct. Even when evaluative expressions appear in both contexts, however, it seems that they are being used to make importantly different sorts of judgments in each context.

We all recognize that there are occasions when a single evaluation is insufficient to capture all that needs to be said morally about a situation. A classic example is the well-intentioned person who manages inadvertently to cause great harm. Focussing first just upon the conduct, there may be a number of evaluative judgments that are relevant. As the preceding discussion of the nature of human action makes clear, during any given interval of time a person may exemplify an indefinite number of act-types. For example, Arnold may move an index finger, squeeze the trigger of a gun, fire the gun, shoot Harold, and cause great harm to Harold. But at the same time he may be keeping a promise (to keep his child safe), saving a life (his child's), awakening a neighbor, disturbing a neighbor's sleep, etc. If we focus on any one of these act-tokens, we get an incomplete picture of what Arnold did during the interval in question. Looking at the fact that he caused great harm, we might say that what he did was dreadful. And certainly it was. But if in causing the great harm to one person he was also keeping a very important promise and saving a life, what he did was also (in
certain respects) laudable. For certain purposes it may be important to arrive at some kind of "all things considered" evaluation of what an agent did on a given occasion. We may, for example, wish to encourage similar conduct in the future. For such purposes it would be useful to have a less "fine-grained" concept of an act--one that will allow us to include the exemplification of several act-types as part of "the same act." One fairly obvious candidate would be the entire set of act-tokens performed by an agent at a time and representable on an actual act-tree. As a matter of convenience I will adopt the convention of referring to such a set of act-tokens simply as an "act." When I wish to refer to an act-type or act-token, I shall do so explicitly.

In order to arrive at an all things considered judgment about the moral worth of an act, one must balance the morally good constituent act-tokens against the morally bad ones. Such an evaluation may seem impossible, given that an act may contain an indefinite number of act-tokens. But fortunately only a small number of the act-tokens involved in any particular act are likely to be morally relevant. The other act-tokens may be safely ignored. Which act-tokens one deems relevant is a function of one's moral theory. One of the tasks of a normative moral theory is to identify morally relevant act-types. For example, hedonistic utilitarianism includes "bringing about (or causing) pleasure" on its list of morally relevant act-types. Kantianism does not, but includes "respecting the rational autonomy of another" on its list.

In addition to assessing an act, one may be interested in assessing the person who performed the act. We may, for example, wish to decide
whether the person ought to be blamed or punished for the act. For these purposes the various moral assessments (including an "all things considered" evaluation) of an act may be helpful, but rarely conclusive. Assuming that the harm caused is sufficiently great to warrant overall condemnation of an act, we may have prima facie reason to condemn the agent as well. But whether the agent is indeed worthy of condemnation will depend on other factors as well, not the least of which is the agent’s motivation for performing the act. Suppose, for example, that the agent was actually trying to show kindness to one who had previously been unkind, but that the effort "misfired" in some unanticipated manner, resulting in grave harm to the intended beneficiary. In that case we might be inclined actually to praise the agent, despite the unfortunate nature of the conduct performed.

As in the case of conduct, we may have reason to make an "all things considered" evaluation of an agent. Perhaps we wish to know whether the person is the type that we want to befriend or marry. Such a judgment should rest upon a composite of all relevant character traits. Although a single act, or a pattern of conduct, often provides reliable evidence for an inference to an underlying character trait, this is not always so. Sometimes good people do bad things. They may even have one or two undesirable character traits. Likewise, the most dastardly among us may occasionally do good, or may have some worthwhile trait of character. If we are to hope to do justice to all that needs to be said morally about most situations in which conduct occurs, we need a moral language that allows us to condemn,
condone, or praise conduct without necessarily condemning, condoning, or praising the agent, and vice versa.

In addition to making possible the kind of sensitive moral evaluation just described, the combined use of act and agent evaluations may also facilitate the enterprise of moral education. We want our sons and daughters to grow up to be autonomous, compassionate, courageous, honest, just, temperate and trustworthy people. But we also want them to be able to make correct moral decisions in individual cases. We know that having the desired character traits is no guarantee that correct moral decisions will be made. Humans are fallible after all. On the other hand, there is more to morality than simply knowing what is the right thing to do. The manner in which (i.e. from what motives and with what emotional tone) morally right conduct is performed is also important. In fact, absent adequate moral motivation, there is no guarantee that one who knows what is the right thing to do will do the right thing. Both of these goals, developing the moral personality and providing guidelines for practical decision-making are important.

Appealing to paradigms of virtuous men and women (Christ, Muhammad, Gandhi, Mother Teresa), along with illustrative stories of how such persons conduct themselves in various circumstances, seems well suited for the first goal. Here the emphasis is not so much on what is done, but how it is done. Of course, parents are important moral paradigms for their children. Children learn as much by watching the manner in which their parents meet various obligations and duties as they do by observing what duties and obligations are there to be met.
Although an important tool for teaching the subtle nuances of moral "style," the appeal to paradigmatic personalities and the language of agent evaluation that accompanies it is ill-equipped for the task of teaching practical decision-making skills. If one’s question is whether to abort a mentally retarded fetus, it is of little help to be told to be virtuous (e.g., compassionate, honest, resolute, etc.), or even to act as the virtuous (e.g., compassionate, honest, resolute, etc.) person would act (how would the virtuous person act and why?). Here an appeal to various moral rules and use of the language of conduct evaluation seems particularly helpful. In the very early stages of moral development parents tend to focus more on the "content" of morality than on style. "Don’t lie" (or "we don’t lie"), "keep your promises" (or "we keep our promises"), and "don’t hit your brother" (variously expressed) are common instructions provided for the education (and survival skills) of the one and two-year-old. No doubt children are "tuned in" even at this early stage to elements of moral style, but parents seem more interested in laying down the ground rules. It is soon discovered that an appeal to such simple rules is inadequate. At first we (parents, teachers, etc.) try to handle complications by building in relevant exceptions: "Don’t lie except when the stakes are low (it’s a "white" lie) and you can thereby make someone else very happy (e.g., Aunt Mildred will take pleasure in thinking that you like her dress, and your mother will be spared the humiliation that would result if you were to offend Aunt Mildred). Later still it may become clear that no learnable list of exceptions will handle all relevant cases. This is the point at which one introduces the
terminology of "moral considerations," "weighing," and "prima facie" duties, obligations, etc. Throughout the process, however, this aspect of moral education has a distinctive focus on the evaluation of conduct (or types of conduct). It seems useful, therefore, to have available both the language of conduct evaluation and the language of agent evaluation when teaching the content and style of morality.

Although there are good reasons for drawing a distinction between act and agent evaluation, the distinction is not entirely unproblematic. We begin by noting that when one looks at the overall situation in which an agent performs an act, it is not always clear what features of the situation pertain to the act and what features pertain to the agent.

As we noted earlier, a single basic act-token may generate a number of higher-level act-tokens. Thus, Arnold may squeeze the trigger of a gun, fire a gun, shoot Harold, wound Harold, kill Harold, murder Harold, eliminate a business competitor, cause grief to Harold's spouse, and startle Harold's neighbor, all by merely moving his index finger. This "accordion effect" of actions appears to be bidirectional. Starting with the basic act-token, one may stretch the action accordion outward from the agent so as to include causal effects of the basic act-token. Thus, moving an index finger may generate firing a gun, wounding Harold, killing Harold, and startling Harold's neighbor. But one may also stretch the accordion inward to include various mental events or states of the agent. Thus, if Arnold killed Harold intentionally (and without justification), he also murdered Harold by moving his index finger. And if Arnold murdered Harold out of spite or from a motive of revenge, one could include the motive as well: it was a
spiteful, malicious, or vengeful murder. It even seems possible to go further and incorporate reference to dispositions or character traits. The performance of certain act-tokens seems to entail that the agent has a corresponding character trait. Examples are courageous, cowardly, honest and dishonest act-tokens.

One might go even further and suggest the possibility of moving the basic act-token itself into the inner arena. Suppose we humans at some point evolve into creatures capable of causing physical changes in our environment through sheer efforts of will unmediated by our own bodily movements. By simply willing that Harold's heart stop beating, our futuristic Arnold could bring about Harold's demise. Some, we may suppose, would be better at effecting desired changes than others (just as some are now better marksmen with a gun than others). Now suppose that futuristic Arnold tries unsuccessfully to stop Harold's heart. Is this not attempted murder? Criminal attempts are normally thought to require the performance of an act. But here the act cannot be identified with any change in the world external to the agent; there was no such change. The most likely candidate for an act is some internal psychological event, perhaps the effort of will exercised by Arnold, or some resulting psychological event, like that of imagining a beating heart stopping. I see no reason to resist this conclusion and insist that human action requires bodily movement or the occurrence of any other external event.46

These observations suggest that the line between act and agent is not all that clear. Moreover, one might suspect that this unclarity will adversely
affect any attempt to distinguish sharply act from agent evaluation. Do we, for example, assign the psychological states of intention or knowledge to the agent, for purposes of evaluation, or do these belong to the act? And what about motives? If a particular murder was "malicious" or "vengeful," is that an additional feature to take note of in evaluating the act, or is that more properly addressed when evaluating the agent? And if we turn our evaluative eye toward honest or courageous acts, does the underlying character trait itself enter into our evaluation of the act? Or are we to evaluate the act independently of the trait? These questions raise serious concern about the tenability of insisting upon a sharp distinction between act and agent evaluation.

One strategy for maintaining a sharp line between act and agent evaluation must, I think, be resisted. It may be granted that ordinary language action descriptions include psychological components. Nevertheless, it may be suggested, for purposes of moral (and legal) evaluation, we should depart from ordinary usage and purge the "act" of all psychological elements. This proposal makes sense, some may argue, for at least two reasons. First, the line between the non-psychological ("objective") component and the psychological ("subjective") component is a clear one that can be consistently drawn across all types of conduct subject to moral or legal evaluation. Second, the value assigned to the non-psychological component of any act is entirely independent of the value assigned to the psychological component. In some cases, as in murder, the complex of an undesirable objective state of affairs and an undesirable mental state is considerably worse than either the state of affairs or the
mental state alone. But the explanation for this is simply that value is additive, so that if one adds the negative value of the objective state of affairs to the negative value of the accompanying mental state, the net result is a complex whole that is much less valuable than either component standing alone. There is, however, no reason to think that the value of the objective state of affairs is in any way affected by the value of the accompanying mental state, or that the value of the mental state is affected by that of the objective state of affairs. Nor is there reason to think that some third value is somehow produced by the interaction between the objective state of affairs and mental state. For these reasons we should depart from ordinary language usage and identify the "act," for purposes of moral and legal evaluation, with a non-psychological objective state of affairs (e.g. bodily movement, together perhaps with the causal consequences of that movement). We may then relegate to the province of "agent evaluation" all psychological events, states, and processes that accompany the act. This is tidy, but is it defensible?

It must be granted that one can draw a sharp distinction between the objective and subjective components of conduct. One way of doing so is to regard the objective as that which occurs "outside a person's skin," while the subjective is that which occurs "inside the skin." Those making such a recommendation commonly assume that this way of drawing the distinction corresponds rather closely to the common sense distinction between psychological events, states and processes, on the one hand, and behavioral and other external events, states and processes, on the other. However,
Wittgenstein and others have cast serious doubt on this way of distinguishing the psychological from the non-psychological. 47

Even if one ignores this complication, there is another problem with the recommendation: the assumed value independence of the psychological and non-psychological does not hold. The case of murder can be misleading in this regard. When a murder occurs someone is caused to die, typically as a result of the bodily movements of an agent. It seems reasonable in such a case to judge the causation of death to be bad. This evaluation stands, one might argue, regardless of the mental state accompanying the agent’s bringing about the death. It seems sensible, then, to recognize the possibility of two entirely independent evaluations: one of the "act" (causing the death of another) and one of the "agent" (intending to bring about the death of another). 48

This analysis becomes strained in the case of crimes involving wholly innocent (value neutral) objective states of affairs. The classic example is that of perjury, which may be defined as "making a statement, whether true or not, on oath in a judicial proceeding, knowing it to be false or not believing it to be true." 49 Making a statement under oath is an evaluatively neutral event. It only becomes suspect if accompanied by a suspect mental state. The law of criminal attempts furnishes many additional examples of wholly innocent objective states of affairs that support negative evaluations only when accompanied by certain mental states. Consider the "act" of placing sugar in a cup of coffee, on its face an entirely innocent act. But if the agent mistakenly believes the sugar to be arsenic and places it in the coffee with the intent to poison another, the innocence vanishes. We now
have a most evil act indeed: attempted murder.

Now I submit that in these cases it is implausible to maintain that the negative value of the criminal offense is equal to the negative value contributed by the mental state alone. In the case of perjury the complex state of affairs of making a statement under oath, knowing it to be false or not believing it to be true, is considerably worse than simply knowing the statement to be false, or not believing it to be true, or even (while not believing it to be true) intending to make the statement under oath, but failing to do so. Furthermore, the complex state of affairs of intending to poison another and placing sugar in a cup of coffee in an effort to realize that intent is considerably worse than simply deciding or intending to poison another without engaging in any overt bodily movement toward that end.\(^{50}\)

The objective state of affairs contributes in some way to decrease the value of the overall complex state of affairs, even though in isolation it is value neutral. Whether the additional value is to be assigned to the objective state of affairs, to the mental state, or to the union of the two, something analogous to a chemical reaction takes place when the objective and subjective components are brought together in a certain manner. Something is lost when the attempt is made to abstract and evaluate each component in isolation from the other.\(^{51}\) This recommendation for preserving a sharp distinction between act and agent evaluation will not do.

Perhaps the beginning of wisdom here is to recognize that an act is itself a feature of an agent. According to the theory of action sketched in the preceding section, an act-token is an exemplification of a property (act-
type) by an agent. Thus, in a sense, every evaluation of an act is to some extent an evaluation of the agent who has performed the act. Of course, we sometimes wish to take note of and evaluate more enduring features of an agent, such as dispositions or character traits. But the distinction between evaluating such enduring traits and evaluating more transitory features may be more one of degree than kind. In addition, some features such as the motive with which an act is performed may be relevant to the evaluation of both an act and the agent performing the act. People who kill others because they do not like the color of their skin have not only performed a most abhorrent act, but have revealed themselves to be of vile character as well. My conclusion is that the distinction between act and agent evaluation, although a useful device for drawing our attention to complexities of moral evaluation that may otherwise go unnoticed, and insuring a balanced approach to moral education, is not as sharp as is sometimes thought.
ENDNOTES

1I am, to some extent, just carrying out a research program suggested by Alvin Goldman. See A. Goldman, A Theory of Human Action 224-26 (1970) [hereinafter cited as Goldman].

2Goldman 10.


4See Goldman 1-10.

5For a nice summary of the arguments of both sides of this debate, see L. Davis, Theory of Action 28-38.

6Here I am omitting the temporal references for the sake of simplicity. One should not forget, however, that such references are always required.

7These are not the only important relations holding between act-tokens, but they are fundamental in that other important relations are definable by reference to them.

8Goldman 21.

9Goldman 21-22.

10To be more precise, each of the pairs of generational acts in this example are related by what Goldman refers to as "causal generation." I shall have more to say about this species of level-generation, as well as the others that Goldman distinguishes, momentarily.
Note that it would make no difference to the point that I am raising if we were to substitute the act of John's causing the death of Harold for the act of his killing Harold. The same question could be asked, viz., when did the act of causing Harold's death occur?

That Goldman is in fact making this assumption is born out by his recursive definition of 'act-token,' the fourth component of which is: "If A is a temporal part of an act-token, then A is an act-token." Goldman 45.

Whether he can likewise ignore the possibility of non-act temporal parts of acts in the context of giving a recursive definition of 'act-token' that stipulates that any temporal part of an act-token is an act-token is somewhat more questionable. See Goldman 45. In order to avoid circularity it would seem that a characterization of "temporal part" that does not assume that only acts are temporal parts of acts is required.

Of course there may be a number of other acts that are causally generated by this act. By extending his arm out the car window John may cause a pedestrian to mistakenly think that he is being waved at, in which case John's act of extending his arm out the car window will have causally generated the act of causing a pedestrian to mistakenly think that he is being waved at.

The last type of case is noteworthy, for it involves as an augmenting circumstance a co-temporal act. The co-temporal acts of shooting a basketball and jumping (one can shoot a basketball "while also" jumping and vice versa) may each generate, via augmentation generation, the act of jump-shooting. If shooting is taken as the generating act, the co-
temporal act of jumping is regarded as an augmenting circumstance. If jumping is taken as the generating act, the co-temporal act of shooting is regarded as an augmenting circumstance. Goldman calls this species of augmentation generation "compc ind generation." Goldman 28.

22 Goldman 28.

23 See Goldman 28-30, for a discussion of the issues involved in taking augmentation generation to be a species of level-generation.

24 Once again, for a detailed discussion of tree-diagrams, see Goldman 30-38.

25 Goldman 31.

26 Id.

27 Id.

28 Goldman 64.

29 Id.

30 Goldman 18.

31 Goldman 65.

32 Id.

33 Goldman 64.

34 See Goldman 65-66.

35 Goldman 67. The clause in square brackets was added. Rather than have this proviso as an unstated background assumption, it seems better to include it explicitly in the definition.

There may be a question about what it is for a property exemplification to "depend" on S's level-generational knowledge or cause-and-effect knowledge. It could be argued that the exemplification of some of the act-types that Goldman lists as basic depend on cause-and-effect knowledge. The ability to shrug one's shoulder, pucker one's lips, and
wrinkle one’s nose seems to require some knowledge of how to do these things. Or consider the act of saying ‘hello.’ Talking is (at least in large part) an acquired skill that presupposes considerable cause-and-effect knowledge (as well as knowledge of language conventions). It is not entirely clear why saying ‘hello’ is to be deemed a basic act, while saying ‘hello’ in the way that Jimmy Carter says ‘hello,’ when performed by an accomplished impersonator is not. Think of the child learning to say ‘hello.’ In the process of learning the child may concentrate on each syllable separately before learning to say the entire word as a "unit." Or compare the beginning typist with the expert touch typist who types whole words or phrases without thinking of each letter. Does the child’s act of saying ‘hello’ depend on cause-and-effect knowledge? What about the expert typist’s typing the word ‘hello’?

One suggestion is that we require that exemplification of basic act-types not be mediated by conscious inference from or access to level-generational or cause-and-effect knowledge. Such a reading of the dependency relation would certainly draw the distinction between basic and non-basic act-types in a way different from that of Goldman. On this view, it may turn out that a concert pianist’s playing the C-scale, (or even an entire musical score?) is a basic act. Since my reliance on the Goldmanian theory in what follows does not require a resolution of this issue, I merely wish to flag the potential problem.

36 Goldman 70.

37 Goldman 72.

38 This example is a variant of one discussed by Goldman at 60-61. It should be noted in connection with condition (b) that Goldman recognizes that more needs to be said about the nature of "the characteristic way" in which action-plans cause exemplifications of basic act-types. However, he does not think it the job of the philosopher to provide a more detailed account of this characteristic manner of causation. The task of filling in the details of the relevant causal mechanism is one for the scientist (e.g., the neurophysiologist). See Goldman 62-63.

39 Goldman 45.

40 Goldman 57.

41 For a detailed account of intentional action, see Goldman 56-63.
42 See the definition of 'basic act-token,' supra 50.

43 It is, of course, a matter of controversy whether in such cases one of these uses is primary and the other derivative.

44 The term is Feinberg's. J. Feinberg, Deserving 134.

45 According to the analysis of human action offered in the previous section, basic act-tokens are necessarily intentional. On this view even when the accordion is entirely collapsed, it includes a mental element.

46 Goldman entertains the possibility of "mental acts." But his analysis precludes both wants and beliefs, from being acts, since neither are exemplifiable at will, nor are they generated by some property that is exemplifiable at will. Goldman 69-70, 92-93.


48 Of course this way of describing the case greatly oversimplifies things. If causing the death of one saves the lives of many, we may judge the objective state of affairs as good (on balance). And of course an agent who caused the death of one in order to save others may likewise be deemed praiseworthy (or at least not worthy of condemnation).

49 See Husak, Philosophy 124. The Model Penal Code definition requires that the statement made be false. MPC § 241.1. On this definition one might argue that the objective state of affairs of making a false statement under oath will nearly always be at least prima facie bad, since it may be relied upon, thereby distorting an official decision-making process with potentially significant consequences for those involved. To this I would say that perjury has not always been defined to require that the statement made be false, and in any event, the example of criminal attempts, which clearly may involve evaluatively neutral objective states of affairs is sufficient to make the point I wish to make.

50 I think similar comments are in order for murder, for that matter.

51 Similar remarks apply even if we move the basic act-token into the inner arena. Suppose I want to see Harold dead and even form the intent to kill him myself if the opportunity arises. Suppose further that I have the power to stop his heart from beating by going through a mental exercise
(e.g., imagining a beating heart stopping). However, the exercise must be performed while in close proximity to the "target," and even then does not always work. Now suppose I see Harold enter an adjoining room. I position myself as close as possible to the room Harold has just entered, muster my powers of concentration, and imagine a beating heart stopping. Nothing happens. Harold exits unharmed. We might divide my attempt on Harold's life into two components: the intent to kill Harold and an "innocent" mental act of imagining a beating heart stopping. Again, I would suggest that the disvalue of the complex of imagining-a-beating-heart-stopping-with-the-intent-to-kill-Harold is substantially greater than that of my *simply* intending to kill Harold.
Chapter III

3.1. Introduction

Having laid the necessary groundwork in the previous two chapters, we are now in a position to begin our quest for a plausible theory of criminal law justification defenses. Despite a considerable amount of interest among contemporary criminal law scholars in the distinctions among the various defense categories, especially those of justification and excuse, few attempts have been made to articulate and defend a comprehensive theory of justification. One of the few scholars to have made such an effort is Paul Robinson. His theory will be the focus of this chapter’s investigation. Criticisms of Robinson’s theory will provide the motivation for, and some guidance in the development of, a more satisfactory theory.

3.2. Justification and the Harm Requirement

Recalling the distinctions made in Chapter I, Robinson’s theory may be classified as a "strong objective" theory of justification. An objective theory holds, in effect, that the mental state of the accused at the time the charged act was allegedly committed is irrelevant to whether the accused has a valid justification defense to the charge. A strong theory endorses the
incompatibility thesis, viz., the thesis that only one of two incompatible acts can be justified. Robinson combines a harm-negating rationale with an objective theory of harm to yield an objective harm theory of justification:

\[ J(da) \leftrightarrow O(c) \& \sim H(ac) \]

[Read: \( d \) is justified in performing \( a \) if and only if circumstances \( c \) obtain such that the performance of \( a \) in \( c \) is not harmful.]

According to Robinson, justification defenses compensate for the overly rigid nature of written codes. Through the agency of the justification defense judges and jurors have an officially recognized opportunity to exercise sensitive moral judgment:

... [J]ustification compensates for the inherent limitations of a written criminal code. It excludes from the jurisdiction of the criminal law those cases where a defendant engages in conduct which the code prohibits, but which, because of special circumstances, does not in fact harm society or its members and which therefore should not be prohibited or punished. Inherent in this theory of justification is the view that harm is a prerequisite to criminal liability. As a penal code prohibits only harmful acts, so does justification, in refining the application of the code, exculpate nonharmful acts.\(^2\)

Several comments about this passage are in order. I shall assume that when Robinson says that harm is a prerequisite of criminal liability, he intends the harm requirement to place moral constraints on the imposition of criminal liability. In other words, we should read the statement as saying that harm is a morally necessary condition of this form of State coercion, so that the criminal law may prohibit and punish only harmful acts. So understood, the harm requirement is designed to be an effective critical tool for evaluating actual or proposed criminal legislation, as well as the
interpretation and application of legislation by the courts.

The actual language of the above quote is consistent with a purely descriptive interpretation, but such an interpretation makes the thesis that harm is a prerequisite uninteresting, if not patently false. If the claim is that legislators never criminalize nonharmful conduct, it is clearly false. To see this we have but to remind ourselves that there was a time when some of our own laws forbade blacks from sharing the same water fountain with whites, or sitting next to whites on a bus. Of course the claim may be that legislators never criminalize conduct unless they regard it as harmful. This may be true. But even if it is, such an empirical claim could not serve as an effective critical tool for evaluating actual or proposed legislation. The thesis that harm is a "prerequisite" for criminal liability in this sense seems hardly worth the ink it would take to state it, and certainly less than that required to comment on it.

Robinson's position that harm is a moral prerequisite of criminal liability should not be confused with that according to which harm is merely a morally relevant reason in favor of, or a *prima facie* moral requirement of, criminal liability. For Robinson harm is a *necessary condition* of the permissible imposition of criminal liability.¹

When Robinson says that conduct expressly prohibited by statute may nevertheless fail to provide grounds for criminal liability because it does not occasion harm, he must mean that under the circumstances in which the conduct was performed there was no *net* (or on balance) harm. A person who kills an aggressor in defense of that person's family causes harm in a
fairly obvious sense. But if the aggressor is thereby prevented from inflicting an even greater harm, the act of killing is not on balance harmful.⁴

Surprisingly, Robinson does not devote much energy to explicating the central concept of harm. The best he does is to say that, as he is using the term, it "has its common meaning 'physical or mental damage,' 'an act or instance of injury,' or a 'detriment or loss to a person.'"⁵ Not only individual persons may be harmed in this damage or injury sense. So may groups of individuals, institutions, and even society itself.⁶ The problem with this characterization of harm, and especially its inclusion of "societal harm," is that it is so general as to border on vacuity. The danger posed by such a concept is that, when employed in a fundamental principle of liability, that principle may be used to sanction or exclude almost any piece of criminal legislation. When that happens, the concept of harm functions as a "conclusory concept," announcing the author's conclusion that liability should or should not be imposed, rather than functioning as a meaningful premise from which to argue for such a conclusion.⁷ Consider the following attempt by Robinson to illustrate the value of the "straightforward and distinct meaning" he assigns to "harm":

[O]ne could argue to a legislature that homosexual acts do not constitute a societal harm and should not be prohibited. One argues this by making and trying to document the claim that homosexual acts do not damage or injure society. It is not necessary to repudiate the requirement of societal harm to support legalization of homosexual acts.⁸

The problem with this illustration is that the concepts of societal damage and injury are just as vague, and hence just as subject to a question
begging conclusory use, as is the concept of societal harm itself. A person who, for whatever reasons, maintains that homosexual acts harm society is likely, for the same reasons, to think that such acts damage or injure society. It would seem that something more needs to be said about the concept of harm, and in particular societal harm, if a harm requirement employing that concept is going to be a useful critical tool.

No doubt there are limits to how much precision one can expect in an analysis of a term like "harm." But we might be able to improve somewhat on Robinson’s characterization of harm by importing an analysis of harm offered in a recent work by Joel Feinberg. Certainly the most comprehensive account of criminal harm to date is presented in Feinberg’s four volume treatise, *The Moral Limits of the Criminal Law*, especially volume one, *Harm to Others*. Writing in the tradition of John Stuart Mill, Feinberg attempts to answer the following question: "What sorts of conduct may the State rightly make criminal?" Alternatively, the task he sets himself is "to trace the contours of the zone in which the citizen has a moral claim to be at liberty, that is, free of legal coercion." He invokes a "presumption in favor of liberty," by which he means that "whenever a legislator is faced with a choice between imposing a legal duty on citizens or leaving them at liberty, other things being equal, he should leave individuals free to make their own choices." Factoring in the liberty presumption, the question to be addressed becomes: "what kinds of reasons can have weight when balanced against the presumptive case for liberty?" Answers to this question come in the form of "liberty-limiting principles."
Feinberg’s own response to the question is a position identified as "liberalism," i.e., "the view that the harm and offense principles, duly clarified and qualified, between them exhaust the class of morally relevant reasons for criminal prohibitions." 14

A liberty-limiting principle states neither a necessary nor a sufficient condition for justified State coercion. It does not state a sufficient condition because "in a given case its purportedly relevant reason might not weigh heavily enough on the scales to outbalance the standing presumption in favor of liberty." 15 Nor does it state a necessary condition for the justification of State coercion since "there is nothing in a principle’s formulation to preclude any or all of the others from being valid too. If more than one coercion-legitimizing principle is correct in stating an always relevant reason for criminalization, then no one of them can state a necessary condition for all justified criminalization." 16

Feinberg identifies a number of commonly-proposed liberty-limiting principles. 17 For our purposes it is his harm principle that is of immediate interest. The harm principle states that "[i]t is always a good reason in support of penal legislation that it would probably be effective in preventing (eliminating, reducing) harm to persons other than the actor (the one prohibited from acting) and there is probably no other means that is equally effective at no greater cost to other values." 18 We cannot begin to enter into the complexities that surround Feinberg’s explication and defense of this principle, but it will be useful to take a look at how he defines the central concept, "harm."
Feinberg distinguishes three meanings of "harm." In what I shall call its "core" sense, harm is "the thwarting, setting back, or defeating of an interest." To have an interest in X is to have a stake in X's well-being, where to have a stake in X is to stand to gain or lose depending on the nature or condition of X. In a second, normative sense, "[t]o say that A has harmed B . . . is to say much the same thing as that A has wronged B, or treated him unjustly. One person wrongs another when his indefensible (unjustifiable and inexcusable) conduct violates the other's right . . . ."

Although it is rare for someone to wrong others without harming them in the core sense of setting back their interests, it is not uncommon to set back their interests without wronging them, for "some actions invade another's interests excusably or justifiably, or invade interests that the other has no right to have respected." Paradigmatic of the latter category are invasions of interests to which the victim has consented. According to Feinberg, neither harm as setback of interests nor harm as wrong is the type of harm required for purposes of a harm principle that is to serve as a guide to the moral limits of criminal legislation. For this purpose we need a third sense of "harm." Legal (or criminal) harm is the intersection of the other two types of harm: "only setbacks of interests that are wrongs, and wrongs that are setbacks to interest, are to count as harms in the appropriate sense."

There are, then, four distinct ways to argue that particular conduct did not cause criminal harm: (1) argue that there was no invasion (or setback) of interests, (2) argue that the person whose interests were invaded had no right that the interests not be invaded, (3) argue that any rights-violating
invasion was justified, or (4) argue that any rights-violating invasion was excused. To argue in the first way is to argue that there was no "core" or setback harm. To argue in any of the latter three ways is to argue that even if there was a setback harm, it was not wrongful.\textsuperscript{25}

The preceding analysis suggests five additional interpretations of "harm." One might identify harm with a setback of interests that (perhaps defensibly) violates a moral right of another (e.g., there has been no consent, waiver, or assumption of risk). One might identify harm with an unjustified (but perhaps excused) violation of a moral right of another (without regard to whether there has been a setback of interests). One may take harm to be a setback of interests that unjustifiably (but perhaps excusably) violates a moral right of another. Two other possibilities are that harm is an indefensible setback of interests (that may not violate a moral right), or an unjustified setback of interests (that may be excused or may not violate a moral right).\textsuperscript{26} The fifth, sixth and eighth interpretations agree in treating excuses as not negating criminal wrongdoing.

Of the eight meanings of "harm" just distinguished, it would seem that only the first and fourth are viable candidates for inclusion in a formula that "justification negates harm." All of the others expressly require absence of justification as an element of the definition of "harm." Their inclusion would, therefore, render the formula vacuously true. Moreover, one could not understand the claim that harm is being negated without a prior understanding of the nature of justification. Assuming that the above list of meanings of "harm" is comprehensive, if the thesis that justification
defenses negate harm is to prove enlightening, the harm in question must be either that of a setback of interests, or a setback of interests that violates a moral right. On either of these interpretations harm requires a setback of interests. It would seem that only sentient creatures can have interests (or rights). Thus, the concept of a harm to society requires some explication.

In what sense does society have interests? It would appear that any such interests must ultimately be derived from the interests of individual sentient members of the society in question. I have an interest in my physical integrity. Other persons in society have a similar interest in their physical integrity. There is a societal interest in protecting individuals from physical injury only because individuals have an interest in being free from such injury. Individuals also have a stake in the fair and efficient operation of certain social institutions (e.g., the I.R.S and the courts). Any interference with such fair and efficient operation constitutes a societal harm only to the extent that the interests of individuals in the welfare of those institutions are set back. If no one had a stake in the operation of a certain public institution (e.g., a bureau for the proliferation of the letter Z), interference with the operation of that institution would not constitute a societal harm. The requirement that societal harm be analyzed in terms of setbacks to the interests of individuals may seem like a relatively innocuous restriction, but it will be seen to have significant implications for Robinson’s effort to reconcile criminal attempt liability and liability for other inchoate offenses with his harm requirement.

Finally, when Robinson says that harm is a prerequisite of criminal liability, he means that actual harm is required. It is not enough that
conduct threatens harm, or that it evidences a dangerous character trait:

There are . . . risks of harm that do not cause harm in themselves. For example, no harm results from a person shaking his fist at a rain cloud honestly expecting to cause the destruction of the world. If one adopts the theory that dangerous but nonharmful conduct should be punished, then why not punish one who shakes his fist at a rain cloud, since he has shown his dangerous intentions? Or why not punish a man for having sexual intercourse with a twenty-two year old woman he reasonably believed to be under sixteen, since such conduct is similarly dangerous? We do not do so because the harm requirement has not been met.28

With these comments as background, we are in a position to begin evaluating the thesis that justifications negate criminal harm. Perhaps the place to begin is with the underlying premise that actual harm is a morally necessary condition of criminal liability. Despite widespread acceptance of this thesis, a strong prima facie case can be made against such a requirement. At bottom the objection is that making serious liability consequences contingent on the occurrence of harm treats relevantly similar individuals in significantly and arbitrarily different ways.

That the difference of treatment can be significant, no one would contest. One who intentionally fires a gun and succeeds in killing the intended victim is guilty of murder; while one who fires and misses is guilty of attempted murder. Most American jurisdictions assign quite different penalties to murder and attempted murder. Under the American Law Institute's Model Penal Code, the maximum penalty for murder is life imprisonment or (in some cases) death; while the maximum penalty for attempted murder is ten years.29 One who negligently causes the death of
another may receive a maximum term of five years; while one who engages
in the same type of negligent conduct, but who is lucky enough not to harm
anyone, is guilty of no crime.\textsuperscript{30}

As an aid to highlighting the apparent arbitrariness of such liability
distinctions, consider an analogy. Imagine a set of lotteries that are set up
so that one becomes eligible for a given lottery by engaging in a designated
form of conduct. Once one has entered a lottery, the ultimate outcome is
determined by a random drawing, the payoffs being various penalties. Each
lottery has a range of possible penalties, the minimum being "no penalty,"
with the most severe penalty for some lotteries going as high as life
imprisonment, or even death. Whether one enters a given lottery, we may
suppose, is within one's control. But once entered, the penalty assigned is
completely fortuitous. The payoff differential for some lotteries can be
enormous. It is possible, for example, for two persons to gain entry by
engaging in precisely the same type of conduct, one receiving no penalty,
while the other receives life in prison, or death.\textsuperscript{31}

We can imagine any number of reasons for entering a given lottery.
In some cases the eligibility-producing conduct may be performed in
circumstances (e.g., in the "heat of passion") such that the potential lottery
penalties simply are not taken into account. In others, conduct that makes
one eligible for a lottery may have payoffs other than just the penalties
offered by the lottery, and these additional payoffs may induce some agents
to enter despite the risks of an undesirable lottery penalty. Furthermore, in
those cases where a conscious decision is made to risk a lottery penalty, the
decision is likely to be influenced by the near universal human tendency to
discount the probability of an unfavorable outcome.

The analogy to a criminal justice system that makes significant differences of liability contingent on whether conduct in fact causes harm should be obvious. I am driving negligently or recklessly down a narrow and winding mountain road. As I round a curve, my car suddenly lurches across the center line. But, as luck would have it, there is no oncoming car. I regain control and continue on my way. Jones is not so lucky. While driving in a similar negligent or reckless fashion down the same mountain road, Jones loses control as he rounds the same curve. This time there is an oncoming car. As a result of the ensuing collision, the driver in the other car is killed. Jones is guilty of vehicular homicide. By virtue of engaging in precisely the same type of conduct, negligent or reckless driving, Jones and I have entered the criminal justice lottery. Forces beyond our control determine whether our engaging in this conduct will result in a significant penalty assignment. The forces of nature make the assignment, but whether each of us receives a penalty is as much beyond our control, and so is just as arbitrary, as would be a true lottery assignment of the penalty.\(^\text{32}\) The obvious question to ask is why, assuming we would not be willing to tolerate a true lottery assignment of penalties, we should be willing to tolerate the analogous assignment based upon the equally fortuitous occurrence of harm. Absent some good reason for doing so, it would seem that we ought not allow the occurrence of harm to dictate differences of liability. And this is precisely what a harm requirement does.
In a recent article, Earnest van den Haag offers an argument in defense of the justice of capital punishment that might be thought relevant here:

Maldistribution of any punishment among those who deserve it is irrelevant to its justice or morality. Even if poor or black convicts guilty of capital offenses suffer capital punishment, and other convicts equally guilty of the same crimes do not, a more equal distribution however desirable, would merely be more equal. It would not be more just to the convicts under sentence of death... 

To put the issue starkly, if the death penalty were imposed on guilty blacks, but not on guilty whites, or, if it were imposed by a lottery among the guilty, this irrationally discriminatory or capricious distribution would neither make the penalty unjust, nor cause anyone to be unjustly punished, despite the undue impunity bestowed on others.  

The principle of justice being relied upon in this passage is what Feinberg has called "strong retributivism," the view that "all evil or, more generally still, all fault deserves its comeupance; it is an end in itself, quite apart from other consequences, that all wrongdoers (or faulty doers) be made to suffer some penalty, handicap, or forfeiture as a requital for their wrongdoing." Feinberg contrasts strong retributivism with "weak retributivism," the view that "if someone must suffer, it is better, ceteris paribus, that it be the faulty than the meritorious." Feinberg observes that weak retributivism is a comparative principle, in that it tells us how to deal with situations "in which someone or other must do without, make a sacrifice, or forfeit his interest." It is really a principle of distributive justice. Feinberg recognizes this when he says that the principle of weak retributivism "commonly governs the distribution of that special kind of
benefit called 'the benefit of the doubt,' so that, where there is doubt, for example, about the deterrent efficacy of a particular mode of punishment for a certain class of crimes, the benefit of that doubt is given to potential victims instead of convicted criminals.37 By contrast, strong retributivism is a noncomparative (nondistributive) principle. It applies to situations "where to take from one is not necessarily to take from another."38 It is clear that van den Haag is relying upon the noncomparative nature of strong retributivism in the argument quoted above.

I do not intend to adjudicate the merits of strong and weak retributivism here. What I wish to point out instead is that there is a third principle of fairness or justice that Feinberg's dichotomy ignores. This principle, like weak retributivism, is a comparative principle, but unlike weak retributivism, it does not compare the treatment of the blameworthy offender with that of non-blameworthy potential victims. Instead, it compares the treatment afforded equally blameworthy offenders. The principle in question may be viewed as supplementing, and restraining, the application of both strong and weak retributivism. In a nutshell, the principle requires that if there are good reasons for the blameworthy to suffer (either because they deserve to, or because if they do not, the innocent will), then ceteris paribus, equally blameworthy offenders should be made to suffer equally. It is this principle, which I think both Feinberg and van den Haag could embrace, that is being relied upon in the "lottery objection" argument against a requirement of actual harm for criminal punishment. A presumptive case having been made against such a harm
requirement, the burden of proof is now upon those who would require actual harm for criminal liability.

Much to Robinson's credit, he is not content simply to say that net harm is a morally necessary condition of criminal liability. He provides five arguments for this conclusion. Unfortunately, the arguments are not persuasive.

[1] All would agree that the criminal law seeks to prevent harmful results rather than to punish evil intent that produces no harm. . . . If the criminal law is extended to punish bad intent alone or the mere possibility of harmful conduct, it goes beyond its accepted role, appears unfair and overreaching, and ultimately loses its credibility and integrity.

[2] If one views deterrence as the proper function of the criminal law, a harm requirement is appropriate. To the extent that the criminal law punishes nonharmful conduct, it weakens the stigma and deterrent effect of criminal conviction for harmful conduct. [3] If a defendant who has caused no harm feels that he is punished unjustifiably, rehabilitative efforts will be hampered. Indeed, one may ask: If no harm has been caused, what harm will be deterred by punishment, and what harm-causing characteristic will be rehabilitated? [4] If one believes that the role of the criminal law is to provide retribution, a harm requirement is also proper; in the absence of harm there is nothing for which to seek retribution. . . . [5] The consistency of a requirement of harm with these fundamental purposes of the criminal law is reflected in the fact that harm has, from the earliest of civilized times, been treated as a de facto requirement. In ancient Israel where murder was a recognized societal harm but assault was not, the victim's death was a prerequisite to imposition of any criminal sanctions, even if the attacker struck the blow with intent to kill.39

The first two arguments simply beg the question against those theorists most likely to challenge a harm requirement. The first may be cast as a dilemma: The criminal law may only punish evil intent or harmful
conduct. It may not punish evil intent. Therefore, it may only punish
harmful conduct. Despite Robinson's assurances, it is not at all obvious that
the criminal law may not punish evil intent. On some interpretations (not,
as we shall see, Robinson's), this is precisely what the criminal law does
when it punishes such inchoate crimes as conspiracy and attempt.
Furthermore, the first premise of the argument sets up a dichotomy that
simply ignores other possible objects of punishment. One such object is
conduct that is undesirable, though not harmful. Certain types of "morals
legislation" (e.g., prohibition of private deviant sexual acts engaged in by
consenting adults) might fall into this category, as would conduct that
substantially threatens harm (e.g., reckless driving) without actually causing
harm. Another possibility ignored by the premise is the punishment of
psychological features of a more enduring nature than that of an evil intent
(e.g., character traits). I am not here arguing that these alternatives are
legitimate objects of criminal punishment. My point is that many people
think that they are, and to simply assume that they are not, in an argument
designed to establish that they are not, cannot be persuasive.

The second argument likewise begs the question against those most
likely to challenge a harm requirement. It assumes, without argument, that
the proper goal of criminal punishment is deterrence from harmful conduct.
Why not include deterrence from evil thoughts or from other forms of
undesirable conduct? Moreover, even if we assume that deterrence from
harmful conduct is the goal of punishment, it is far from clear that the most
effective way to accomplish this is by punishing only conduct that results in
harm. In fact, it would seem that if one wished to decrease the incidence of harmful conduct it would be more effective to punish conduct that substantially threatens to cause harm in addition to punishing harmful conduct. We could go further. If, contrary to fact, there were some way to ascertain reliably when a person is merely contemplating a harmful act, punishing such bad thoughts would further decrease the incidence of harmful conduct. Surely harmful conduct would be less likely to occur if merely thinking about doing harm were punished than if punishment were limited to cases of conduct actually causing harm.

There is reason to think that at least for certain crimes, most notably crimes of vehicular homicide involving negligence or recklessness, an even stronger claim can be made. Not only will limiting punishment to instances in which negligent or reckless conduct causes harm result in a significant detrimental impact on deterrence, but there is little or no additional deterrence achieved by punishing more severely harmful negligent or reckless conduct than can be achieved by a lesser penalty for the same negligent or reckless conduct whether or not it causes harm.\(^{40}\)

It also seems doubtful that the stigma attached to harmful conduct must be, or is even likely to be, weakened by punishing anything other than harmful conduct, including dangerous conduct and harm-directed thoughts. So long as it is clear that the punishment is for threatening or planning to do harm, punishment in the latter two types of cases seems likely to reinforce the stigma attaching to harmful conduct. It demonstrates that society regards harmful conduct as such a serious matter that it is willing to sacrifice valuable liberty and privacy interests in order to reduce the
frequency of its occurrence.

In response to the third argument, we might acknowledge that anytime a defendant is punished but feels that the punishment is unjustifiable, rehabilitation may be impaired. But the real question is whether defendants will feel that only punishment for harmful conduct is justifiable. I see no reason to think that this is so. Suppose A plans to murder B, takes aim with his trusty rifle, shoots, but misses the mark. Further suppose that B is completely unaware of the attempt on his life.41 Or suppose A is driving so recklessly as to endanger the lives of several innocent persons. Would A really feel that he was being punished unjustifiably if he were punished for trying to kill B or for endangering the lives of others? I doubt it. In addition, the argument may prove too much. For if a defendant's belief of suffering injustice impairs rehabilitative efforts and this in turn is a good reason for not punishing, then it would seem that there is good reason not to punish anyone who thinks, for whatever reason, that the punishment is unjustified. Thus, there is good reason not to punish members of the lower socio-economic classes who view the criminal law as a tool of oppression wielded by those with economic and political power, and who therefore think their punishment unjust. There is good reason not to punish a Klansman who takes the life of a black man, thinking that the victim is subhuman and the act of violence not really murder. Etc.

An answer to the apparently rhetorical question, "If no harm has been caused, what harm will be deterred by punishment, and what harm-causing characteristic will be rehabilitated?" might be: (i) the future harm that
would otherwise have resulted had the defendant not been punished for trying unsuccessfully to cause harm or for thinking about causing harm, and (ii) the character trait or disposition to cause harm which was manifested in the attempt to cause harm or in the contemplation of harm. 42

The fourth argument amounts to little more than a dogmatic assertion of the harm requirement. Why should harm be required for retribution? Might not one seek retribution for evil thoughts, for unsuccessful attempts, for immoral (but non-harmful) conduct, etc.? One standard interpretation of retributivism is that punishment requires that the agent be morally blameworthy. But it is far from clear that blameworthiness requires that the agent cause harm.

The fifth argument differs significantly from the other four arguments offered by Robinson. The first four arguments attempt to show that making criminal liability contingent on causing harm, although perhaps fortuitous, is not arbitrary, for reasons can be given to justify a harm requirement by appealing to the traditional goals of criminal punishment. By contrast the fifth argument relies upon a claimed de facto requirement of harm from the "earliest of civilized times." The idea seems to be that some evidence for the legitimacy of requiring harm is that societies have always in fact thought that it was legitimate to do so. Such a moral consensus ought, it is suggested, to count for something. Suppose, for the sake of argument, we grant that a consensus would show something. One example furnished from the law of ancient Israel hardly establishes the claimed moral consensus. If there ever was such a consensus on a harm requirement, it seems clear that it no longer exists. The most obvious counterexamples are the inchoate
offenses (e.g. attempt, conspiracy, and solicitation).

Having reviewed the arguments offered by Robinson on behalf of the harm requirement, it seems fair to say that if a harm requirement is warranted, additional arguments are needed beyond those that Robinson provides.43

Are there perhaps other arguments that could be brought to bear in support of such a requirement? An argument that has been made, and that could be cast as a direct response to the lottery objection, goes as follows. Although the same fortuity may be involved in a lottery assignment of penalties as in a harm assignment, most people do not see things this way. People tend to assume that they have more control over chance events than they in fact have. It has been observed, for example, that most people will bet more in a dice game before the dice have been rolled than after. When rolling the dice, they tend to roll softly for low numbers and much harder for high numbers. Apparently they think that they can somehow influence the outcome of the roll of the dice.44 This assumption of control over chance events may help explain why juries are reluctant to impose severe punishment unless harm has been caused. If juries refuse to convict those who fail to cause harm, at least in cases where severe punishment is the likely consequence of conviction, it might be argued that the end result is the same as it would be if the law officially required harm anyway. Moreover, the argument continues, the law's refusal to require harm, in light of the defacto harm requirement imposed by juries, is likely to breed disrespect for the law. The law should, therefore, officially require harm for
In response it should be noted first that this "jury nullification" argument applies only in cases where relatively severe punishment is in the offing. Second, jury nullification is not likely to present a serious problem in the context of intentional crimes. Thus, the case for punishing attempts, even at a rather severe level equal to that currently imposed for completed offenses, is not threatened by the argument. Finally, in the case of nonintentional crimes (i.e., crimes involving negligence or recklessness) the problem of jury nullification can be addressed by imposing relatively mild penalties in the case of harmful and nonharmful conduct alike. As was noted above, the deterrent impact of the law will not be significantly affected by doing so. The jury nullification argument will not, therefore, support harm as a necessary condition of criminal liability.

Another argument for actual harm as a necessary condition of liability derives from the "inherent limitations" of the power of government over its citizens. Absent such a requirement, government could become a most intrusive and oppressive force, to the point of punishing harmless private immoralities and engaging in Orwellian thought control. A harm requirement is justified, therefore, in the interest of protecting individual liberty.

Perhaps the main objection to this argument is that a harm requirement is overly protective of individual liberty. One who conspires with others to bring about a serious criminal harm and who takes steps sufficient to evidence a firm commitment to see the plan through, does not have liberty unduly restricted if the State sees fit to impose liability for the
conspiracy. Nor is the liberty of one who tries, but fails, to kill another
unduly restricted by the State's decision to impose liability for an attempted
murder. Similar remarks could be made about one who carries, without
using, a dangerous weapon. In none of these cases has actual criminal harm
resulted, and yet it seems perfectly appropriate for the State to intervene and
impose liability even though actual criminal harm has not resulted. The
interest in liberty can be adequately safeguarded by a requirement that the
State have sufficient evidence of criminal disposition or intent, and by
setting a high standard of proof. This might be translated into an overt
behavior or act requirement. In many cases, however, there will be proof
beyond a reasonable doubt that a given defendant has a relevant disposition
or intent, and in those cases there would seem no liberty-based reason to
require the occurrence of actual harm.

Those who wish to defend the harm requirement against such
eamples typically respond by stretching the concept of harm in an effort to
avoid the putative counterexamples. One way to do this is to define "harm"
so that it includes cases of unreasonable risk of harm. Another move is to
invoke a "legal presumption" of harm, which is just a way of saying that
there may in fact be no harm, but the law will pretend that there is
anyway. Yet another strategy is to rarify the interest allegedly violated,
so as to make it difficult to prove that no such interest is in fact invaded or
set back. This is often done by introducing the notion of "immaterial" or
"intangible" harm. The move is further facilitated by countenancing
harm to society. This is the strategy employed by Robinson. His response
to each of the above putative counterexamples is that actual, albeit intangible, societal harm is caused in each case.

One must be wary of attempts to deal with the problem of counterexamples by redefining key terms. If the new definition departs too much from the ordinary meaning assigned to the terms, one may be left wondering whether one really understands the statements in which those terms are used. In extreme cases the process of redefinition may yield concepts that have little or no meaning. Such concepts are ideally suited for the kind of question-begging conclusory use referred to earlier. In such cases simple formulas often mask complex decision-making processes. It would better serve the ends of justice if these processes were acknowledged and made the explicit bases of legal conclusions. The prospect of the concept of harm functioning as a conclusory concept is particularly worrisome in the context of an argument that regards the harm requirement as an important check on authoritarianism. The usefulness of a harm requirement for this purpose would seem to vary inversely with the degree of vagueness tolerated in the concept of harm. Robinson's concept of intangible societal harm certainly looks suspiciously like a conclusory concept, and in fact I think that it is. We shall pursue this theme in the following section.

3.3. Societal Harm and Criminal Attempts

A takes careful aim with a .22-caliber pistol, shoots, but misses the intended target, V. In an effort to blow up V's house, A plants several sticks of dynamite and lights the fuse. A police officer on the beat sees
what is about to happen and intervenes before the dynamite can explode. In an effort to blow up V's house, A plants several sticks of what is believed to be dynamite and lights the fuse. As luck would have it, the sticks are a carefully disguised theater prop made of wood. No explosion takes place. In yet another effort to kill V, A constructs a doll after V's image and carefully places pins in strategic locations. V is unaffected by this ritualistic attempt on his life. A, a married woman, finds herself very much in love with another man, B. She discloses her predicament to her spouse and proposes that instead of choosing between her two lovers, she be married to both. With her spouse's consent A enters into what she thinks is a criminally prohibited bigamous marriage with B. However, prior to the second marriage the state legislature, unbeknownst to A, repeals legislation that had previously made bigamy a crime.

In each of these cases, A tried to do something that A thought would constitute a crime. But in each case A failed. In the first two cases A simply failed. But in the latter three cases it would seem that A could not have succeeded. These three cases represent three commonly distinguished types of "impossible attempts." The first is typically referred to as one of "factual" impossibility, the second, "inherent" impossibility, and the third, "legal" impossibility.54 There is general agreement that criminal liability is appropriate in "simple attempt" cases. But debate continues regarding which, if any, types of impossible attempts should trigger criminal liability. Interestingly, Robinson views the impossible attempt debate as one about whether each type of impossible attempt involves sufficient harm to support
liability. This suggests that all attempts produce some kind of harm. The kind of harm produced in varying degrees by all criminal attempts is "intangible societal harm."^55

What might this intangible societal harm be? Following an earlier suggestion, societal harm should be regarded as a function of setbacks to the interests of individuals. Societal harm is at bottom harm to individuals. Keeping this in mind, one way societal harm may be intangible is if the relevant harm to individuals is intangible. A number of possibilities suggest themselves. First, intangible harm may be "imperceptible" harm. One way in which harm can be imperceptible is if the object of the interest set back is imperceptible. Examples of such intangible interests would be an interest in controlling the use of an idea for a play or a sophisticated computer program, or a person's interest in a good reputation. A harm to such an interest may be substantial, even though "imperceptible."

Another way in which harm can be imperceptible is that the degree to which a tangible interest is set back is so minor as to be negligible, even though the object of the set-back interest is quite perceptible. A minor scratch on the arm or a bruise might qualify as an imperceptible harm in this sense.

Jeremy Bentham suggests yet a third way in which harm may be imperceptible. Bentham distinguishes two main categories of harm, or as he prefers to call it, "mischief." Primary mischief is that suffered by "an assignable individual, or a multitude of assignable individuals." Secondary mischief is that which "extends itself over the whole community, or over some other multitude of unassignable individuals."^56 Primary mischief
may be further divided into that which is *original* and that which is *derivative*. Original primary mischief is "that which alights upon and is confined to any person who is a sufferer in the first instance, and on his own account: the person, for instance, who is beaten, robbed, or murdered." Derivative primary mischief is that "which may befall any other assignable persons in consequence of [the person suffering the original primary mischief] being a sufferer, and no otherwise." Two types of secondary mischief may also be distinguished: *danger* and *alarm*. Danger is the increased chance (or probability) that unassignable individuals will suffer the type of original primary mischief caused by the act. Alarm is the apprehension or fear on the part of unassignable individuals that they will suffer such mischief. Bentham identifies two ways in which an act can increase the danger of future mischief similar to the original primary mischief caused by the act: (i) it can suggest the feasibility of performing such acts in the future, and (ii) it can undermine the operation of motives that normally restrain persons from engaging in mischievous conduct. Finally, we may also distinguish the danger and alarm attributable to the agent from that attributable to other members of society who may be tempted to follow the agent's example.

A man physically assaults a young woman, stealing her money and jewelry. The original primary mischief consists of the physical and psychological pain of the assault, and the anguish experienced by the woman at having lost valuable possessions. In addition, there are assignable individuals related in various ways to the woman who also experience pain.
as a result of the assault. She may have a spouse who suffers psychological anguish and who shares in the financial loss. Other relatives and friends may suffer through empathy, or more directly, as in the case of children who stand to lose because less money is available to spend on their own welfare. The aggregate of all such painful consequences experienced by others appropriately related to the young woman constitutes the derivative primary mischief of the assault. Aside from the suffering caused to assignable individuals, there is the fear generated throughout the community that similar acts of violence will occur in the future. This constitutes the alarm of the act. Furthermore, the act will likely have increased the real probability of such future assaults either by suggesting to the offender or to others the feasibility of performing such acts, or by reducing the strength of restraining motives. This constitutes the danger of the act.

The third sense in which harm may be imperceptible corresponds to Bentham's secondary mischiefs of danger and alarm. Danger is an imperceptible "harm" because it is not really an actual harm at all, but merely the probability of harm. Alarm is an actual harm, but it (like danger) may be regarded as imperceptible, since it affects unassignable individuals. The fear may be there, but we would be hard-pressed to say who it is that experiences it. There are, then, three ways in which actual harm may be intangible because imperceptible: the interest may be intangible (i.e., the object of the interest set back is imperceptible), the degree of set-back to tangible interests may be negligible, or the individuals whose interests are set back may be imperceptible because unidentifiable.
Another possibility is that intangible harm is *indirect* harm. Social institutions designed to promote or protect "primary" interests of individuals, some of which may be imperceptible in one of the ways distinguished above, may be harmed in the sense that their operation is impaired. Such institutional harm is indirect because individuals do not have an interest in their existence and functioning *per se*. Their interest in the institution is tied to the primary interests the institution promotes or protects.

Let us consider whether any of these meanings of "intangible societal harm" might be usable by Robinson. We begin with the interpretation of intangible societal harm as imperceptible individual harm. It seems most unlikely that the intangible harm in question is *minor* harm to tangible individual interests. Nearly everyone, including Robinson, would agree that liability should attach in both of the simple attempt cases described earlier. But in neither case is *any* actual perceptible harm caused. Nor does it seem that these cases must involve a setback to primary individual interests in some *imperceptible object*. Consider the attempted shooting. V is the most likely person to have an intangible interest violated. But V need not be affected in any way by the attempted shooting. V need not even be aware of the attempt. No property interest or reputation interest need be implicated. V’s autonomy or privacy interests need not be invaded. Moreover, one looks in vain to find any other individual’s primary intangible interests that must be affected by the attempt.

Some might be tempted by the suggestion that *danger* is the relevant "intangible harm" that is present in varying degrees in all criminal attempts.
It is almost certain, however, that Robinson would not be among them. Recall the earlier quote: "There are . . . risks of harm that do not cause harm in themselves. . . . Why not punish one who shakes his fist at a rain cloud [expecting to thereby cause the destruction of the world], since he has shown his dangerous intentions? Or why not punish a man for having sexual intercourse with a twenty-two year old woman he reasonably believed to be under sixteen, since such conduct is similarly dangerous? We do not do so because the harm requirement has not been met." 61

It is interesting that Robinson should pick these two examples to illustrate that mere dangerousness, not being an actual harm, is insufficient to support criminal liability. For these are precisely the kinds of examples that many commentators would classify as inherently and legally impossible attempts. The man who shakes his fist at the cloud, like the person who places pins in a doll, is choosing a means that "a reasonable man would view as totally inappropriate to the objective sought [inherent impossibility]." 62 Similarly, the man who has intercourse with a twenty-two year old incorrectly thinking that she is under sixteen, like the woman who marries another incorrectly thinking that she is thereby committing the crime of bigamy, has "done everything which he [or she] had meant to do and had thereby not committed a completed crime [legal impossibility]." 63

On the one hand, Robinson implies that all attempts cause actual societal harm when he says that the debate over attempt liability is a debate over the sufficiency of the harm caused. On the other hand, when illustrating the claim that mere dangerousness is not enough to support criminal liability, that actual harm is required instead, he offers examples of
impossible attempts, implying that these attempts involve *no* harm. What is
going on here? Robinson’s claim that the debate over attempt liability
focuses on whether the harm caused is *sufficient* to support liability suggests
that he has a sliding scale of harm in mind. This is consistent with there
being no harm caused in certain extreme cases. Perhaps Robinson thinks
that whatever it is that constitutes the intangible societal harm caused in
varying degrees by criminal attempts has reached the vanishing point in
these two examples. But we still do not know the identity of this mysterious
intangible harm.

Although Robinson does not do so, we should distinguish two kinds
of risks.\(^ {64} \) There is what we may call the "conduct risk." That is the
chance that harm of the type intended (destruction of the world) will occur
as a result of the specific conduct in question (shaking the fist). But there is
also the chance that harm of the type intended will occur, given that the
agent is the sort of person who is willing to take steps (shake his fist) in an
effort to bring about that harm. Call this "agent risk." Persons who pose a
serious agent risk we might call "dangerous." Conduct that itself poses no
serious risk of harm may nevertheless provide evidence that a certain agent
does.

It seems clear that no conduct risk of harm is posed by a person’s
shaking a fist at a cloud or having consensual intercourse with a twenty-two
year old. There may, however, be a significant agent risk of harm. Suppose
that once it becomes clear that fist shaking is ineffective, the person is likely
to change methods and engage in conduct that does pose a significant risk of
harm (e.g., exploding a nuclear device). Suppose the next young woman really is under sixteen. One who thinks that liability is inappropriate in these two cases might account for this by maintaining that whether or not attempt liability is appropriate depends entirely on the degree of conduct risk. In neither case does the conduct in question cause or risk actual harm, even though it may evidence a significant agent risk. By contrast, the conduct of shooting at another and missing, or lighting a fuse attached to dynamite does pose a substantial risk of harm.

Neither conduct risk nor agent risk are equivalent to Bentham’s danger. Danger may, however, be partially a function of both. Of the two types of risk, agent risk would seem to be the more weighty determiner of the degree of danger. At any rate, Robinson is apparently averse to any "risk" account of criminal harm, be it conduct risk, agent risk, or danger, and so would not identify intangible societal harm with any of these.

Could it be that what Robinson has in mind is Bentham’s alarm? Perhaps, but there are two problems with attributing this interpretation to him. The alarm caused by an act is a kind of pain. It is the fear experienced by unidentifiable persons that they may experience the kind of original primary mischief caused by the act. According to Bentham, whether such fear exists, and if so to what degree, is a function of the apparent state of mind of the agent when performing the act. In particular, alarm depends upon the agent’s apparent intentionality, consciousness, motive, and disposition. Alarm is a function of the agent’s apparent state of mind because alarm is a function of the apparent danger posed by the agent (and others who may be encouraged by the agent’s example to behave similarly),
and danger is a function of the *actual* state of mind of the agent.68

The consequences for attempt liability are significant. It seems safe to assume that the apparent mental state of the agent will generally reflect the actual mental state of the agent. The agent will appear to intend to cause harm because he does intend to cause harm. Evidence sufficient to make members of the public believe that the defendant is dangerous will exist precisely because the defendant is dangerous.

A shakes his fist at a cloud, B places pins in a doll, C has intercourse with a twenty-two year old woman, D places several pieces of wood next to a brick building and sets fire to them. Are any of these acts cause for alarm? Absent some information about the mental state of the agent, we cannot say. But if we somehow find out that A was trying his best to bring about the destruction of the world, B was intending to murder V, C believed the young woman to be under sixteen years of age, and D thought the wood was dynamite and intended to blow up the building, we now have reason to be alarmed. The degree of alarm may vary from case to case. We may decide that the alarm attributable to some of the agents is so minor as not to warrant criminal liability. But the point remains that whether alarm occurs, and if so to what degree, is a function of the apparent mental state of the agent, which in turn typically depends on the actual mental state of the agent. If, as Robinson clearly holds, harm is purely a function of objective factors, so that the mental state of the agent is entirely irrelevant to whether harm occurs, alarm cannot be what Robinson has in mind when he speaks of "intangible societal harm."69
That Robinson would reject an alarm interpretation is reinforced by noting that the two examples that he appeals to in arguing against the relevance of dangerousness to attempt liability, the fist shaking and putative rape cases, both could have significant alarm consequences. They are, however, offered as paradigms of non-harmful conduct.

In addition to the two *ad hominem* arguments just presented against attributing an alarm interpretation of societal harm to Robinson, there is another reason for not doing so. Alarm does not accompany all liability-triggering attempts. Alarm depends upon unidentifiable members of the public being apprehensive about the danger posed by a given agent, or by other agents who may engage in similarly threatening conduct. But apprehension requires awareness. If A were unsuccessfully to attempt to kill B without B or anyone else finding out about the attempt, there would be no alarm attributable to the attempt. If alarm is the intangible societal harm upon which attempt liability turns, it would follow that the mere fact that one is successful in concealing an attempt precludes criminal liability. Surely we do not want to base liability on how good criminals are at covering their tracks.70

Having exhausted all of the imperceptible harm interpretations of "intangible societal harm," we come to the indirect harm interpretation. One suggestion is that the indirect, and hence "intangible," harm caused by criminally punishable attempts is one to the social institution responsible for protecting important individual interests from violation, viz., the criminal law. It might be argued that willful attempts to engage in conduct believed to be criminal undermine respect for the law (an actual harm), thereby
diminishing the deterrent force of the law. There are, however, problems with this argument.

First, it seems to confuse harm caused by the attempt *per se* with harm caused by a failure to punish the attempt. The latter kind of harm can never satisfy Robinson's harm requirement. The harm requirement imposes moral limits upon the kinds of conduct that may be *criminalized*. It says, in effect, "thou shalt not criminalize conduct unless it is harmful." Failure to punish *any* type of prohibited conduct, especially if such failure is recurrent, may undermine respect for the law. But this only shows that once the law has determined that conduct ought to be criminalized and has assigned a penalty to the performance of that conduct, it should, in Justice Holmes' words, "keep its promise" to punish violators. It does not show that it was permissible to *make* the promise in the first place, i.e., to criminalize the conduct. In short, the fact that a failure to *punish* an attempt undermines respect for the law cannot suffice as the harm required for purposes of determining whether it is permissible to *criminalize* the attempt.

Second, this criterion would not distinguish inherently or legally impossible attempts from simple attempts or factually impossible ones vis-a-vis the appropriateness of liability. The person who shakes a hand at a cloud in a sincere effort to destroy the world, or who sticks pins in a doll in an effort to kill an adversary, is manifesting the same disrespect for the law as is the person who shoots but misses or tries to explode wooden dynamite sticks. The same is true of one who marries another mistakenly believing that the crime of bigamy is thereby being committed, or one who has
intercourse with a woman mistakenly believing her to be under sixteen. If the failure to punish the open display of disrespect manifested by simple and factually impossible attempts undermines respect for the law, I fail to see why a failure to punish inherently and legally impossible attempts would not do so as well.

It might be suggested that when assessing the merits of attempt liability, including liability for various impossible attempts, one consider the possible adverse affect of failing to criminalize such attempts on the ability of the criminal law to reduce the frequency of serious harm which everyone will agree is the proper concern of the criminal law: the statutory harm of completed offenses. If it could be shown, for example, that failure to punish attempted murder has an adverse impact on the ability of the law to reduce the incidence of murder, then it might be thought permissible to punish such attempts. On this view, the indirect, and hence "intangible," harm caused by criminal attempts is an increased likelihood of future tangible harm. But of course this is just Bentham’s danger, and we have already found reasons to reject danger as an interpretation of Robinson’s intangible societal harm. In addition, it is doubtful that this type of indirect harm will provide a satisfactory basis for distinguishing among the various types of attempts vis-a-vis liability. Not only will punishing simple and factually impossible attempts likely decrease the incidence of completed offenses, but the same is true of punishing inherently and legally impossible attempts as well.

Having examined the arguments for an actual harm requirement, I conclude that the case has yet to be made for such a moral requirement of criminal liability. It seems that there are cases where it is appropriate to
impose liability in the absence of actual harm. Criminal attempts are
paradigmatic of such cases. If criminal attempts cause intangible societal
harm, the nature of this harm remains a mystery. It is difficult to resist the
suspicion that the concept of intangible societal harm is in reality a conclu-
sory concept that disguises the absence of true harm, and is used to
announce the conclusion that liability is deemed appropriate nonetheless.72

3.4. Objective Harm and Self-defense

The view that justified conduct is conduct that results in no net
objective harm does seem to explain many of the paradigm cases described
at the beginning of Chapter I. The prison inmate who escapes in order to
avoid a life-threatening fire, the druggist who administers a prescription
drug in order to save the life of a heart attack victim, the driver of the car
who leaves the scene of an accident in order to rush a passenger to the
hospital, and the "arsonist" who blows up houses in order to save the
community from the ravages of an oncoming forest fire, may all be said to
have avoided a greater harm than would have likely occurred had they
conformed to the letter of applicable criminal statutes.73 The use of
defensive force to protect one’s self or others is also often cited as a
paradigm of justified conduct. No doubt many such cases can readily be
explained on the no-net-harm rationale. One such case is the defense of
one’s family from the life-threatening attack of a single aggressor. Clearly,
if a person must take the life of one in order to preserve the life of many,
Another is the use of deadly force in order to defend a single life from the attack of an aggressor. Most of us would agree that defensive force in such cases is justified. The no harm analysis can readily account for this result: the value of the two lives cancel each other out. The net harm is zero.  

But not every act in defense of self or others will so readily yield no net harm. For we need not confine ourselves to multiple defender or one-on-one attacks. Suppose it is necessary to kill five aggressors in order to save a single life. The use of such defensive force seems clearly justified. If the no harm analysis is to account for such multiple aggressor cases, it must either discount the harm caused the aggressors, or it must find some additional harm avoided by the use of defensive force, other than merely the loss of the defender’s life, which when added to the loss of that life is greater than the harm consisting in the loss of five aggressors’ lives. 

Following the first suggestion, one common way to get the result required by the no harm analysis is to combine a forfeiture of rights theory with a concept of harm that requires rights violation. According to this view, the aggressors, by virtue of their wrongful conduct, have forfeited or waived their right to life. Given that harming another requires that one violate a right of that other, it follows that taking the life of the aggressors causes them no harm. Thus, regardless of the number of wrongful aggressors, the calculation of harm resulting from the defensive use of deadly force always comes out in favor of the use of such force. On one side of the scales we have zero harm caused, on the other side the harm of
killing one who has not waived his or her right to life. Killing the aggressors is clearly justified.

There does, however, seem to be a problem with this particular version of the forfeiture rationale. Rights are either to be identified with important individual interests (e.g., the interest in freedom from physical pain, in self-expression, in continued existence, in exclusive and uninterrupted use of possessions, etc.) or they are a mechanism for providing extraordinary protection of such interests. But if rights are identified with important individual interests, the idea that I may forfeit a right makes no sense. Do I really forfeit my interest in liberty or in a continued existence when I wrong another? Not at all. I still have a very substantial stake in my continued liberty and existence. If we are to appeal to the forfeiture of rights as a way to account for the asymmetry between an aggressor and a potential victim vis-a-vis the harm resulting from loss of life, it would seem that we must regard rights as mechanisms for protecting important individual interests.

Forfeiture of rights does make sense if rights are thought of in this way. On this view rights protect individual interests by imposing especially strong obligations upon others to respect (i.e., not impair) those interests. Forfeiture of a right is accomplished by engaging in conduct that has the effect of eliminating or suspending the special protection of interest(s) which it is the function of the right to afford. It is important to recognize that it is the conduct of the agent that determines whether a given right has been forfeited. Forfeiting is something agents do.
Now in the case of an aggressive attack, what is the nature of the conduct that triggers the alleged forfeiture? Obviously, the aggressor need not actually complete a criminal offense (e.g., murder) against the defender before the defender may respond. Conduct that constitutes a criminal attempt would seem sufficient to trigger a forfeiture. Suppose that A, a jilted lover, pulls out a loaded revolver and points it at B with the intent to fatally shoot B. Has A forfeited the right to life so that B may shoot A in self-defense? I should think that if ever the right to life is forfeited, it is in this case. Now suppose that A, who has forgotten to load the weapon, pulls out the revolver and points it at B with the intent to fatally shoot B. B is unaware that the gun is unloaded. Has A forfeited the right to life so as to permit B to shoot in self-defense? Again, it would seem so. Finally, imagine the same situation except that this time B knows that A has forgotten to load the gun. B knows that A poses no immediate threat when the gun is aimed. May B nevertheless shoot and kill A in "self-defense"? If we take the forfeiture of rights theory seriously, it would seem that we must allow B to kill A even in this last case. A’s conduct in the latter two cases seems indistinguishable, at least in any way that might be thought relevant to the issue of forfeiture. The only distinguishing feature is one over which A has no control, viz., B’s knowledge that the gun is unloaded. In both cases A took aim with a gun A believed to be loaded with the intent to kill B. In both cases A has done something equally wrongful, as evidenced by the fact that both acts constitute criminally punishable attempts, even though one involves so-called "factual impossibility." But surely it is wrong to allow B to kill A in the third case. B’s life is not in danger and B
knows this. If B shoots anyway, B has unnecessarily killed A. No American jurisdiction would grant B a valid justification defense in such a case. Yet the forfeiture theory seems unable to explain the difference between the second and third cases.

There is a move open to one who wishes to maintain a forfeiture rationale in the face of such examples. The move is to reject the theory of harm implicitly relied upon above, according to which harming someone requires that one violate a right. The reason for recognizing a right is that an individual interest is sufficiently important either by itself or when aggregated with other similar interests to generate obligations on the part of others to respect that interest. Clearly, these rights-generating interests may be significantly impaired even in cases where no protective right is being violated because the right has been forfeited. Why not say in such a case, what it seems perfectly natural to say, that the person whose important interest has been set back has been harmed even though no (legal or moral) complaint may be based upon the violation of a right? On this view, even though an interest may not, because of forfeiture, have the benefit of the extraordinary protection afforded by a right, the absence of the right does not mean that "anything goes." The interest must still be accounted for in any overall moral assessment of whether it is permissible to invade or impair the interest.

Recognizing that individual interests may have significant value independent of any protective right could account for such typical requirements as that the legitimate use of defensive force must be based
upon a reasonable belief that such force is necessary to repel an imminent attack and that the degree of force is not excessive given the nature of the harm threatened. Even if an aggressor has forfeited a right to life, life has significant value. One would not be permitted to take it in a case where one knows that a person poses no genuine threat (e.g., the gun is unloaded), the threat is remote, or use of non-deadly force would suffice to repel the attack. 80

Although this interpretation of the forfeiture rationale has the virtue of placing reasonable limits on the use of defensive force in situations where an aggressor has allegedly forfeited a right to life, it has one troublesome implication. According to this view, one would expect that as the number of aggressors increases so would one's hesitancy to regard an act of self-defense as justified. That is, not only would the value of the aggressors' interest in life impose constraints upon the kind of defensive response permissible (e.g., use of non-deadly force if that will suffice), but if the number of aggressors were sufficiently large, the sum of their interests in a continued existence might outweigh a defender's right to life. But in fact the extent to which one is justified in using deadly force to defend oneself from attack does not seem to vary with the number of aggressors envisioned. The defender appears to be just as justified in taking the lives of ten thousand in self-defense, as in taking the life of one for the same purpose.

In response one might simply insist that a moral right always outweighs competing interests that are not themselves protected by a moral right. Perhaps this could be defended as part of the "logic" of rights. If this
move is plausible, it provides the forfeiture theory a way to account for the permissibility of using deadly force against multiple guilty aggressors, regardless of the number of aggressors involved. But even if the forfeiture theory can plausibly account for justified self-defense against threats posed by multiple *guilty* aggressors, it cannot account for cases of justified self-defense against threats posed by *innocent* aggressors.

Suppose that as a result of psychosis or somnambulism A honestly believes that B is about to kill A. A, therefore, pulls out a revolver and takes a bead on B. May B justifiably use deadly force in self-defense? It would seem so. But it is most implausible to suggest that A has done something that constitutes a forfeiture of the right to life. Or consider the following casebook hypothetical: "A five-year-old child picks up a loaded pistol and starts running toward a police officer. The child shouts that she is a robber who will shoot the police officer dead, then giggles, and pulls the trigger. The first shot misses, but she laughs, keeps approaching, and says she will shoot again." If the officer shoots the child in self-defense, and we decide that doing so is justified, does it really make sense to suggest that this is because the child forfeited a right to life? The innocent aggressor problem cannot be avoided by claiming that no net harm results when one innocent person (the "victim") takes the life of another innocent person (the "aggressor") in self-defense, on the grounds that the value of the two innocent lives cancel each other out, because there may be cases of multiple innocent aggressors. Imagine several five-year-olds approaching the police officer with pistols in hand.
Perhaps, instead of focussing our attention exclusively on the conduct (or status) of the aggressor, in order to decide whether use of defensive force is justified, we ought to be focussing upon the defender as well. The objective harm theory of justification suggests that the defender has a right (i.e., is justified) to use defensive force in certain circumstances because doing so does not result in net harm. The challenge then becomes one of explaining how it is that the lives of the aggressors are worth less than the life of the defender. This seems to require some way to discount the lives of the aggressors, and this in turn suggests a forfeiture rationale. But if we abandon the objective harm theory of justification, we may be encouraged to look for some alternative explanation of the right enjoyed by the defender to use defensive force.\textsuperscript{82}

If an attempt to discount the value of the lives of aggressors will not satisfactorily account for multiple aggressor self-defense cases, are there any other ways to argue that such cases yield no net harm? One way to do so would be to argue that the death of the defender is not the only harm avoided or reduced by the act of self-defense. Perhaps the danger or alarm that would otherwise be produced by an aggressive attack would be reduced by the act of self-defense. But, as our earlier discussion revealed, Robinson would not be willing to include danger in the calculation of actual harm, since danger is merely the probability of future harm. And the inclusion of alarm would seem incompatible with a purely objective theory of justification.\textsuperscript{83}
3.5. Justification and Excuse

We have been examining the case that might be made for an objective theory of justification by appealing to an objective harm requirement of criminal liability. Our tentative conclusion, based upon a detailed examination of one such attempt, is that a plausible case has yet to be made. Indeed, the theory seems ill-equipped to provide a satisfactory rationale for cases of self-defense. Putting aside the question of whether justification defenses can be grounded on a defensible objective harm requirement, are there any other reasons for preferring an objective theory of justification? The arguments presented earlier on behalf of a harm requirement could, I suppose, be recast simply as arguments for an objective component of criminal liability, without necessarily identifying this component with harm. Justifications could still be regarded as negating this objective component, whatever one chooses to call it. These arguments would, however, be subject to many of the same objections raised to their original counterparts. Another reason, suggested by Robinson, for preferring an objective theory is that such an account of justification is necessary to preserve a sharp distinction between justification and excuse, a distinction that he thinks is to be drawn along act evaluation versus agent evaluation lines. Apparently he thinks that preserving the latter distinction requires that the evaluation of conduct be based exclusively on objective factors.

According to Robinson, failure to preserve a sharp distinction between justification and excuse has the following undesirable consequences for our system of criminal justice: (1) it undermines the deterrent function of the criminal law by confusing conduct that should be
condemned (and so deterred) but excused, from conduct that should be
couraged (or at least tolerated) because justified; (2) it encourages
injustice because judges fearful of sending the wrong message to the public
may refuse to exculpate defendants who should be excused; (3) it impairs
public confidence in the integrity of the courts, as a result of the confusion
alluded to in (1) and the potential for judicial inconsistency raised by (2); (4) it undermines our ability to readily distinguish those non-punishable
defendants in need of treatment from those that are not in need of
treatment; and results in (5) confusion over the permissibility of using
defensive force, as well as (6) confusion over the permissibility of third-
party assistance or interference.

The claim that the deterrent function of the criminal law would be
undermined by a failure sharply to distinguish justification from excuse is
plausible only if we assume that defendants will continue to be acquitted in
cases where they unjustifiably violate criminal statutes. If the law were
simply to do away with all excuses, there would be no legal distinction
between justification and excuse, and yet the deterrent function would not
be impaired. In fact, it may be slightly reinforced. Eliminating excuses
would also solve the judicial inconsistency problem, and the resulting
impairment of public confidence in the judiciary. The elimination of
excuses would, however, lead to intolerable injustice, and would likely
undermine public confidence in the legal system that would be at least as
detrimental as any attributable to a perceived judicial inconsistency. So, if
we assume that there are compelling reasons for preserving excuses, the
potential for a detrimental impact on deterrence and a judicial inconsistency born of refusal to honor excuses that the law recognizes, are serious concerns. The question, however, is whether it is necessary to adopt an objective theory of justification in order to avoid these undesirable consequences. I do not think so, but a full response will have to await presentation of an alternative theory of justification in chapter V. 90

Although in general it may be true that, whereas defendants acquitted on grounds of justification are not in need of treatment, those acquitted on excuse grounds do require some sort of treatment, one must be careful not to assume that all (and only) excused defendants require treatment. Surely we ought to distinguish, vis-a-vis treatment-worthiness, the insane defendant and the defendant who acted under duress. 91 Since Robinson seems to assume that an inference from "excused" to "treatable" is always warranted, he overstates the importance of a justification versus excuse distinction for the task of identifying defendants in need of treatment.

Evaluation of the fifth and sixth alleged consequences (confusion over the permissibility of using defensive force and confusion over the permissibility of third-party assistance or interference) will require a more elaborate response.

With regard to using defensive force to protect one's person or property, Robinson maintains that one is justified in using defensive force to repel an attacker if, but only if, the attack is not justified. A consequence is that defensive force may be justified against an excused attacker, but never against a justified attacker. 92 As for third-party liability, Robinson's view is that a third-party may assist any act that is justified, but may not assist
one that is merely excused. Moreover, third-parties may not interfere with the performance of justified acts. However, if an act is not justified, but merely excused, third-parties may interfere with its performance.93

Commenting on a raging fire example, Robinson has this to say about the right to use defensive force:

A person may or may not have an obligation himself to burn the field and save the town, but if burning the field is justifiable, it would be inconsistent to conclude that a person may lawfully prevent it. This is not to say that anyone who prevents the burning should be punished; if he was insane, or ignorant of the justified nature of the act, he might be excused. But only one of the acts can be justified, and the criminal code should require the nonjustified act to yield to the justified.94

In this passage Robinson is endorsing what we earlier identified as the "incompatibility thesis":95

(IT) If two acts are incompatible, then if the performance of one is justified, the performance of the other cannot be justified.

Recall that two acts are "incompatible" if the successful performance of one precludes the successful performance of the other.96 It follows from (IT) that if an attack is justified, the victim of that attack may not use defensive force to prevent the attack from being successful. (IT) has implications for third-party interference as well, assuming the interference comes in the form of an incompatible act. But is (IT) plausible? Consider the following two examples. In the first A is pushed into a deep and narrow well by C. B, who is working at the bottom of the well, looks up to see A plummeting down and realizes that if A’s fall is allowed to continue uninterrupted B will
certainly be killed by the impact (let's assume that there is insufficient room to move out of the way). Suppose that if A lands on B, A will survive the impact (B's body cushioning the fall) and that B realizes this. If B were to pull out his trusty laser pistol and disintegrate A in "self-defense," his doing so would seem to be justified (not merely excused). Now let us take the example a step further. A, who has been pushed down the well, observes that B is about to disintegrate him. A pulls out his laser pistol, sets it on "stun," shoots B before B can get off a shot, and then uses B's body to cushion his own fall. A's act would also seem to be justified. And yet the acts of A's shooting B and B's shooting A are incompatible. So it seems false that if two acts are incompatible, they cannot both be justified.97

In the second example a set of Siamese twins are born sharing a single heart and liver. The heart, as luck (or God) would have it, is enclosed in one twin, while the liver is enclosed in the other. Let us suppose that there are no other available donors, that if an organ transplant is not performed promptly both twins will die, and that the risks associated with transplanting the heart and liver are identical. Clearly, there is no single optimal state of affairs that the attending physician can bring about. There are two equally desirable (and incompatible) states of affairs, either of which the physician would be justified in bringing about. The physician would be justified in operating to save one of the twins (at the expense of the other), no matter which twin is chosen. Once again, we have two incompatible acts and yet, contrary to (IT), the fact that one is justified does not preclude the other from being justified.
Turning now to the issue of third-party rights to assist or interfere, we find Robinson saying the following:

If an act is justified, the conduct of anyone assisting in the act should also be justified. But if an actor is excused, the excuse should appropriately be limited to him and should not extend to others. If X robs a bank because he is insane, Y, his accomplice, is not shielded by X’s insanity. Unfortunately, the individualized character of excuse is often applied to justification because of confusion in current legal theory; it is currently no defense that the actions of the principal, identical to the defendant’s are deemed justified. It is possible, then, that the acts of one person will be found justified, while the identical acts of another will not. 98

In this passage Robinson is highlighting the first of the two issues mentioned above in connection with third-party rights, the right to assist. According to Robinson, justifications operate to "shield" third-parties from liability in a way that excuses do not.

It may be worthwhile to speculate about why Robinson thinks an act’s being justified serves to shield an accomplice from criminal liability. The answer suggested in the above quote is that he regards both the principal and the accomplice as performing the same act. 99

If we assume that being justified is a genuine property of acts, or is a function of such properties, and we appeal to the Leibnizian principle that identical acts have all the same properties, it would be inconsistent to hold that the principal’s act is justified, while holding that the identical act performed by an accomplice is not justified. If this is what is meant by the "shielding effect" of justifications, it may be viewed as a joint implication of the thesis that being justified is a property of acts, or a function of such properties, and Leibniz’s Law. So understood, the liability shield is a
conditional one. It may be expressed by a conditional statement to the
effect that if two acts are identical, then if one is justified, so must the other
be. To derive a non-conditional conclusion regarding the non-liability of
any particular accomplice, we need two additional statements: (i) one
asserting that the principal’s act is justified, and (ii) another asserting that
the act of the accomplice is identical to the act of the principal.

It is worthwhile to observe that if the shielding effect of justifications
is interpreted in this way, as in effect based upon a consistency requirement,
a similar (though slightly more complex) shielding principle for excuses
may be derived by conjoining the thesis that being excused is a property of
agents, or a function of such properties, with Leibniz’s Law. The result is a
conditional statement that if two agents are identical and they perform
identical acts, then if one has an excuse for performing the act, the other
must also have an excuse for performing the act. Deriving from this
principle a non-conditional conclusion regarding the non-liability of an
accomplice requires three additional statements: (i) one asserting that the
act of the accomplice is identical to the act of the principal, (ii) one
asserting that the principal has an excuse for performing the act, and
(iii) one asserting that the accomplice is identical to the principal.

Now one might think that all of this is quite silly. After all, in any
real case where the liability of an accomplice is at issue the principal and
accomplice will be different people. No doubt that is true. But the same
can be said of the acts performed by any two agents in any real case where
accomplice liability is at issue. Consider A and C working side by side to
burn B’s field. Aside from the fact that they may have different beliefs about the nature of their actions or the nature of the surrounding circumstances, and their actions may be differently motivated, it is clear that the bodily movements of the two actors take place in different physical locations and may in fact not even resemble one another (suppose A uses matches to light a fire at the edge of the field, while C throws a hand grenade from the comfort of his automobile). On the theory of human action presented in Chapter II, the acts of A and C cannot literally be identical. On the other hand, we do compare acts performed by different people in different locations, just as we compare their beliefs, motives, hair color, etc. We very naturally say that A performed the same (or a different) act, with the same (or different) beliefs and motives. There must be some way of making sense of these comparisons even if A and C cannot literally share identical beliefs, motives and actions.

The obvious move to make at this point is to appeal to the type-token distinction utilized in Chapter II. We must distinguish belief, motive, and act-types from belief, motive, and act-tokens. When Robinson claims that two people working side by side perform the same act, he might be understood as saying that they exemplify the same act-type. But of course at any given time each person may be exemplifying an indefinitely large number of act-types. Some of them may be the same as those exemplified by the other person; some may not. While working side by side in the field, A and C both exemplify the act-type of flexing the muscles in their right arm. But one could not plausibly argue that since A and C performed (exemplified) the same act-type, if A’s exemplification of the act-type was
justified, so must C's exemplification be justified. Suppose A's flexing his muscles level-generates the act-tokens of starting a fire and saving the town, while C's flexing his muscles level-generates the act-tokens of hitting B with an axe and killing B. We have, then, the following dilemma: a liability shielding principle for justifications can be derived from Leibniz's Law, together with the claim that two act-tokens are identical. But in no real case of accomplice liability will the act-tokens of the principal and accomplice be identical. On the other hand, there will be cases where the principal and accomplice exemplify identical act-types, but application of Leibniz's Law to individual act-types does not yield a plausible liability shielding principle.

If we are to make sense of Robinson's argument, we must appeal to the more "coarsely grained" concept of an act described in Chapter II; one that allows us to include the exemplification of several act-types as part of the "same act." Using this concept, a determination of whether an act is justified may be analyzed as an "all things (act-tokens) considered" evaluation. Identity of acts performed by principal and accomplice is no more likely than is identity of individual act-tokens. But we may still make sense of the locution, "same act." Suppose we refer to those morally relevant act-types that bear on the issue of whether conduct is justified as "justification relevant" act-types. The claim that A and C are performing the same act (so that if A is justified, C must be as well) may then be understood as a claim that (i) A's act includes act-tokens of justification relevant act-types by virtue of which the act is justified, (ii) C's
act includes act-tokens of the same justification relevant act-types, and (iii) C’s act includes tokens of no other justification relevant act-types. Sameness of act does not require act identity. To make this clear, one might use the expression, "relevantly the same." Determining whether A and C’s acts are relevantly the same requires that one identify those act-types that are justification relevant. And, as already noted, this identification requires (or presupposes) a normative moral theory. 103

According to Robinson, the justification relevant property is "produces no net harm." Moreover, as we have seen, he adopts an objective account of harm. 104 So, at least one of the act-types exemplified by both agents, if their acts are to be deemed relevantly the same, must be "produces no net harm." This requirement obviously is not sufficient to operate as a criterion of relevant sameness of act. Suppose A burns B’s field, thereby saving the town, while C breaks into a cabin in order to avoid freezing to death. Since the context is criminal law justification, and since (on Robinson’s account) justification presupposes that a criminally forbidden and harmful act has been performed, it might make sense to say that two acts are relevantly the same only if (i) they both include tokens of the same criminally forbidden act-type(s), (ii) these tokens level-generate tokens of "avoiding harm h," or "bringing about benefit b," and (iii) the latter tokens level-generate tokens of "producing no net harm." 105 We will not concern ourselves here with the details of what might be necessary in order to provide a complete analysis of relevant sameness of act. The point to notice is that according to this interpretation of Robinson’s claim that A and C are performing the "same act," the claim presupposes a "harm criterion" of
Theorists who agree with Robinson (not to mention those who do not) that acts are justified, while agents are excused, but who are inclined to reject Robinson's objective (harm) theory of justification, will not accept this criterion of relevant sameness of acts. They may wish to distinguish between the act of C, who burns B's field in order to harm B, and that of A, who burns B's field in order to save the community, even though the act of each saves the community. They would not be persuaded by Robinson's claim that since A and C are performing the same act, if A is justified, C must also be justified. They would insist that A and C are not performing the relevantly same act. As we have seen, Robinson's position to the contrary assumes that for purposes of deciding whether two acts are the same vis-a-vis justification, the only thing that matters is whether they have the same consequences vis-a-vis harm. But this amounts to assuming an objective (harm) theory of justification. Objective theories of justification cannot, therefore, derive any support from alleged consequences of adopting an objective (harm) criterion of relevant sameness of conduct. We must conclude that Robinson's effort to support his objective theory by appealing to its ability to dispell confusion that allegedly exists regarding the conditions of justified use of defensive force and accomplice liability does not succeed, for it assumes the truth of the very type of theory it is designed to support.

Earlier it was suggested that a liability shielding principle analogous to that attributed to justifications could be formulated for excuses. It is now
time to elaborate on that suggestion. We may begin by granting that the two agents in any realistic accomplice liability situation will not be identical. But there is no reason why a concept of relevant sameness of person may not be substituted in a way analogous to the earlier substitution of relevant sameness of act. Let us suppose that A and C are working side by side in an effort to burn B’s field. A, we shall suppose, is suffering from a psychological disability resulting in a delusional belief that a raging forest fire is threatening the town. If we invoke the shielding principle for excuses formulated earlier, modified by substituting "relevantly the same" for "identical," it would follow that A’s excuse can "shield" C from liability if (i) the acts performed by A and C are relevantly the same (e.g., burning the field is not necessary to save the town), (ii) by virtue of the disability and resulting delusion, A has a valid excuse, and (iii) C exemplifies all and only the same excuse relevant properties as exemplified by A (e.g., C too has a delusion that there is a raging forest fire, etc. which resulted from the same type of psychological disability).

It would seem, then, that if we take the liability shielding effect of justification to rest upon a consistency requirement applied to evaluations of conduct, it fails to provide an interesting contrast between justifications and excuses, since an analogous liability shielding principle for excuses may also be formulated. This is true whether or not one adopts Robinson’s harm criterion of sameness of act.
3.6. The Case Against an Objective Theory

Having examined and critiqued the best case offered to date for an objective theory of justification, Robinson's defense of an objective harm theory, it is now time to take a look at what can be said against objective theories as a class. The first two arguments that we shall examine derive from two different conceptions of justification: an "implicit elements" conception and a "license" conception. According to the implicit elements conception, justifications are implicit exceptions or qualifications to prohibitory norms that are incompletely expressed in criminal statutes. According to the license approach, a justification is to be understood as a license, liberty, or privilege to violate a prohibitory norm that is fully expressed in a criminal statute. These two conceptions form the basis of two arguments against an objective theory of criminal law justification that I shall call the "privilege" and "exception" arguments. Both arguments come from George Fletcher.

3.6.1. The Privilege and Exception Arguments

The first argument is a conceptual argument, based upon what it is to "exercise a privilege":

[T]he act of "exercising" or "acting under" a privilege presupposes knowledge of the justifying circumstances. To derive this conclusion, we need only inquire what we mean when we say that someone has "acted in" self-defense or has "exercised a privilege" generated by the imminent risk of danger to others. What we mean in these cases, I take it, is that the actor advert to the relevant circumstances in acting. If he is ignorant of the impending aggression or risk to others, he may act so to benefit himself or others, but it is not correct to say that he has "exercised" his privilege. . . . [W]hat we do with
privileges is exercise them. We do not "obey" or "comply" with them as we do with norms. We "exercise" them, and this notion has built into it the requirement that we know of the circumstances that permit us to justify our conduct.109

I think that one could challenge this analysis of "exercising a privilege." Imagine, if you will, someone immigrating to the United States who is totally ignorant of the scope of First Amendment protections. Coming from an oppressive country, the person mistakenly thinks that openly criticizing certain government policies, would likely result in severe punishment. Nevertheless, the person has occasion to do so. As expected, there is an arrest. At the preliminary hearing the defendant is astounded to hear the attorney from the public defender's office argue in defense that "my client was merely exercising the First Amendment privilege." I do not think that such a statement violates any linguistic rules concerning the proper use of "exercising a privilege."

But even if we grant Fletcher that exercising a privilege does require that the exerciser "advert to the relevant circumstances in acting," he has provided no reason for thinking that valid justifications require the exercise of a privilege. Why is it not sufficient that the conduct falls within the scope of a privilege? If being privileged to do X is having a right to do X, there is no reason to think that being privileged in general requires knowledge (or even belief) of the privilege-creating circumstances. I may have property rights of which I am unaware. Suppose my great aunt, whom I despise, dies leaving to me some parcel of land without my knowledge. In celebration of her death, I set fire to a section of the property that I now own. Others may not burn the land, but I am privileged to do so. Being
privileged (and perhaps even exercising that privilege) does not require knowledge of the circumstances by virtue of which I am thus privileged. I should think that an objectivist like Robinson might say similar things about the privilege of self-defense and other justification defenses. There may be circumstances by virtue of which I am privileged (i.e., have a right) to use defensive force even though I am unaware of the relevant circumstances. So much for the privilege argument.

Fletcher's second argument focuses on the conception of justifications as exceptions to prohibitory norms:

As exceptions, these claims should be available only to those who merit special treatment. Injecting the element of merit makes the actor's intent relevant, for objective circumstances alone cannot establish the special merit of someone who treads upon the protected interests of another. For someone who violates a norm to deserve exceptional treatment, he or she must at least know of the circumstances supporting the claim of an exception.110

This argument seems to assume that justifications are special dispensations, rather than defenses resting on entitlements. On the dispensation view, one must "earn" the right to rely upon the defense by having a "meritorious intent (or knowledge)." But if justifications are based on entitlements, then it is not at all obvious that one must "earn" the right to assert them. If Victim has a right to kill Aggressor because Aggressor is about to kill Victim, why should it matter that Victim is unaware that her life is in jeopardy? It is far from obvious that the simple fact that such defensive killings are exceptional mandates knowledge of justifying circumstances. Of course, Fletcher might respond that the concept of "exercising a right"
requires such knowledge. But that move will be no more successful here than in the privilege argument. There may be good reasons for including a meritorious intent, knowledge, or motive among the validity conditions of a justification defense, but these two arguments do not provide them. 111

3.6.2. Justification and Act Evaluation

One of the assumptions operating throughout Robinson’s discussion of criminal law justification defenses is that justifications essentially involve evaluations of conduct. In an effort to preserve what he sees as an important distinction between justification and excuse, Robinson opts for an objective theory of act evaluation and individuation that is then incorporated into an objective theory of justification.

In Chapter II reasons were given for thinking that the distinction between act and agent evaluation is not as sharp as is sometimes thought. We also cautioned against any attempt to define "act" so as to purge it of all mental elements, with an eye towards simplifying the task of act evaluation. The theory of human action developed in that chapter suggests that if one attempts to purge human acts of mental elements one will be left with pale behavioral act ghosts (bodily movements and their objective consequences), not the genuine article. And if one attempts to engage in moral or legal evaluation of these act ghosts, one will almost certainly omit much that ought to be considered in any overall moral or legal evaluation of the performance-of-an-act-by-an-agent.
Consider the example of the "arsonist," Arnold, who blows up several houses in the path of an oncoming fire in an effort to save the rest of the community. Unpredictably, the wind changes, diverting the fire in such a manner as to bypass the community. Arnold’s deliberate destruction of the houses was not necessary after all. Had the wind not changed, it seems clear that on Robinson’s theory Arnold would have a justification defense. His conduct would have resulted in no net harm. But an event which Arnold could not have reasonably predicted and over which he had no control, the change in direction of the wind, prevents Arnold from having a justification defense.

Nearly everyone (including Robinson) would agree that it would be unjust to punish "unlucky" Arnold for having unnecessarily destroyed the houses. Of course this need not be the result of denying Arnold a justification defense. As Robinson would be quick to point out, Arnold may have an excuse for his unfortunate conduct, based upon his reasonable but mistaken belief that blowing up the houses was necessary to save the rest of the community. But somehow this does not seem an entirely satisfactory response. Consider a variation on the arson theme. Arnold is a terrorist who, completely unaware of the oncoming fire, blows up precisely the houses necessary to save the rest of the community. If, as Robinson suggests, it is only the objective harm-relevant consequences of the conduct that matter, Arnold the terrorist has a justification defense to criminal liability. Not only that, but if we take seriously other things Robinson says about justified conduct, it would seem that we ought to regard the terrorist’s conduct as "correct behavior," to be "not only tolerated but encouraged."
Meanwhile, unlucky Arnold, who blew up the houses in order to save the community but whose plan was unpredictably thwarted by Mother Nature, has no justification for what he did. His conduct is to be condemned. Once again, he may avoid punishment on excuse grounds. But I think that we would understand his puzzlement if he were to query why his conduct was being condemned, while that of his terrorist counterpart was not.

Is there not a sense in which what unlucky Arnold did was praiseworthy, while what terrorist Arnold did warrants our strongest condemnation? Granted, the objective harm-relevant consequences of unlucky Arnold’s conduct were undesirable, while those of terrorist Arnold’s conduct were desirable. But something important seems to be omitted when we confine ourselves to an evaluation of the objective harm-relevant consequences of conduct.

If we were to examine unlucky Arnold’s conduct in some detail, we would find that it consists of a number of act-tokens, some of them being quite praiseworthy. One of the things that he did was attempt to save a large segment of the community from the ravages of fire. Another was to exercise reasonable care in ascertaining the need to destroy the houses. Yet another was to bring about no more destruction than seemed necessary. All of these act-tokens are desirable and are the sorts of acts one would want to encourage. Of course, another constituent act-token was that of bringing about the unnecessary destruction of houses. Another was bringing about a substantial amount of net harm. Certainly these tokens are undesirable. But if we were interested in an "all things considered" moral assessment of
Arnold’s conduct, we would do well not to exclude the first three act-tokens from our purview.

No provision is made within Robinson’s defense classification schema for features of conduct, other than harm-relevant consequences, to relieve an agent of criminal liability for engaging in that conduct, if the conduct produces the kind of harm that a criminal statute is intended to prohibit. But this seems a serious omission. Suppose that, to slightly modify the earlier hypothetical, Arnold reasonably calculates the speed and direction of the oncoming fire and determines that two houses must be blown up in order to save the community. One of those houses happens to be his own. Contained within the house are many items having great personal value (e.g., family photographs, video recordings of his children at various stages of development, personal letters, assorted gifts, the only existing manuscript of his first publishable book, etc.). According to Arnold’s (reasonable) calculations, if he were to take the time to retrieve any of these items, there would be insufficient time remaining to set the explosive and blow up the two houses. With great anguish Arnold chooses to save the community. Immediately after blowing up the houses, the wind changes, etc.

We will assume that on balance Arnold’s conduct caused more harm than would have resulted had he not acted as he did. On Robinson’s theory, if we are to relieve Arnold of liability for blowing up his neighbor’s house, we must do so on excuse grounds. The conduct, since it resulted in net harm, is to be condemned; but we may let Arnold off the hook on the grounds that he was suffering from such a disability at the time as to prevent
his conduct from being the result of a meaningful exercise of free will. But where is the disability? And why would one think that Arnold did not freely choose to do what he did? Moreover, his conduct had many desirable features. It was a careful and reasonable response to a serious threat to the welfare of others. It was performed with the best of motives. It involved great self-sacrifice. Yes, it resulted in net harm, but considering what was potentially at stake, a modest amount. Wasn’t Arnold’s conduct, all things considered, more desirable than not? Would we not want anyone similarly situated to act precisely as did Arnold? Why refrain, then, from classifying his conduct as "correct," as conduct to be "not only tolerated but encouraged," in short, as "justified"?

So even if one insists on regarding justifications as essentially involving conduct evaluations, there is good reason not to focus merely upon "objective" features of those acts. However, one need not wed justification defenses to act evaluation in the manner suggested by Robinson and many other legal and moral theorists. The view to be developed and defended in the following two chapters departs from "main line" thinking about justification defenses on just this very point. The account to be defended is what I call a "character trait" or "motivational" theory of justification. It has much in common with character trait theories of excuses. But even theorists who espouse the latter rarely embrace the former. In this regard I strike my own path.
ENDNOTES

[1] Robinson’s first, and most detailed, exposition of his theory occurred in a 1975 article, *A Theory of Justification: Societal Harm as a Prerequisite for Criminal Liability*, 23 UCLA L. Rev. 266 (1975) [hereinafter cited as *Justification*]. He has since refined the theory. One such refinement is his recognition of a distinction between justifications and offense modifications. Nevertheless, the theory remains substantially as it was originally formulated. See his discussion in Defenses chs. 2 & 3.


[4] In fact, it is essential to Robinson’s view of justifications that some harm be occasioned by the conduct that is to be justified, and that the harm occasioned be normally sufficient to support liability. Otherwise, his own distinction between justifications and offense modifications breaks down. An offense modification is based upon a claim that the conduct in question did not cause statutory harm. A justification, by contrast, acknowledges that the conduct caused statutory harm, but argues that the conduct also resulted in the avoidance of an even greater harm or the promotion of a good sufficient to outweigh the statutory harm caused. Offense modifications and justifications both rest on a "no harm" rationale, but the harm they negate differs.


[6] Id.
This is not an uncommon move in the law. One of the more notorious examples is the tort concept of "proximate cause." See W. Prosser, Handbook of The Law of Torts 244-50 (4th ed. 1971).

Justification 267 n.7 (emphasis added).

Supra note 3.

As Feinberg construes the task, it is "a quest not for useful policies but for valid principles." Harm 4. In saying this, he is relying on Ronald Dworkin's distinction between policies and principles. Harm 246 n.3; R. Dworkin, Taking Rights Seriously 22-28 (1978).

Harm 7.

Harm 9.

Harm 9.

Harm 14-15.

Harm 10.

Harm 10. The last sentence in this quote seems to assume that in order for something to be a "morally relevant" liberty-limiting reason, there must be some circumstances in which it would be individually sufficient to outweigh the presumptive case for liberty. This is why, if there is more than one always relevant liberty-limiting reason, Feinberg thinks that none can be a necessary condition for justified state coercion. But why need we assume this? Suppose that some factor, e.g., the invasion of an interest (harm), is always required in order to justify legal coercion. Also suppose that in some circumstances, even though not sufficient by itself to overcome the presumptive case for liberty, if conjoined with certain other factors (e.g., offense or immorality) it is sufficient to win the day. Finally, suppose that these other factors are never individually or jointly sufficient to warrant legal coercion; harm is always required. This is certainly conceivable. If it were to obtain, I see no reason to deny that harm, as well as the other "coercion relevant factors" are "morally relevant" liberty-limiting reasons.

Harm 26-27.
In regarding both justification and excuse as negating criminal wrongdoing, Feinberg's view is in sharp contrast to that of George Fletcher. G. Fletcher, Rethinking 759. Feinberg also regards consent as negating wrongdoing. On this he and Fletcher agree. Rethinking 552-54, 566, 698. But, in contrast to Fletcher, Feinberg seems to view consent as neither a justification nor an element of the offense definition. Id.

These last two interpretations may or may not be genuine options. Whether they are would seem to depend on whether every indefensible or unjustified setback of interests must necessarily violate a moral right, i.e., whether in order for a setback of interests to be indefensible or unjustified, it must violate a moral right.

Recall that to have an interest in X is to have a stake in the well-being of X, and to have a stake in X is to stand to gain or lose depending on the nature or condition of X. Supra 78.

Justification, 269. In this respect Robinson's harm requirement departs from Feinberg's harm principle, which takes prevention of actual harm or the unreasonable risk of harm to be a relevant reason for criminal coercion. Feinberg, Harm 11.

MPC §§ 5.05(1), 6.06, 210.6.

MPC §§ 6.06, 210.4.

Such a lottery would correspond to a jurisdiction that fails to criminalize attempts. One who attempts murder, but fails, receives no
penalty; while one who attempts and succeeds may receive life imprisonment or death. That no American jurisdiction goes so far as to eliminate attempt liability is some concession, I would say, to the obvious injustice of treating relevantly similar persons in radically different ways. The lottery objection being developed above is designed to suggest that injustice remains so long as any significant difference of treatment exists between those who engage in conduct that fails to cause harm and those who engage in the same type of conduct that happens to cause harm.

32 The point being made here is independent of the issue of detection. We may assume that law enforcement officers observe from above the driving of Jones and myself. In a Model Penal Code jurisdiction, if my driving were judged to be merely the result of gross negligence, I would not be chargeable with any criminal offense. If it were deemed reckless, then I would probably be guilty of "recklessly endangering another person," a misdemeanor ordinarily punishable by a maximum one year sentence. Jones, on the other hand, could face up to ten years if convicted of manslaughter (reckless homicide), and five years if convicted of negligent homicide. MPC §§ 210.3, 210.4, 211.2, 6.06, 6.08.


34 J. Feinberg, Deserving 217.

35 Deserving 218.

36 Deserving 218.

37 Deserving 219. The application that Feinberg makes of the principle of weak retributivism seems to me less than persuasive. After all, we are contemplating trading off a certainty of harm that the criminal will suffer against a (perhaps very low) probability of deterrent efficacy of imposing that harm, and thereby reducing the amount of harm suffered by innocents. It is not obvious how this calculation comes out.

38 Deserving 218.

39 Justification 266-67.
Three factors contribute to this initially surprising result: (1) the relative infrequency of homicides that result from negligent or reckless driving, (2) the near universal human optimism that inclines people to underestimate the probability that their risky behavior will in fact result in serious harm, and (3) the "natural sanctions" that attach to such risky behavior in the event it leads to harm (e.g., property damage, potential for serious physical injury or death), which make any additional legal sanctions superfluous. S. Schulhofer, *Harm and Punishment: A Critique of Emphasis on the Results of Conduct in the Criminal Law*, 122 U. Pa. L. Rev. 1497, 1543-44 (1974). A particularly good example to illustrate the claims made in the text would be the proposal to punish parents who negligently (or recklessly) fail to use child restraining devices, but only in cases where there is a resulting injury to the child. It is hard to believe that there would be no gain in deterrence from implementing some mild punishment (a fine) for anyone who is caught not using the device. Moreover, given the obvious "natural sanction" that attaches to nonuse (severe injury or death to a child), it is doubtful that punishing much more severely in those cases where serious injury results will have any detectable deterrent effect. In light of these reflections, one must wonder about the wisdom of a California prosecutor's recently reported decision to charge the parents of a child killed in an auto accident with manslaughter for failing to use a child restraining device.

Of course Robinson would object that such an unsuccessful attempt does cause a kind of societal harm. See infra 96.

Similarly, there is no reason to think that the goal of isolating the criminally dangerous mandates punishing only those who cause actual harm.

For a detailed examination of the issue of whether an emphasis on harmful consequences in the criminal law can be justified, see S. Schulhofer, *Harm and Punishment: A Critique of Emphasis on the Results of Conduct in the Criminal Law*, 122 U. Pa. L. Rev. 1497 (1974). Although there are some exceptions, Schulhofer concludes that for the most part such an emphasis is not warranted.

For a detailed discussion of the "jury nullification" problem in the context of a discussion of the deterrence rationale, see S. Schulhofer, *Harm and Punishment* 1522-57.

Except perhaps where the death penalty is a possibility. See Schulhofer, *Harm and Punishment* 1603-04.

*Supra* 88.


This is Eser's move: "Consequently, 'criminal harm,' in its most general sense, is the negation, endangering, or destruction of an individual, group, or state interest which was deemed socially valuable, in harmony with the Constitution and, therefore, protected by a criminal sanction." A. Eser, *"Harm" in the Concept of Crime* 413.

Eser uses this ploy in response to examples of what he calls "simple conduct" crimes like carrying a dangerous weapon, perjury and forgery. *"Harm" in the Concept of Crime* 388-90.

See A. Eser, *"Harm" in the Concept of Crime* 388-89.


Consider the potential for authoritarian abuse associated with post-revolution Soviet law's definition of "crime" as "any act dangerous to the Soviet system." A. Eser, *"Harm" in the Concept of Crime* 360. In addition to the problems discussed in the text, infra, it seems to me that a definition of "harm" that encompasses injury to "intangible societal interests" might yield a harm requirement so anemic as to constitute no meaningful restraint on authoritarianism.

LaFave and Scott 440-46.

*Justification* 269.

Principles].

57Principles 143-44.

58Principles 144. Feinberg suggests that typically there is no interest in avoiding such anxiety or discomfort per se, although the anxiety or discomfort may, if severe enough to be debilitating, impair other interests, and so something that causes the anxiety or discomfort may be harmful in that way. See J. Feinberg, Harm 45-50.

59Bentham specifically identifies four such motives: benevolence, self-preservation, fear of shame, and fear of the divine displeasure. These motives are associated with the physical, political, moral, and religious sanctions, respectively. Principles 145-47.

60Principles 145-47.

61Harm 269.

62LaFave & Scott 445.

63LaFave & Scott 442.

64That Robinson conflates conduct and agent risk seems evident when he asks: "If one adopts the theory that dangerous but nonharmful conduct should be punished, then why not punish one who shakes his fist at a rain cloud, since he has shown his dangerous intentions?" Justification 269.

65As will become apparent in the course of my later endorsement, and defense, of a character trait theory of criminal responsibility, this is not the position I would take. Whether liability is appropriate in such attempt cases depends, on my view, on whether the attempt adequately evidences a blameworthy character trait. Thus, as between conduct risk and agent risk, I would suggest that it is agent risk that ought to control our attempt liability decisions.

66In this way one can account for the predominant view that simple attempts are appropriate occasions for criminal liability, whereas inherently or legally impossible attempts are not. See LaFave & Scott 425, 439, 446. The predominant view also has it that cases of factually impossible attempts should be treated like simple attempts vis-a-vis liability. Id. 439. On the
conducted and risk account this is because factually impossible attempts pose substantial risks of harm. This seems plausible in light of cases like that of attempting to rob someone who just happens not to possess any money. There is a certain risk of harm involved in any attempt to force someone to turn over their money, even if it is not always the harm of losing money. But what about the case of one who lights wooden sticks of "dynamite"? One possibility is that such a case should be treated as an inherently impossible attempt, like the voodoo doll example.

67J. Bentham, Principles 152. Intentionality, for Bentham, includes acting with "direct" and "oblique" intention. These two types of intentional action seem to correspond with what the Model Penal Code calls acting "purposely" and "knowingly" (with respect to a material result of the conduct). Bentham's consciousness corresponds (I think) with Model Penal Code knowledge of material attendant circumstances. The goodness or badness of intentions is a function of the (directly or obliquely) intended consequences (not the actual consequences) of the act. Motives, not being intrinsically good or bad, derive their goodness or badness from the quality of their consequences. The same is true of dispositions, which are "a kind of fictitious entity, feigned for the convenience of discourse, in order to express what there is supposed to be permanent in a man's frame of mind, where, on such or such an occasion, he has been influenced by such or such a motive, to engage in an act, which, as appeared to him, was of such or such a tendency." Principles 125. The goodness or badness of a man's disposition is, accordingly, indicated by the quality of the motive and intention with which he acts.

68Principles 152.

69Recall that the definition of "objective theory of justification" offered in § 1.2 excludes mental states from the validity conditions of the defense. Should it be the case, as many philosophers have argued, that mental states are not to be identified with "private" events, states or processes that exist or occur only "inside the skin," then at least some mental states might be (or have) behavioral features. In that case, we must understand Robinson, and other "objectivists," as maintaining that the validity conditions of justification defenses exclude those external features of a person that are (features of) mental states.

The main thrust of the "objective" theory is to deny that an agent's intentions, beliefs, motives, and other mental states (however construed) are relevant to determining the availability of a valid justification defense. In
fact, I think that most objectivists do adopt a "private state" view of the mental, and that this is what motivates many of them to opt for an objective theory.

Of course A would not be apprehended and hence would not be punished for an undetected attempt. But whether A is caught and punished is a separate issue from whether A is criminally liable to being punished (i.e., ought to be punished).

Actually, this is not quite accurate. Danger, for Bentham, is the increased probability that the original primary mischief of an act will befall unassignable individuals. In the case of an attempt, there is no original primary mischief. So the "danger" of an attempt would have to be the increased probability that the original primary mischief typically resulting from the completed offense will befall unassignable individuals.

There is another possible interpretation of the concept of "intangible societal harm" that should be mentioned, although I do not think that it is one that can be used to support a purely objective theory of criminal law justification. According to this view, the act-types proscribed by criminal statutes are to be selected by reference to the relative frequency with which tokens of those types result in actual harm. The "seriousness" of a given offense would, on this view, be at least partly a function of the degree of harm typically resulting from tokens of the proscribed act-type and the relative frequency with which tokens of that type do result in such harm.

The death of a human being is a serious harm indeed. If one is interested in reducing the incidence of deaths resulting from the actions of fellow humans, one may look for act-types, tokens of which tend frequently to result in such harm. Of course, one set of such act-types consists of those that require the actual killing of another human. The relative frequency (probability) of tokens of these act-types resulting in death will be 1! This will be true whether or not the act-type in question involves any of the Model Penal Code culpability elements of negligence, recklessness, knowledge, or purpose. But surely we do not need to limit ourselves to cases of actual homicides. The function of the law of criminal attempts is to identify relevant act-types, tokens of which do not necessarily result in actual statutory harm, but which do so frequently enough. Thus, one who unsuccessfully tries (i.e., acts with the intent) to kill another may be punishable because that person has exemplified an act-type, tokens of which very frequently do generate homicides.
If one were to follow this approach, one might be able to draw distinctions among the various types of impossible attempts on the grounds that, even though in the particular circumstances in which the attempts are made none of them can succeed, some of them are tokens of act-types with a substantially higher frequency of success than others. Thus, one who attempts to rob a person who just happens to have no money at the time has exemplified an act-type (trying to rob someone believed to possess money), tokens of which very frequently generate thefts. The man who shakes his fist at a rain cloud, expecting to thereby cause the destruction of the world, may be thought not to have exemplified an act-type with tokens that result in such harm with a sufficiently high relative frequency to warrant criminal prohibition. On these grounds one might distinguish "factually" from "inherently" impossible attempts. Similar comments would apply to other categories of attempts as well, e.g., attempts that simply fail and legally impossible attempts.

This way of distinguishing the various categories of attempts has some promise, but puzzles remain. In particular, the criteria for selecting the relevant act-types is far from obvious. Why not, for example, pick "trying to rob someone who does not possess money at the time" as the relevant act-type for purposes of estimating the relative frequency of harm caused by the type of factually impossible attempt discussed above?

Whatever promise the "frequency" interpretation of intangible societal harm may hold for an account of criminal attempt liability, it is doubtful that can be used to support an purely objective account of criminal law justification. Presumably, the frequency approach would attempt to identify "justification relevant" act-types by reference to the relative frequency with which tokens of those types result in no actual net harm, despite the violation of a criminal statute. Just as in the case of selecting the act-types to be proscribed by statute, one could require that the relative frequency with which tokens of justification relevant act-types result in no net harm be 1. This could be accomplished by requiring in each case that no net harm in fact result from the violation of the relevant statute. But this requirement makes no more sense in the context of identifying justification relevant act-types than in the case of identifying "criminalization relevant" act-types. Just as there may be good reason to discourage certain attempts to do harm that fail, so there may be good reason to encourage certain attempts to reduce net harm that fail. In that case, the frequency approach will likely select, as justification relevant, act-types that include mental states of the agent. A person who violates a criminal statute because the
person honestly (and reasonably?) believes that doing so is necessary to reduce overall harm ought to avoid liability for reasons analogous to those offered in support of attempt liability. If this argument is sound, the frequency interpretation will not support a purely objective theory of justification.

73*Supra* 1-2.

74Actually, it is not quite so clear as I suggest. Suppose the single aggressor is a medical researcher who is on the verge of finding a cure for AIDS. Further suppose that without the researcher’s continued efforts, a cure is unlikely to be found for at least a decade. If we include the probable savings of thousands of lives if the researcher is allowed to live, there may in fact be net harm that results from the use of defensive force.

75This assumes that one human life is equal in value to another. Some might object to any procedure that attempts to quantify the value of a human life, urging that the value of human life is "priceless." As I think that there are more serious problems for the no harm analysis, I only mention this potential objection in passing.

76See Feinberg, Harm 33, 36; *supra* 78.

77Another possibility might be to invoke Feinberg’s principle of "weak retributivism," viz., the thesis that if someone must suffer, it is better, *ceteris paribus*, that it be the guilty (or blameworthy) than the innocent (or non-blameworthy). *Supra* 84. This principle does not appeal to rights forfeiture to explain the asymmetry between guilty aggressors and innocent victims. It can, therefore, avoid the objections in the text that focus upon the concept of forfeiture. It cannot, however, provide a satisfactory account of self-defense against single or multiple innocent aggressors. See *infra* 113.

78Of course I do not intend this statement to rule out forfeiture by omission. The point is that the occurrence of the state of affairs that triggers a forfeiture must be within the forfeiting person’s control. It must be "up to me" whether I forfeit a right.

79Factual impossibility is not recognized as a defense to an attempt charge. See discussion of the difference vis-a-vis liability for factually, legally, and inherently impossible attempts, *supra* 95.
Alternatively, if rights are regarded as having the status of *prima facie* moral constraints on the actions of others, we may say that the wrongful act of an aggressor reduces the weight (or force) of the aggressor's *prima facie* right to life without extinguishing it. Thus, a potential "victim" who knows that the aggressor poses no genuine or immediate threat may not use deadly force in "self-defense."


See the discussion of self-defense *infra* 258.

It is not clear (to me at least) what the "frequency" approach discussed, *supra* note 72, would recommend regarding the multiple aggressor self-defense cases, except that it would not likely restrict the availability of self-defense to cases that result in no actual net harm (however that is to be understood). If we assume that certain types of acts of self-defense tend to result in no net harm, then presumably one would want to encourage tokens of those act-types regardless of whether in a given instance no actual net harm results. Thus, the frequency approach once again provides no support for a purely objective account of criminal law justification (in this case, self-defense).

We might follow Fletcher in referring to this component as "external (or objective) wrongdoing." Rethinking 471, 475-78.

*Justification* 276 n.45.

*Justification* 279 n.52; Defenses §§ 25(b), 177(c).

*Justification* 277-78.

*Justification* 279.

It is worth noting that it is not enough that the law officially recognize the two different grounds for exculpation. The basis for particular exculpations must be well publicized in order to insure that the public is clear about what each acquittal really means. Otherwise, the deterrent function may suffer despite an official recognition of the distinction between justification and excuse. The need for publicity would seem to favor special verdicts, patterned after the "not guilty by reason of insanity" verdict, over the general verdict which does not specify the
specific grounds for acquittal.

90\textit{infra} 257.

91According to Robinson, the agent who raises an excuse has admitted a "personal weakness" which "should lead to some sort of treatment." \textit{Justification} 279 n.2. This may seem reasonable for such disabilities as intoxication, somnambulism, automatism, hypnotism, subnormality and mental disease or defect, but it seems of dubious application to the state of coercion that grounds a duress excuse. On the typical formulation of duress the relevant state of coercion must be "caused by a threat that a person of reasonable firmness in [the agent’s] situation would not have resisted." Defenses § 177(a). If a person of reasonable firmness would not have resisted, it is hard to see why the resulting state of coercion is (or indicates) any abnormality that requires treatment. In an apparent attempt to deal with this kind of objection, Robinson suggests that for some excuses, such as duress, society may be willing to forego treatment and "be content to allow the abnormalities to continue in the belief that they will rarely produce harm because of the infrequency of such situations." \textit{Justification} 279 n.52. But again, why should one who does precisely what a person of reasonable firmness would have done be considered to suffer from an abnormality at all? Has Robinson confused the statistical and evaluative meanings of 'abnormal'? The situation of duress may be statistically infrequent, and in that sense states of coercion could be said to be abnormal, but it does not follow that the person suffers from some (evaluatively) abnormal condition that even in theory requires treatment.

Consider another example. Many theorists, including Robinson, are willing to recognize reasonable mistake as to the presence of justifying circumstances ("putative justification") as the basis of an excuse. But where is the abnormality here? It seems ludicrous to classify simply being mistaken as an abnormality warranting treatment. But I see no other candidate for the relevant disability. Of course, another possibility is that duress and putative justification ought to be regarded as justifications, rather than as excuses, in which case these two examples do not threaten the thesis that all excuses evidence some type of "personal weakness" or disability.

92How he proposes to support the claim that self-defense is justified against innocent (excused) aggressors (especially if there is more than one) is a mystery to me. See the discussion of this point in the preceding section, \textit{supra} 113.

Robinson, *Justification* 278 (emphasis added).

*Supra* 6.

*Id.*

The example comes from R. Nozick, *Anarchy, State, and Utopia* 34 (1974). Here is another example that makes the same point: A is coming down a mountain road, rounds the corner and sees B in his car at the edge. There has been a landslide. The angle of the slide with the road is such that if A’s car hits it, the car will careen into B’s, knocking it and B off the edge. The force of the collision will, however, stop A’s car from going over. A applies breaks, but to no avail. It would seem permissible for B to disintegrate A and A’s car. But A may also stun B and use B’s car to stop his own.

*Justification* 279.

Additional support for this claim can be found in a footnote on the same page as the quoted passage. After lamenting the failure of current law to consistently apply the distinction between justification and excuse to third-party liability issues, Robinson writes: "Thus the present system might, in the raging fire example, consider one actor justified in burning the field yet still consider a companion unjustified in assisting in the identical act." *Justification* 279 n.51.

Robinson also regards the shielding effect as a consequence of the "universality of justification." Excuses, being personal to the agent (i.e., lacking universality), do not for that reason have a liability shielding effect. If A were excusably mistaken about the need to burn B’s field in order to save the community (e.g., he was suffering from a psychotic delusion), A may avoid liability, but C is not shielded by A’s excuse. Unless C has a personal excuse himself (e.g., he too is suffering from a delusion), C may be held liable for helping to start the fire. *See Justification* 279 n.51.

In fact, on the Goldmanian theory described in Chapter II, the simple fact that two different agents are involved is sufficient to guarantee the non-identity of the acts. *See supra* 31.
An individual act-token may be said to be "justified" if it is a constituent of a justified act.

This usage allows for the possibility that there are morally relevant act-types that are not relevant to whether an act is justified. After all, the predicate "is justified" is only one of several moral predicates applicable to acts.

Supra 55.

This explains why he insists that if A and C work side by side to burn B's field and doing so saves the town, such issues as C's prior fault or evil intent or motive do not prevent C from being justified in burning the field, if A is justified in doing so. See Justification 279 n.51, 287, 289.

Of course he would not require that all of the consequences of two acts be the same in order for the acts to be relevantly the same. Suppose that A and C are working side by side burning the field, but that C is a minor league pitcher who was to be the starting pitcher in a game that day. Instead of going to the game, C decided to help A (who is not a minor league pitcher). As a consequence of C's helping to burn the field a different pitcher had to be substituted to start the game, but otherwise things turned out as they would have if C had started (the team lost). In this case C's act clearly has at least one consequence that A's lacks, but surely this difference would not prevent Robinson from regarding their acts as "relevantly the same."

It is interesting to note that if we apply this criterion of act sameness to Robinson's own example of two agents working side by side to burn a field in order to save the community, it does not follow that the acts of the two agents must be the same. Suppose C was on his way to defuse a nuclear time bomb and his stopping to help A burn the field resulted in the destruction of a city much more populated than the community saved by burning the field. In that case there would be a significant difference between the acts of A and C vis-a-vis net harm caused. Presumably, A would be justified in burning the field, while C would not. In this example, condition (iii) of the proposed set of necessary conditions of relevant sameness of act is not met. Query: What does this example show about the claim that if the principal is justified in performing a, then an accomplice must also be justified in assisting in the performance of a?

See supra 121.
The terminology is Husak’s. D. Husak, Philosophy 189.

For a discussion of some possible implications of each conception, concluding with a tentative recommendation of the license conception, see Husak, Philosophy 190-92.

G. Fletcher, Rethinking 564-65.

Rethinking 565.

See infra 234-35, for further discussion of this issue.

Robinson, Justification 278.

Robinson, Justification 274.

Analysis 221.
Chapter IV

4.1. Character Traits and Excuses

It is difficult to provide a satisfactory account of the logic of criminal law justification defenses without taking a stand on the logic of closely related defenses. This is particularly true of criminal law excuses, which are frequently classified, along with justifications, as "exculpatory" defenses. With this in mind, this chapter is devoted to presenting an account of the logic of criminal law excuses.

Following H.L.A. Hart's suggestion, it has become common in criminal theory to regard a theory of excuses as part of a broader theory of criminal responsibility, where the latter is concerned to provide an account of all psychological requirements of criminal liability. In addition to clarifying the nature of excuses, such a theory of responsibility will have implications for the appropriateness of including Model Penal Code mens rea elements in statutory definitions. It will also bear on such matters as attempt liability and mistake of fact defenses. I shall argue that it has important implications for justification defenses as well.

Given the resurgence of interest in "virtue ethics," it is not surprising that some contemporary philosophers have placed moral character at the center of their account of criminal responsibility. Assuming a proper
understanding of the nature of moral character traits, this emphasis upon moral character is warranted. Although there may be other ways of incorporating an emphasis on moral character into an account of criminal law excuses, a *prima facie* plausible way to begin is by endorsing the following three theses:

**PT:** Criminal punishment is morally justified only if the person punished performed an act that evidenced an undesirable character trait.

**CT:** A character trait is a disposition of a person.

**ET:** Criminal law excuses defeat a normally warranted inference from conduct to an undesirable character trait.

A number of questions immediately arise when one considers these principles: (1) Why would one think that criminal punishment is morally justified only when conduct evidences an undesirable character trait? (2) What type of character trait is required? (3) What kind of disposition is a character trait? (4) What kind of conduct normally warrants an inference to the relevant kind of character trait? (5) How precisely do excuses "defeat" this normally warranted inference? One contemporary legal philosopher who accepts (under appropriate interpretations) PT, CT and ET is Michael Bayles. In an effort to find satisfactory answers to the questions raised by these theses, we shall take a brief look at Bayles' theory of criminal responsibility. Dissatisfaction with some of his answers to the questions will set the stage for further exploration.3

The response to the first question is that Bayles thinks PT is plausible because it follows from two other theses that he accepts:
PT₁: Criminal punishment is morally justified only if the person punished performed an act that evidenced that person's blameworthiness.⁴

BT: A person is blameworthy if and only if the person has an undesirable character trait.⁵

Bayles does not attempt to derive these two principles from further premises. It may be that he thinks they receive support "indirectly," by virtue of being fundamental constituents of a theory of criminal responsibility that has great explanatory power or overall plausibility. In order to assess Bayles' argument for PT, we must assess these two premises, as interpreted by Bayles. We begin with his answers to the second and third questions identified above.

Unfortunately, Bayles does not provide a clear answer to the third question regarding the kind of disposition that is a character trait. Some of his examples suggest that the relevant disposition is a behavioral disposition,⁶ as opposed, for example, to a cognitive or motivational one.⁷ But beyond this, it would be risky to speculate. As for the second question, we should really distinguish two versions: (2a) What type of character trait is required for blameworthiness? (2b) What type of character trait is required for morally justified criminal punishment? An answer to (2a) can be gleaned from BT, together with Bayles' claim that "'character trait' is not, as in the Aristotelian view, restricted to traits which people can voluntarily control possessing or manifesting in behavior. Instead, it refers to any socially desirable or undesirable disposition of a person."⁸ If we take this description of character traits seriously, it would seem that we are
forced to include as undesirable dispositions, and hence as character traits, epilepsy, low intelligence, and relatively permanent mental abnormalities. Such a broad conception of character traits might not by itself be problematic. But when it is combined with BT and other theses that Bayles accepts, it has some very counter-intuitive implications. Before looking at these implications, however, we should take note of Bayles’ answer to (2b).

In the context of discussing the category of impossible attempts sometimes referred to as "inherently" impossible attempts (e.g., trying to kill someone by sticking pins in a doll), Bayles says: "Although people who hope to kill others by sticking pins in dolls may not have admirable characters, they are more a nuisance than a harm. For punishment, one must look not simply to whether their intent or purpose indicates an undesirable trait, but also to whether this trait is likely to produce criminal harm. The criminal law should be restricted to those undesirable dispositions which may result in harm." Thus, the kind of character trait (disposition) required for morally justified punishment is one by virtue of which the person is dangerous (likely to cause harm). Of course, a person with a dangerous disposition will also have a socially undesirable one. So, the type of character trait necessary for blameworthiness is also necessary for justified punishment.

Now let us look at some of the implications of Bayles’ broad conception of character traits. According to BT, having an undesirable character trait is both necessary and sufficient for being blameworthy. Thus, we must conclude that a person who has epilepsy, low intelligence or some enduring mental abnormality is for that reason blameworthy. On the
other hand, one who suffers a temporary mental disturbance is not on that account blameworthy, since the temporary nature of the affliction prevents it from being a behavioral disposition. In fact, looking at ET, if criminally prohibited conduct were to result from a temporary mental disturbance, the agent may have an excuse. Despite the prima facie oddness of these implications, Bayles is prepared to accept them.¹⁰

I suspect that one reason he is prepared to accept them is that he does not limit the blame that one must be worthy of, in order to be justifiably punished, to moral blame. Blame, says Bayles, is "a negative attitude towards a person for a character trait."¹¹ On the face of it this description of blame ignores some important distinctions. First, not just any negative attitude constitutes blame, even if the attitude is directed towards another on account of what we all would classify as a character trait. Suppose I just don't like people who are callous. I may nevertheless (perhaps because I am a "hard determinist") not blame them for what they do or for what they are. Not liking and blaming are distinguishable negative attitudes toward persons. Second, not only do we commonly recognize a distinction between blame and other nonblaming negative attitudes, but we typically distinguish moral from nonmoral blame. If the second baseman fumbles the ball, allowing the winning run to score, we may blame him for the loss. But the blame we thereby pin on him would not normally be moral blame. Nor would we typically say that he was "blameworthy" for the loss. We usually reserve "blameworthy" for cases where we think the person in question is worthy of moral blame. Of course there could be situations where the
second baseman is morally blameworthy for fumbling the ball. Suppose, for example, that at the time of the fumble he was trying to impress a fan with some fancy footwork, or he was drunk, or the fumble was attributable to some other moral fault.

When Bayles says, in the context of a theory of excuses, that criminal punishment is morally justified only if the one punished is blameworthy, it would seem that either he does not have moral blameworthiness in mind, or his view of the nature of moral blameworthiness is suspect. There is some reason to think that it is the former. In the context of explaining Hume’s account of blame and excuse, an account with which Bayles has much sympathy, he observes that "[m]ost contemporary philosophers note that blame may be appropriate in nonmoral contexts and then focus on moral blame. While this procedure is perfectly acceptable, one must keep clearly in mind that the account given is only for a subclass of blameworthiness. However, Hume’s view is not so restricted." The same, I think, is true of Bayles’ view. The adoption of a definition of "character trait" that rejects a voluntariness/control requirement and includes conditions such as epilepsy and enduring mental abnormalities as character traits, together with the corresponding definition of blameworthiness articulated in BT, result in a thesis that justified punishment requires what we may call "generic" blameworthiness, rather than moral blameworthiness.

Of course punishment need not be justified in every case where a person evidences generic blameworthiness. This is implicitly recognized in PT₁, which states that blameworthiness is only a necessary condition of justified punishment. Punishment is one of many organized social
responses to the generically blameworthy. There may, in a given case or class of cases, be good policy reasons for employing one of the other social responses available. It may, for example, make more sense to employ institutions better suited for moral education in the case of the youthful offender, and the psychological treatment afforded by a secure mental hospital may be more appropriate for the mentally abnormal offender.  

Should only generic blameworthiness be required for justified punishment? Why not require that the person punished be morally blameworthy? What precisely turns on which type of blameworthiness is required? According to ET, excuses defeat an inference from conduct to character trait, and hence (on Bayles' view), from conduct to generic blameworthiness. But, of course, if excuses negate generic blameworthiness, they also negate moral blameworthiness, since the latter type of blameworthiness is a species of the former. One who (unlike Bayles) requires moral blameworthiness for justified punishment and who (like Bayles) views excuses as defeating an inference from conduct to blameworthiness, will recognize a broader range of excusing conditions than does Bayles. For example, such a theorist may include epilepsy and permanent, severe mental disability as excusing conditions, while Bayles would not do so. No doubt Bayles would not recommend punishment as the appropriate response to the epileptic or the mentally impaired, so one might still wonder if anything of importance hinges upon whether one requires generic or moral blameworthiness. There are at least two, perhaps three, potentially significant consequences.
In order to set the stage for the first consequence, we must look briefly at Bayles' rationale for ET. It would seem to be one derived from the general justifying aims of punishment, which Bayles apparently takes to be the correction and suppression of undesirable character traits. He says, for example, that "[b]lame and punishment are not directly for acts but for character traits." The claim is reiterated in connection with his discussion of attempt liability, where he says: "If punishment is a response to undesirable character traits rather than dangerous acts, then punishment for impossible attempts seems to make more sense." Again, in the context of a discussion of the mentally abnormal offender, he says, "if punishment would not be effective in correcting or eliminating the undesirable character traits, then punishment would not be an appropriate social response. For many persons who commit wrong acts, punishment is an appropriate response because it corrects their character trait or prevents the development or expression of similar character traits of others." On this view, the reason we are not justified in punishing individual offenders who have an excuse is that their excuse blocks an inference from prohibited conduct to an undesirable character trait. Without a good reason to think that the offender has an undesirable character trait, one lacks a good reason to think that punishment will serve the general justifying aims and so punishment is not justified.

If we apply this rationale, along with the broad conception of character traits and the generic blameworthiness requirement endorsed by Bayles, we may conclude that in those cases where a valid excuse is recognized, the State is morally precluded from providing a social response
of any sort that is designed to correct or suppress an undesirable character trait, be it punishment or some other form of treatment. This is because a valid excuse defeats an inference to the existence of any kind of undesirable character trait. Thus, contrary to the views of some scholars, it is morally impermissible for the State to subject anyone with a valid excuse to forced treatment, even if its intent is nonpunitive.  

Excuses remove punishment, along with other forms of coercion, from the menu of morally permissible responses available to the State. But if we apply the above rationale, conception of character traits and generic blameworthiness requirement to such undesirable traits as epilepsy, severe retardation and other enduring mental disabilities, we cannot automatically remove punishment from the menu of potentially permissible State responses to conduct caused by such conditions. Whether punishment is justified depends on whether it will likely be an effective way to alter those traits or suppress their expression. If, for example, a mentally ill defendant's condition can be effectively improved by criminal punishment, punishment may be morally permissible even if at the time of the criminal act the defendant could not have acted otherwise. This appears to be an implication of Bayles' rejection of the Aristotelian voluntariness/control requirement for blameworthiness [VT].

There is a third possible implication of Bayles' adoption of generic, as opposed to moral, blameworthiness as a requirement for justified criminal punishment. Although PT1 and ET require the performance of an act, BT and CT apparently do not. According to Bayles' version of CT,
character traits apparently are behavioral dispositions. On any plausible analysis of dispositions, a person may have a disposition and never manifest it, just as a cube of sugar may be soluble and yet never dissolve because never placed in water. This seems innocuous enough. But the absence of an act requirement in BT would appear to be more troubling. After all, we normally think of persons as being blameworthy for doing or failing to do something. Nevertheless, although Bayles insists that criminal *punishment* would never be justified in the absence of a criminally harmful or threatening act, it is clear that his definition of blameworthiness does not require an overt manifestation. Consider his description of the distinctions among blame, expressions of blame and punishment:

One must distinguish blame, expressions of blame, and punishment. Blame is a negative attitude towards a person *for a character trait*; it is appropriate if the person is blameworthy. Even if a person is blameworthy, one may or may not express blame verbally. Finally, punishment is a particular social response to blameworthy *conduct*—that exhibiting an undesirable character trait.

Later, in response to the very charge that his theory "would support punishment or treatment without the requirement of an overt act," Bayles has occasion to refer back to these distinctions: "This criticism ignores the previous distinction between blame or blameworthiness, expressing blame, and punishment or other social responses. Blameworthiness is only a necessary, not a sufficient, condition for punishment." Implicit in this response is an admission that although justified punishment requires performance of an overt act, blameworthiness *does not*. In thus divorcing blame and blameworthiness from any necessary connection with conduct,
Bayles is (consciously) following Hume. In *A Treatise of Human Nature*, Hume says:

If any action be either virtuous or vicious, 'tis only as a sign of some quality or character. It must depend upon durable principles of the mind, which extend over the whole conduct, and enter into the personal character. . . . Actions are, indeed, better indications of a character than words, or even wishes and sentiments; but 'tis only so far as they are such indications, that they are attended with love or hatred, praise or blame.

On this view, people (or their character traits) are blameworthy in a primary, nonderivative sense; acts are blameworthy only in a secondary, or derivative sense. People are blameworthy for what they are, not for what they do (or do not do). What they do (or do not do) is only relevant as a window to their soul.

Let us consider once again our initial question: Why would one think that criminal punishment is morally justified only when conduct evidences an undesirable character trait? Bayles' answer would appear to be that criminal punishment is morally justified only if the person punished is generically blameworthy and to be generically blameworthy is simply to have an undesirable character trait, i.e., a socially undesirable behavioral disposition. The reason that only those who are generically blameworthy may be justifiably punished is that the general justifying aims of criminal punishment (along with other coercive State responses) are the correction and suppression of socially undesirable (and dangerous) behavioral dispositions. If the State lacks reason to believe that a person has such a disposition, then it lacks reason to punish. Absent a good reason to punish, punishment is morally unjustified.
According to ET, valid excuses defeat a normally warranted inference from conduct to an undesirable character trait. As Bayles correctly points out, the theoretical possibility of other inference paths to character trait does not mean that criminal punishment is justified in the absence of conduct. But even if we assume that the criminal law is properly concerned solely with inferences from conduct to character, two questions remain to be answered if we are to complete our survey of Bayles’ account of excuses: (4) What kind of conduct normally warrants an inference to an undesirable character trait? (5) How precisely do excuses defeat this normally warranted inference?

If we begin by assuming that the kind of character trait that we are interested in identifying is one by virtue of which a person is dangerous (i.e., disposed to cause harm), then it might make sense to suggest that the relevant kind of conduct is behavior that causes or substantially risks harm. A little reflection is all that is required for one to realize that the fact that a person has engaged in a single instance of harmful (or harm threatening) behavior, without more, is not a very reliable basis upon which to infer the presence of a disposition to cause harm. After all, innocent accidents do happen. If we add, however, that the behavior in question was an instance of, for example, purposely or knowingly causing harm to another, then (other things being equal) we can be much more confident of the presence of an undesirable disposition to cause harm. According to Bayles, these psychological elements, indicate attitudes towards harm, and behavior plus attitude constitutes strong evidence of an underlying character trait.
If we follow George Fletcher in distinguishing incrimination from exculpation, and allocate to the statutory definition the elements of incrimination, we can generate a rationale for including Model Penal Code \textit{mens rea} elements in the definitions of criminal offenses. Incriminating conduct, on the Baylesian model, is conduct that sufficiently evidences the kind of undesirable character trait required to warrant criminal liability. Prohibitory norms, the content of criminal statutory definitions, are proscriptions of act-types, tokens of which are incriminating in the above sense. Thus, a virtue of the Baylesian theory is that it offers a promising account of the problematic concept of incrimination that is central to Fletcher’s theory of criminal liability.

Another virtue of Bayles’ character trait model is its ability to account for our intuitive ranking of the various \textit{mens rea} elements. Take the distinctions among purpose, knowledge, recklessness and negligence. Other things being equal, we rank a homicide committed purposely as more serious than a homicide committed with knowledge, the latter as more serious than one committed recklessly, and one committed recklessly as more serious than one committed negligently. Bayles’ explanation of this ranking is that the different types of \textit{mens rea} indicate different attitudes towards harm. A person who harms purposely \textit{wants or desires} the harm to occur. One who harms knowingly \textit{does not care} whether the harm occurs. One who recklessly harms another displays an attitude of \textit{indifference to a substantial and unreasonable risk} of harm. And a person who negligently harms another displays an attitude of \textit{indifference to failing to observe a standard of care}, which, if observed, would result in an awareness of
substantial and unreasonable risks of harm of which the person should be aware. These attitudes towards the occurrence of harm are in a descending order of undesirability. Since behavior plus attitude provides strong evidence of an underlying character trait, we have good reason for concluding that the character traits corresponding to these *mens rea* elements are also ranked in a descending order of undesirability.

We come now to the final question on our original list: How precisely do excuses defeat the normally warranted inference from criminally prohibited ("incriminating") conduct to the existence of a criminal disposition? On this issue Bayles is not as clear as one would like. In fact, he applies his theory to very few bona fide excuses, preferring instead to deal with other issues of criminal responsibility. Three defenses that he does treat in some detail, and which he classifies as excuses, are mistake of fact, ignorance or mistake of law, and "temporary" insanity. He also discusses "permanent" insanity, but concludes that it ought not provide the basis of a valid excuse. From what he says about these cases, one can reasonably extrapolate to such defenses as duress, infancy, and intoxication.

A young woman is strolling through the woods wearing white gloves. A nearby hunter mistakes one of the gloves for the white tail of a deer, shoots and kills the unsuspecting woman. To a charge of murder ("purposely causing the death of another human being"), the defendant may successfully plead the defense of mistake of fact. The mistaken belief is incompatible with the definitionally required *mens rea*. Recalling our earlier defense taxonomy, the mistake constitutes a "denial." Clearly, not
just any factual mistake will constitute a valid defense. Suppose, to take an obvious example, the hunter mistakenly believed the target to be a man. Although Bayles observes that to constitute a valid defense, "a mistake must be about part of the definition of the criminal act," he incorrectly classifies valid mistake of fact defenses as excuses. Still, his explanation of why mistake of fact defenses negate liability is in accord with his overall character trait theory. A mistake that is about part of the definition of the criminal act defeats a warranted inference to an undesirable (dangerous) character trait, while a mistake that is not about part of the definition does not defeat such an inference. Of course, the reason that the inference is defeated in the case of a valid defense is because the relevant mens rea, which we saw earlier is required to warrant an inference from a single instance of harmful behavior to character trait, is shown to be absent.

Bayles correctly points out that a relevant mistake need not be reasonable in order to support a valid defense, unless "the crime be one for which there is a lesser offense of doing the act negligently, then an unreasonable mistake should not prevent conviction for the lesser offense." Again, the explanation of why an unreasonable mistake does not provide a valid defense to a crime of negligence is that such a mistake is compatible with the statutory culpability requirement of negligence. This is what one would expect, if mistake of fact defenses operate as defensive denials. Harm can be caused in many ways. Only if harmful conduct includes MPC mens rea is it "incriminating." Valid mistake of fact defenses, by showing that MPC mens rea is absent, negate the incriminating dimension of crime.
Could a defendant possess a relevant *mens rea* when performing a harmful act and still be entitled to a valid defense? Certainly. Suppose the defendant is a three year old who purposely shoots and kills a sibling. No modern jurisdiction would hold the child liable. How are we to arrive at this result? If we adopt the Baylesian character trait model, we might say that the presence of the relevant *mens rea* does not in this case have its normal criminal significance. Ordinarily, *purposely* causing the death of another human being warrants an inference to a criminal disposition to cause such harm. But in this case we have additional information that defeats this inference. A child of three is simply too malleable to have well-formed character traits or behavioral dispositions. When a certain condition deprives a statutory *mens rea* element of its *normal criminal significance*, rather than being incompatible with the existence of that element, the defense based upon that condition is an excuse rather than a denial.

Consider the defense of duress. Ordinarily, a defendant who starts a fire with the purpose of destroying a building or occupied structure belonging to someone else is guilty of arson. But if a thug issues a credible threat to kill one of your children if you do not burn your neighbor’s house, you have a valid defense to an arson charge. The Baylesian explanation is that the psychological pressure in such a case is so great that "a person of reasonable firmness in [this] situation would have been unable to resist." Given the circumstances, we have reason to believe that the combination of harmful behavior plus *mens rea* lacks its
normal criminal significance of indicating a disposition to set fires to the buildings of others. Thus, duress is a criminal law excuse.

A similar rationale could apply to an intoxication defense. Actually, there are two varieties of intoxication defense. In some jurisdictions intoxication may function as a *denial*, if as a result of intoxication (whether or not self-induced) the defendant lacked the statutory *mens rea* of the charged offense.41 Suppose, for example, the charge is murder and there is evidence that the defendant was so intoxicated as to be hallucinating at the time of the killing. According to the evidence, which we shall assume is credible, the defendant threw a dagger, expecting it to be transformed into a butterfly in flight. Unfortunately, the anticipated transformation failed to take place. Instead, the dagger pierced the heart of its human target, resulting in the death thereof. In such a case the state of intoxication has produced a mental state incompatible with the statutory element of purpose. The character trait rationale for the defense is the same as for any defense that negates the required *mens rea*.

Could such a defendant be guilty of a lesser offense such as manslaughter or negligent homicide?42 That depends. If the intoxication was self-induced and at the time the drugs were ingested the defendant consciously disregarded a substantial risk that doing so would actually imperil the lives of others, then the subsequent intoxicated state would not negate the required *mens rea* for manslaughter.43 Likewise, if taking the drugs posed a substantial risk to the lives of others of which the defendant should have been aware, then a charge of negligent homicide may stand despite the intoxication-produced hallucination. These results are readily
explained by the Baylesian character trait model. The inference from the reckless or negligent act of taking drugs to an underlying undesirable disposition to cause serious harm or death to others is warranted. There is an obvious parallel with unreasonable mistakes of fact here.44

If, however, the intoxication was not self-induced (someone slipped drugs into an unsuspecting person's cup of tea), then whether the resulting intoxication will provide a basis for a valid defense to crimes of recklessness or negligence depends upon the degree and type of mental impairment that results.45 In our example, where the impairment is severe enough to cause the described hallucination, it is likely that a defense will be available. Again, Bayles' character trait model explains this result. Neither the ingesting of drugs nor the conduct performed during a temporary period of severe mental impairment evidences an underlying character defect. This is true even if the defendant has a relevant statutory mens rea during the period of intoxication. Our hallucinating dagger thrower might be described as reckless or negligent, for there is a substantial risk of harm that is apparently being consciously disregarded, or a lack of awareness of a risk of which a reasonable person in the circumstances would be aware. But, given the state of intoxication, any such lack of awareness or conscious disregard fails to have its normal criminal significance of indicating an undesirable disposition. The appeal to intoxication in this case constitutes an excuse.46

Despite the popular slogan, "ignorance of the law is no excuse," the law does recognize ignorance or mistake of law as an excuse in certain
The element of truth in the slogan is that the law *presumes* knowledge of prohibitory norms. The presumption is irrebuttable in the case of crimes that are *malum in se*, such as theft, murder, and assault. But it is rebuttable in the case of crimes that are *malum prohibita*. However, rebutting the presumption requires more than simply showing that one was in fact ignorant of the law's requirements. For such ignorance is compatible with an indifference to whether one's conduct was in violation of the law. On a Baylesian account of excuses this attitude is significant because the various *mens rea* elements are indicative of undesirable criminal dispositions in the case of conduct *malum prohibita* only if accompanied by an attitude of indifference to whether the conduct is prohibited by the criminal law. According to Bayles, what must be shown is that the agent in fact has a *desirable* character trait, namely, that "of being disposed to discover the law and conform to it." In other words, the defendant must demonstrate that a good faith reasonable attempt to discover the requirements of law was made. Only then can we say with any confidence that in light of the ignorance or mistake of law the *mens rea* of the prohibited conduct lacks its normal significance.

Some reference has already been made to Bayles' discussion of excuses based upon alleged mental abnormality. Whether a given abnormality grounds a valid excuse depends, among other things, upon whether it is a temporary or enduring condition. If the condition disappears before the time of trial, it almost always excuses. But if it endures, it does not. Saying that "permanent" insanity is a condition worthy of (generic) blame, and hence does not excuse, does not entail that punishment is an
appropriate social response, however. Whether it is a justifiable response depends on whether it is an effective way to correct or eliminate the abnormality which is an undesirable disposition.53

One interesting feature of Bayles’ theory is that it provides a third alternative to the Hart-Wooton debate over punishment and treatment.54 Bayles describes the debate this way:

The issue has generally been framed as between (1) a thoroughgoing attempt to treat dangerous persons without any consideration of excuses at the conviction stage [Wooton] and (2) a punishment system incorporating the predominant view’s requirement that those who are punished have had the capacity and a fair opportunity to conform to law [Hart]. The Humean [and Baylesian] character view is concerned to respond to undesirable character traits, but it does not eliminate excuses. . . . [D]efenses of lack of mens rea, accident, mistake, abandonment of attempts, and justifications would still have a place in the criminal law, along with others not discussed, such as duress. Nonetheless, some forms of mental abnormality and perhaps other conditions which do not prevent the inference of an undesirable character trait would no longer excuse. The choice of social response would then be determined by its effectiveness in correcting or preventing the undesirable trait as well as other considerations, such as the cure not being worse than the disease.55

An issue that the above passage does not address is how Bayles proposes to deal with appeals to severe mental abnormality as the basis of defensive denials. On the one hand, he says that he recognizes lack of mens rea as a valid defense. On the other hand, enduring forms of severe mental abnormality do not excuse because they do not block an inference to undesirable character trait. There is at least a putative problem here, for it is quite possible for a mental abnormality to negate the statutory mens rea.
Consider the classic case of the insane woman who chokes her husband to death, thinking that she is squeezing a lemon. She clearly does not have the *mens rea* required to sustain a charge of murder. I am sure that Bayles would not want to see her "walk." In fact, the very evidence introduced to negate the *mens rea* will also likely establish that the woman has an undesirable character trait. In that case, some kind of coercive State response may be justified, even if it is not criminal punishment. Bayles' doctrine that lack of *mens rea* is a valid defense (he would call it an "excuse"), must be qualified so as not to constitute a valid defense in cases where the very condition that negates *mens rea* establishes an undesirable character trait. Harmful behavior plus *mens rea* is not the only evidence sufficient to warrant an inference to undesirable character trait.56

This concludes our survey of Bayles' character trait theory of excuses. Enough has been said to lend some *prima facie* plausibility to the view that excuses function to block a normally warranted inference from conduct (including *mens rea*) to an undesirable character trait. One question that remains to be answered is whether, in addition to *generic* blameworthiness, *moral* blameworthiness should be required for morally justified criminal punishment. We examined some possible implications of requiring only generic blameworthiness, but no persuasive reasons have been offered for or against requiring moral blameworthiness as well.

In fact, Bayles does not rule out the possibility that moral blameworthiness is required for justified *punishment*. His point seems to be that we need not assume that it is required in order to understand the logic
of criminal law excuses. Generic blame is only a *necessary* condition of criminal punishment. It would be perfectly consistent with Bayles' account of excuses to maintain that among the additional requirements for justified punishment is the requirement that the person punished be morally blameworthy. Borrowing a suggestion from Joel Feinberg, we might divide the criminal process into two stages. At the first stage we are interested in determining whether a given person is "clutchable" by the State, where clutchability implies that the State may subject the clutched person to some coerced treatment program, if necessary.\textsuperscript{57} We might suggest that generic blameworthiness is not only necessary but sufficient to make someone clutchable, but not sufficient to make them justifiably punishable. For punishment to be warranted, the person must in addition be morally blameworthy.\textsuperscript{58} On this conception, Bayles' theory of excuses is an account of conditions that defeat a normally warranted inference from conduct to clutchability.

In order to be in a better position to determine whether justified punishment requires moral blameworthiness, we need to explore the nature of moral blame and moral blameworthiness. In addition, it will be useful to examine an alternative account of the nature of character traits, with a special focus on moral character traits. Of course, if it should turn out that moral blameworthiness is a necessary condition for justified criminal punishment, one might expect a different answer to the question of what type of undesirable character trait is required for justified punishment.
4.2. Moral Blameworthiness and Motivation

If we are to understand the conditions under which one is worthy of moral blame, it would seem necessary first to explore the nature and function of moral blame. Armed with a better understanding of what moral blame is, and to what end humans blame one another, we may be better placed to decide under what conditions a person is worthy of such. As a starting point it seems reasonable to follow Bayles' suggestion and take moral blame to be a negative attitude towards a person. On a character trait theory this negative attitude is in some way based upon the belief that the object of blame has an undesirable character trait. But as we observed earlier, even if moral blame is some kind of negative attitude towards a person for a character trait, it is not just any negative attitude towards that person. We must, then, find out what distinguishes moral blame from other negative attitudes, including nonmoral (generic) blame.

Let us begin, then, with the distinction between blame (be it moral or generic) and other negative attitudes. It would appear that one who blames another must have some reason for doing so. In this regard, blaming differs from simply not liking someone. I may not be able to say why I do not like someone. And, although I may be challenged to provide a reason, if I am unable to do so, that does not mean that I cannot dislike them. One might object that my dislike is ill-founded ("You should give her a chance; she is really very nice once you get to know her."), but the "logic" of ‘like’ and ‘dislike’ does not preclude disliking for no reason. The situation is quite otherwise with respect to blaming another. If I were to say, "I blame Harold," the obvious response is, "For what?" If I am unable to provide a
fairly specific answer to this question ("He dropped the ball!" or "He lied!"), then it is not merely that my blaming him is perhaps ill-founded, I have not succeeded in blaming him at all. To blame, then, is to condemn someone for some specific reason.

On a character trait theory the blame-supporting reason is that the person blamed has (evidenced or manifested) a "character flaw." But, as our discussion of Bayles' theory revealed, the notion of a character flaw can be given a rather broad interpretation. If one goes so far as to equate a character flaw with any socially undesirable (behavioral) disposition, even epilepsy and severe mental disorders become grounds for (generic) blame. What, then is the difference between moral and generic blame?

In a now classic discussion of moral blameworthiness and obligation, Richard Brandt divides attitudes of moral blame into two categories: those of the offender and those of other persons toward the offender. Among the attitudes of self-directed moral blame, Brandt lists guilt, shame, remorse, disappointment, and anger. Of these the most characteristic is remorse. Attitudes of other-directed moral blame include distrust, lack of confidence, suspicion, repugnance, disgust, and contempt. But most characteristic of one who morally blames another is the feeling of retributive indignation, "indignation containing some desire for its object to suffer."

Brandt's claim that retributive indignation is somehow central to the concept of other-directed moral blame is intriguing. However, although in many cases one who morally blames another may adopt an attitude of
indignation containing some desire for the object of blame to suffer, such a desire does not always accompany moral blame. Parents, for example, often morally blame their children for offenses without desiring that they suffer. Whether retributive indignation is sufficient for moral blame is less clear. But it seems at least conceivable that one experience indignation accompanied by a desire that another suffer without morally blaming that person. My resentment of a sibling's preferential treatment at the hands of my parents may rise to the level of indignation accompanied by a desire that my sibling suffer, even though I do not morally blame my sibling for her good fortune.62

If one looks at the list of other-directed moral blaming attitudes listed by Brandt (distrust, lack of confidence, suspicion, repugnance, disgust, contempt, and retributive indignation), while some of them (lack of confidence, repugnance) may be acceptable responses to the epileptic or the severely retarded person who uncontrollably or inadvertently causes harm, most seem clearly inappropriate (distrust, suspicion, disgust, contempt, retributive indignation). Why is that?

One suggestion, which I think has a great deal of merit, is that such negative attitudes as contempt and retributive indignation are responses to perceived motivational defects. Moreover, not just any motivational defect tends to prompt such a response. Typically, it is those that we would classify as defects of "moral motivation" that are relevant. Someone who lacks sufficient motivation to mow his lawn may evoke irritation or dislike, but rarely contempt, and probably never retributive indignation. The reason contempt and retributive indignation seem to be inappropriate responses is
that laziness is not generally considered to be a defect of moral motivation, i.e., a lack of sufficient motivation to comply with the requirements of morality. Furthermore, if moral blame is a response to perceived defects of moral motivation, it makes sense to suggest that moral blameworthiness may consist in having, or perhaps evidencing or manifesting, such motivational defects.

Even if other-directed moral blame does not always contain a desire that the object of blame suffer, it is true that knowledge that one is the object of moral blame tends to be uncomfortable, even psychologically painful. So, desire or not, suffering (or at least pain) often results from (an awareness of being the object of) moral blame. Whatever the precise biological, evolutionary, or environmental explanation, we value the esteem of others (especially those with whom we have close emotional ties) and dislike, even abhor, their contempt. We have fairly strong motivation to seek the former and avoid the latter. Thus, the practice of morally blaming one another (assuming there is awareness of being blamed) for having (evidencing or manifesting) defective moral motivation may be an effective way to help cure (or moderate) that defect.

At least as significant as the discomfort that results from knowing that one is the object of others' moral blame is the discomfort associated with self-directed blame. As a result of the moral education process we not only acquire motivation to comply with the dictates of a moral code, but we also acquire a disposition to blame ourselves in cases where we have (evidenced or manifested) defects of moral motivation. It seems likely that this
internalized surrogate parent that we call our "conscience" operates to strengthen our moral motivation by reinforcing compliance with the dictates of a moral code.

This way of describing the relationship between a moral code and a person's motivation to comply with that code suggests that they are distinct. Not every moral philosopher conceives of the relationship in this way. For example, Brandt advocates what we might call a "motivational" concept of a moral code. He distinguishes a personal moral code from a social moral code, and identifies five features that are characteristic of a personal moral code:

[F]or an individual to have a personal moral code is for him (1) to have intrinsic motivation for or against his, and to a large extent, others performing certain kinds of action; (2) to experience guilt-feelings [i.e., to morally blame himself] when his action shows a deficiency of such motivation, disapproval [moral blame] when another person's does, and admiration or esteem when another person's actions show a super-abundance of such motivation; (3) to think the forms of behaviour he is motivated to perform or feels guilty about not performing (etc.) are important; (4) to think his attitudes of all the foregoing types are justified; and (5) to have the linguistic capability to give all this verbal expression.63

A distinctive feature of this conception of a personal moral code is its inclusion of motivational and emotional elements (in particular, the dispositions to praise and blame) as constituents of a personal moral code. The inclusion of these elements distinguishes Brandt's conception from what we might call a "pure intellectualist" conception of a personal moral code.
According to the pure intellectualist conception, a personal moral code is a set of beliefs of the form, "x is m," where x denotes an act-type (e.g., lying or breaking a promise) and m is some moral predicate (e.g., "wrong," "right," "obligatory," "permissible"). On such a conception a person may have a moral code and not be motivated to act in accordance with the beliefs that constitute the code or be disposed to feel guilt or remorse for failing to so act. Perhaps some "sociopaths" are such people. They can list act-types recognized by themselves and others to be morally wrong, but they are simply unmoved by that recognition. Whether the suggested variety of sociopath exists or not, however, the simple fact that the intellectualist concept of a personal moral code is compatible with such people having a moral code suggests that an important feature of morality is omitted if one focuses exclusively upon the moral beliefs of individuals. Part of our interest in moral codes derives from a realization that they typically play an important role in guiding the conduct of individuals. This function of personal moral codes cannot be adequately understood without some account of moral motivation. It is a virtue of the motivational conception of a personal moral code that it regards motivation as at least as important as cognition in accounting for the role of moral codes as guides for conduct. But whether one opts for the motivational or intellectualist concept, some account of the connection between a personal moral code and motivation is called for. In what follows we shall, for the most part, adopt a motivational concept of a personal moral code. But it must be emphasized that this is not required by the overall account of blame and blameworthiness to be developed.
If we were to accept the motivational concept of a personal moral code, how would we understand the concept of a social moral code? Again, Brandt has a useful suggestion:

The similarity or overlapping of individual moral codes allows us to define ‘the moral code’ of a given society, by listing all types of behaviour enjoined or prohibited by the moral code of at least one person, and then arranging the items on the list in order according to the frequency with which they appear in the individual moral codes in that society. We can pick out those items which are virtually universal among adults, and entitle them ‘the social moral code’. 64

Obviously, there is no guarantee that a person living in a given society will have a personal moral code identical with that society’s moral code. But the above definition of a social moral code does guarantee that almost every personal code in a given society will have as a "core" component the social moral code of that society. For the purpose of understanding the nature of blame and blameworthiness, it seems reasonable to confine our attention to social moral codes. For one thing, we can assume that the vast majority of individuals in a society adopt the social moral code of that society. The vast majority of those engaged in blaming will, then, be appealing to their social moral code in some fashion when they blame. An investigation interested in the logic of blame and excuse may find some advantage in examining the phenomenon of blaming and excusing "writ large." But for our special purpose of investigating the logic of criminal law defenses, there is an additional reason for focussing upon social moral codes. The criminal law is clearly a social institution with social goals. To the extent that the criminal law draws upon or supplements morality, it must be social morality
that is in question. Personal code idiosyncrasies simply will not, and probably should not, shape the criminal law.

This is not to deny that an individual may have a personal code with "higher (or lower) standards" (strength of intrinsic motivation) than those of the moral code of the society to which the person belongs. Such a person may be quicker (or slower) to blame than are those who share the standards (strength of motivation) of the social moral code. But even the individual with a deviant personal moral code will recognize certain conditions as grounds for a valid excuse. And I would maintain that the same "logic" of excuses applies here as applies when the standard is the social moral code.

It is important to realize that attributing a moral code to a given society does not imply endorsement of that code. Whether a particular social moral code is justified is a tough question of moral theory. But it seems reasonable to assume that for any given society all of the major competing moral theories will justify the adoption of some moral code or other.

With respect to general features, what kind of social moral code would one expect to be justified by the major competing normative moral theories? There are various possibilities. One is that the code advocated would include only an intrinsic motivation toward act-type(s) identified in the "fundamental" normative principles of the theory. This would mean that monistic moral theories would advocate single-rule moral codes. Dualistic moral theories would advocate two-rule social moral codes, and so on. On this suggestion, a Kantian social moral code would include only an intrinsic
motivation to perform acts based on maxims that the agent could will to be universal laws of nature. Act utilitarians would advocate a code that includes only an intrinsic motivation to perform acts of maximizing human welfare (happiness or whatever). A theory that embraces both a utilitarian and a fairness principle would advocate a code with an intrinsic motivation to perform acts of maximizing welfare and an intrinsic motivation to perform acts that treat people fairly.

Although it is possible that the social moral code recommended will be just the mirror image of the fundamental principles of the moral theory in question, I think it is unlikely. Moral theories, because they strive for comprehensiveness and simplicity, tend to be articulated in very abstract principles, the application of which is difficult for even the most capable of humans in the best of circumstances. Social moral codes, by contrast, are designed to help ordinary flesh and blood human beings make day to day moral decisions. They are intended for practical guidance under circumstances that often are not conducive to the kind of deliberation and detachment required for reliable application of fundamental principles. With this in mind, some act utilitarians have recommended a kind of "compartmentalization" of moral thinking. They suggest that implementation of a multi-rule social moral code composed of "rules of thumb," although perhaps not the preferred method for reaching moral conclusions for an archangel, is the best method available for the vast majority of humans in the kinds of circumstances that typically call for moral decision-making. They cite several reasons for this: shortage of time and energy to perform necessary act utilitarian calculations, inability to
command a clear view of an act's relevant consequences, the tendency to succumb to pressures, temptation, or bias (especially in favor of one's own interests), and the general human tendency to rationalize. In view of these impediments to detached, thorough, reliable act utilitarian calculations, these utilitarians recommend implementation of a multi-rule social moral code that (except perhaps in extraordinary circumstances) precludes direct appeal to act utilitarian calculations.

It seems to me that even a Kantian might concede that asking ordinary human beings in the day to day grind to stop and think through an application of Kant's Categorical Imperative is unrealistic. How many could even follow Kant's arguments for why making a lying promise or committing suicide are forbidden by the Imperative? It might be better to follow the example of the act utilitarians here and propose rules of thumb that, if adopted, will generally get the results that would have been derived by a direct application of the Categorical Imperative.66

What I wish to suggest, then, is that virtually every major normative moral theory would support the implementation of a multi-rule social moral code of the general sort described above, complete with intrinsic motivation for or against identified act-types, dispositions to blame, etc. Even monistic theories will find reason to provide such a code as a practical guide to moral decision-making. Of course the specific act-types identified are likely to vary across theories, as would the degree of motivation that might be assigned to those act-types.
It would seem that the rules to be included in any such social moral code should have the status of *prima facie* rules only. This allows rules (and moral motives) to conflict, as they sometimes do, without inconsistency (and incapacitating psychological dissonance). One virtue of the motivational concept of a moral code is that it provides a mechanism (strength of moral motivation) that can accommodate the varying degrees of *prima facie* moral obligation that one might wish to attach to the rules of the moral code. What determines the strength of the *prima facie* obligation to perform act-type \( x \) will, on this conception, be a function of the strength of intrinsic motivation to perform tokens of act-type \( x \). And just as the strength of intrinsic motivation for or against a given act-type may vary from situation to situation, so too will one's *prima facie* obligation to perform tokens of that type. For example, one who is intrinsically motivated to avoid causing injury to others, and so has a *prima facie* obligation to do so, will find that the strength of motivation (and strength of *prima facie* obligation) will vary with the degree and kind of injury in question.

In selecting a social moral code we may decide to encourage a significantly stronger motivation for or against certain act-types than for others. We may even go so far as to use a special moral terminology (that of "moral rights") when classifying the more strongly motivated act-types. We could, if our normative theory warranted it, adjust the levels of motivation so that people in society would not be motivated on balance to perform acts that "violate rights" simply because the act was motivated by an "ordinary" level of moral motivation. (For example, one might be
motivated so as not to violate rights in the name of promoting overall social welfare.) Moral rights would still enjoy only *prima facie* status, so that if rights were to come into conflict, one right could override (i.e., the motivation to avoid infringement would exceed) another right (infringement of which was less motivated).

Again, let me emphasize that one does not have to accept the motivational conception of moral codes, nor must one agree that the best social moral code would be a multiple *prima facie* rule code, in order to accept the view that moral blameworthiness has essentially to do with the level of moral motivation that a given person has (or manifests).

Moral blame, on the account being recommended here, is a negative attitude towards oneself or another, which is believed to be justified on the grounds that the object of blame has (evidenced or manifested) an insufficient level of intrinsic moral motivation for or against an act-type included in the moral code of the society to which the person blaming and the person being blamed belong. Such blame is justified, and the object of blame is blameworthy, just in case the person blamed in fact has (evidenced or manifested) an insufficient level of intrinsic moral motivation for or against an act-type included in the relevant social moral code.67

This concludes (for now) our discussion of moral blame and blameworthiness. Our proposal is that moral blame is a negative attitude towards (manifestations of) substandard moral motivation and that blameworthiness consists in having (or manifesting) substandard moral motivation. But this discussion needs to be set in the context of the larger
task at hand, viz., articulating a plausible character trait theory of criminal law excuses. We saw earlier that Bayles' conception of character traits, along with the companion concept of generic blameworthiness, had some counter-intuitive implications. Moreover, the question of whether justified criminal punishment requires moral as well as generic blameworthiness was left unanswered. This prompted a quest for an analysis of moral blameworthiness. We have now completed that quest. But what of the related concept of character traits? It would be convenient if one could argue that character traits, at least moral character traits, are simply structures of intrinsic motivation for or against act-types included in a social moral code. It is to that question that we now turn.

4.3. Character Traits as Motivational Dispositions

Perhaps it should not come as a surprise to find that Richard Brandt has defended what he calls "a motivational theory" of character traits. According to this view character trait names designate relatively permanent \textit{dispositions} of the want and aversion kind (or some structure of these), which dispositions reach a certain level of \textit{intensity} and which are \textit{intrinsic}. On Brandt's view, the terms "want" and "aversion" are to be understood as theoretical terms that derive their meaning from the psychological theory in which they are embedded. The primary function of wants and aversions is to explain, in the context of a psychological theory, various behavioral, emotional, and cognitive phenomena. Thus, one who attributes a certain want or aversion to a person is to be regarded as offering what is thought to be the "best (causal) explanation" of certain
Brandt contrasts his theory with, and defends it as superior to, two other theories: The Summary Theory (ST, which comes in a "pure" and a "mixed" form) and the Direct Disposition Theory (DDT). Both versions of ST include as part of the analysis of character trait names the condition that a certain corresponding form of behavior/experience has occurred, in the person, frequently or relatively frequently in the past. The main objection to ST is that we often draw inferences about a person's traits of character on the basis of a single piece of behavior. According to DDT, "a trait is not a motive which may issue in a certain type of goal-seeking behavior depending on what other motives are present; it is a disposition for a relatively specific type of behavior to occur." Consequently, "explanation of behavior will take a different form [from that offered by the motivational theory], being more like the kind of explanation solubility in water gives of the soluble thing's actually dissolving in water." One problem with this theory is that no specific "characteristic" type of behavior can be identified. In fact, almost any form of behavior may manifest a character trait, so long as it requires to be explained by reference to the relevant desire/aversion that constitutes the trait in question. Another is that DDT cannot adequately account for how we make, or revise, trait-estimates in view of single pieces of behavior. The point is substantially the same as that made against ST.

In our earlier discussion of Bayles' account of character traits it was observed that Bayles seems to follow Hume in viewing the relation between
character traits and conduct as an *evidentiary* one. Describing the relation this way is compatible with both the Summary Theory and the Direct Disposition Theory of character traits. It is also compatible with an epiphenomenalism that rejects any direct causal relations between character traits and the conduct that evidences them. Brandt’s view of the relation between character traits and conduct is clearly a causal one. At the same time he thinks that conduct caused by character traits does evidence such traits because the causal attribution is based upon a best explanation mode of inference. It would be useful to have a single term that expresses both a causal and evidentiary relation between conduct and character traits. One term that seems ideally suited for this task is "manifests." Let us then say that conduct *manifests* an underlying character trait (want or aversion) just in case the best causal explanation of that conduct includes the existence of the character trait (want or aversion) as a necessary component of a set of conditions that are jointly sufficient in the circumstances to cause the conduct.

On Brandt’s motivational theory traits of character are taken to be a subset of personality traits, the distinguishing features of character traits being two: "(1) Each is a trait of personality which is normally, in *any* adult (not just in a person with a certain role such as a mother or a nurse or a lawyer), either an important asset or an important liability for cooperative living, from the point of view of society. (2) An expression of any of them in action is within voluntary control, in the sense that a person always can produce [such an action] given appropriate interests (wants, aversions) on his part."
Character trait ascriptions "support," in the sense that they may be "partially explained" by, a set of subjunctive conditionals. By way of illustration, Brandt proposes that the expression "X is sympathetic" may be partially explained by the following set of conditionals:

(1) X would feel disturbed, other things being equal, if he perceived some sentient being to be in acute distress.

(2) X would feel relieved, if he perceived a being in distress in process of being helped, provided he had earlier felt discomfort at the person's distress.

(3) X would be motivated to relieve the distress, if he believed that he could do so and that no one else would if he did not.

(4) X would feel guilty, other things being equal, if he perceived distress he thought he could relieve but did not, provided justifying or excusing considerations were absent.

(5) X would notice a case of distress, if he were presented with it perceptually.

(6) X would remember a previous case of distress, if he had noticed it before, and were now in a position to give relief.

One especially nice feature of the motivational theory is its ability to provide an account of the role of character traits in explaining intentional behavior that is compatible with a thoroughgoing deterministic view of human action.

Despite its initial attractiveness, the motivational account of character traits has been criticized on the ground that it oversimplifies the nature of the psychological structures involved. It is said that some evidence of this can be found by examining Brandt's own effort to provide a sample
(partial) analysis of the specific character trait, being sympathetic. For that analysis is at odds with his description of his program as one that identifies character traits with "intrinsic" desires or aversions of a "standard" intensity. If we look at Brandt's analysis of "X is sympathetic," we find that only one of the six conditionals offered, (3), is about X's desires or aversions. The remaining conditionals describe various emotional and cognitive responses to be expected from one who is sympathetic. Thus, the argument continues, Brandt's own efforts to work out the details of his program belie the adequacy of that program.83

In defense of Brandt it should be pointed out that he does not claim to provide an account of character traits that does not mention psychological events, states, or processes other than desires and aversions. His claim is that character traits are to be identified with desire/aversion structures, but since desires and aversions are dispositions, any attempt to provide an account of character traits will include reference to the characteristic manifestations of the relevant desires or aversions. Some of these manifestations will be behavioral, some emotional, and others cognitive. Brandt would insist, however, that character traits are not to be identified with any combination of these characteristic manifestations.84 There is, then, no inconsistency between his general description of the program and his efforts to provide more detailed sample analyses of specific character traits.

Stephen Hudson has argued that Brandt's motivational theory is seriously deficient, since "it omits all reference to the will and its strength or weakness, and . . . one cannot understand character traits without taking this
into account." According to Hudson, Brandt's failure to account for the relationship between character traits and the will results from (or is evidenced by) his neglect of "certain important distinctions which revolve around differences between human excellences and vices."  

First, Brandt allegedly fails to distinguish subscribing to certain values (characteristic of one who has a human excellence) from having specific goals. This leads Brandt, among other things, to construe most character traits as structures of aversions. For if a person with a given character trait (e.g. honesty) is not inclined toward some specific goal, Brandt assumes that the trait must be a structure of aversions, rather than desires. The procedure is, according to Hudson, "ad hoc." Second, Brandt fails to distinguish between (a) actions typical of a trait, (b) those motivated in a way characteristic of a trait, and (c) those that exhibit a trait. The distinctions may be illustrated by reference to the virtue of courage.  

An action is typical of courage if it is one that a courageous person would perform in the circumstances (regardless of the agent's actual motivation). An action is one motivated in a way characteristic of courage if it is typical of courage and performed for the sorts of reasons that a courageous person would perform it. An action exhibits courage if it is typical of courage, it is motivated in a way characteristic of courage, and it proceeds from a firm and (relatively) unchangeable character. According to Hudson, this three-tier model of the relationship between actions and character traits may be used to illuminate important differences between those character traits that are virtues and those that are vices, differences
that Brandt's motivational theory is not equipped to recognize.

On Brandt's view actions *exhibiting* both virtues and vices may admit of multiple motivation (for Brandt courage, the absence of the aversion which is cowardice, admits of multiple motivation; as does dishonesty, the absence of the aversion which is honesty). But Hudson suggests that a proper account of what it is for an action to *exhibit* a character trait will reveal that only vices admit of multiple motivation. First consider virtues. As noted above, in order for an action to exhibit a virtue (e.g. honesty), the action must "proceed from a firm character." Hudson suggests that this be explicated, in the case of virtues, as a requirement that the agent "be principled and . . . have . . . the strength of will to act as he should despite temptations and the like to do otherwise." 89 Next consider vices. Take dishonesty, for example. How is such a trait exhibited in conduct? Conduct exhibits a vice, according to Hudson, if it exhibits a *deficiency* with respect to a corresponding excellence or virtue. Dishonesty is not, therefore, merely exhibited by conduct performed by one who lacks a commitment to the principles of honesty. One may have the commitment, but lack the strength of will to make that commitment "stick" in the face of strong competing desires. Moreover, one may *simply* lack a commitment to the principles of honesty, or one may actually have a commitment to contrary principles.

There are, therefore, at least three distinguishable types of dishonest persons. "Pliable Dodger" has the resolve, but from weakness of will does not always act in accordance with the principles of honesty to which he is "committed." "Machiavellian Dodger" lacks the resolve in the sense that
the principles of honesty carry no weight in his deliberations. "Libertine Dodger" not only lacks a commitment to the principles of honesty, but has a commitment against those principles. The principles carry "negative" weight in his deliberations. These three Dodgers, according to Hudson, illustrate three different ways conduct can exhibit the vice of dishonesty.

Brandt's theory is not equipped, says Hudson, to make these distinctions. According to Brandt, "[t]he reason why behavior manifesting dishonesty is called dishonest behavior is simply that it would not have occurred had the aversion distinctive of the honest man been present."^90 But this ignores dishonesty of the Pliable Dodger variety, viz., that born of weakness of will. And even if we were to modify Brandt's account so as to require that the aversion distinctive of the honest man be present and effective, thereby allowing that dishonesty could result from an ineffective aversion, the resulting account would still not distinguish Machiavellian Dodger from Libertine Dodger. Both fail, but in distinguishable ways, to have the required aversion. Thus, Brandt's attempt to explicate character traits by identifying them with dispositions of the want and aversion kind fails: "[a]n adequate account of character traits must recognize [the] connections between a person's traits of character, his will, and his choices."^91

I think that the basic program of explicating character traits as desire/aversion structures is fully capable of making the needed distinctions, and of accommodating (and explaining) strength and weakness of will. Let us begin with the claim that Brandt fails to distinguish subscribing to values
from having desires for specific goals. The first thing to note is that the term "values" is ambiguous. It may refer to those things that a particular person values, or it may refer to those things that the person ought to value. There is no reason why Brandt cannot equate valuing X with desiring X. If there is a distinction to be drawn here, it is one that can be drawn among desires (or objects of desires). One need not abandon the desire/aversion machinery to explicate the concept of valuing. Hudson assumes that all desires have particular events or states of affairs for objects. From this he apparently reasons as follows. Values (the objects that we value) are not specific events or states of affairs. The conclusion is that valuing X is not a species of desiring X.

There is, however, no reason to grant the first premise of this argument. We tend to reserve the verb "value" for abstract objects of desire (types of events, or types of states of affairs or features of states of affairs) that are "intrinsically desired." An object is intrinsically desired if it is desired for itself, not just because it is a means to something else desired. We value pleasurable sensual experiences, beautiful sunsets, good conversation. Rarely would we say that we "value" having sex with this woman tonight, watching tonight's sunset, or discussing philosophy with Dan over lunch in about an hour, in the sense that we would list any of these among our "values"--even though they instantiate types of events, states of affairs or features of states of affairs that we would list among our values. Nor would most of us list being independently wealthy among our "values." When Brandt gives examples of objects of desires or aversions that constitute character traits, he clearly has in mind abstract objects that are
intrinsically desired, not particular instantiations or abstract objects extrinsically desired.\textsuperscript{94} Brandt can, therefore, readily embrace (and provide an account of) one interpretation of the thesis that for a person to have a character trait is for that person to subscribe to certain values. And he can do so without abandoning his desire/aversion machinery.

The other meaning of "values," viz., "objects that a person ought to value/intrinsically desire," can also be accommodated. Although in the paper in which Brandt first offered his "analysis" of character traits, and to which Hudson is responding, Brandt does not develop an account of what it is for something to be valuable or intrinsically desirable, he does offer such an account elsewhere. We cannot here enter into all of the details, but roughly the account given is that for something to be intrinsically desirable is for it to be \textit{rationally desired}, i.e., to be the object of an intrinsic desire that has been maximally criticized by facts and logic.\textsuperscript{95} There is, then, no reason to think that the motivational theory lacks the apparatus to distinguish, to the extent that it is useful to do so, subscribing to values from having other types of desires (e.g., extrinsic desires, desires for particular objects, or irrational desires).

What of the remaining charges: that Brandt’s theory incorrectly omits reference to the strength or weakness of will, and that the theory neglects important differences between virtues and vices? We begin with the latter question, since allegedly it is Brandt’s failure to notice the important differences that accounts for his failure to provide for the operation of will.
According to Hudson, Brandt is mistaken in thinking that conduct that *exhibits* a virtue, as well as conduct that *exhibits* a vice, admits of multiple motivation. His strategy is to define exhibition of a virtue to require the satisfaction of three conditions: a typical conduct condition, a typical motivation condition, and a "firmness of character" condition—where the latter is analyzed as being "principled" (subscribing to certain values) and having the strength of will to act in accordance with those principles (values). Moral virtue, on this view, is a combination of values and strength of will. Only if one's proper conduct issues from appropriate motives that are in turn produced by a firm commitment to certain values, does the conduct "exhibit" the underlying virtuous character. Rather than defining exhibition of a vice analogously, as the performance of improper conduct that issues from inappropriate motives that are in turn produced by a firm commitment to some set of (vicious) values (Libertine Dodger), Hudson equates exhibition of vice with conduct that results from any deficiency of moral character that prevents one from having a virtue. Each vice, then, is defined as a deficiency with respect to some corresponding virtue. One's character may be deficient with respect to a given virtue in three distinguishable ways: one may subscribe to the values that are (in part) constitutive of the virtue (and that produce a desire to do the virtuous thing) but lack the firmness of resolve to overcome contrary desires/motives (Pliable Dodger); one may *simply* not subscribe to the relevant values (Machiavellian Dodger); or one may subscribe to *contrary* values (Libertine Dodger).
It follows from the above conception of what it is to exhibit a virtue that conduct exhibiting a virtue will admit of only one type of motivation, viz., the desire to do what accords with the values that are (in part) constitutive of the virtue. Conduct that exhibits a vice, on the other hand, may result from any number of types of motives. One may perform a dishonest act because, even though one subscribes to the values of honesty and so desires to do the honest thing, one also desires to be rich, to be liked, not to be embarrassed, etc., and on various occasions these other desires may be stronger than the desire to do the honest thing. Or one may not care one way or the other about being honest, and so be motivated to perform dishonest acts by a wide range of desires. Finally, one may in fact reject the values of honesty and subscribe to those of dishonesty, in which case one will be motivated to do the dishonest thing (at least in part) out of a desire to do that which is dishonest.

The alleged asymmetry between virtue and vice vis-a-vis motivation is, I think, artificially produced by Hudson's stipulations regarding what it is to exhibit virtues and vices. Looking at the three Dodgers, it seems reasonable to distinguish a "strong" and a "weak" sense of exhibition of vice. The conduct of Libertine Dodger exhibits the vice of dishonesty in a strong sense, precisely the same sense in which the conduct of the honest man (Honest Abe) exhibits the virtue of honesty. Libertine Dodger is paradigmatically and reliably dishonest. This is not true of the other two Dodgers. They each exhibit dishonesty only in a weak sense. One cannot really count on them to do the dishonest thing. In fact, Pliable Dodger may well do the honest thing most of the time, since he does value honesty.
Honest Abe and Libertine Dodger are polar opposites on the moral character trait scale. But just as one can identify two intermediate Dodgers who exhibit dishonesty in a weak sense, so we may distinguish two intermediate Abes. Pliable Abe has a commitment to the values of dishonesty, but on occasion does the honest thing through weakness of will. Machiavellian Abe also does the honest thing, but has a commitment to neither the values of dishonesty nor those of honesty. In fact, as far as their character traits are concerned, Machiavellian Abe and Machiavellian Dodger may be identical.

As noted above, Hudson is right to point out that when Honest Abe does the honest thing he will be motivated to do so because it is the honest thing to do. But likewise, when Libertine Dodger does the dishonest thing he will be motivated to do so because it is the dishonest thing to do. Neither case admits of multiple motivation. Hudson is also right to suggest that the conduct of Pliable Dodger and Machiavellian Dodger may admit of multiple motivation. But the same is true of Pliable Abe and Machiavellian Abe. The alleged motivational asymmetry between virtues and vices does not, therefore, obtain. The apparent asymmetry results from a decision to use "exhibit" asymmetrically.

In light of the above, what are we to say of Brandt’s claim that "[t]he reason why behavior manifesting dishonesty is called dishonest behavior is simply that it would not have occurred had the aversion distinctive of the honest man been present"? First, it should be clear that Brandt does not require that the aversion distinctive of honesty always be effective. He
requires only that the aversion have at least the "standard" level of intensity. Thus, one may do the dishonest thing without exhibiting dishonesty in either the strong or weak sense, so long as the intensity of one's aversion to deceit, etc., is of standard intensity. We might call such a person "Excusable Dodger." Second, if manifesting dishonesty is equated with strongly exhibiting dishonesty, Brandt's claim is clearly false. For, as the cases of Pliable Dodger and Machiavellian Dodger demonstrate, one may perform a dishonest act that one would not have performed if one had the aversion distinctive of honesty without that act strongly exhibiting dishonesty.

Third, if manifesting dishonesty is equated with weakly exhibiting dishonesty, then Brandt's claim seems unobjectionable. So long as the aversion distinctive of honesty is present, even if the dishonest thing is done, it will not be the doing of Pliable, Machiavellian, or Libertine Dodger. It might be the doing of Excusable Dodger, but then dishonesty is not being exhibited.

Finally, Hudson's claim that Brandt's theoretical apparatus cannot draw relevant distinctions among the various Dodgers because (allegedly) Brandt has no way to include operation of the will as a feature of character, is simply false. Suppose we identify the will with a person's values (standing intrinsic desires for abstract objects). We may then distinguish Excusable Dodger and Pliable Dodger from the other two Dodgers on the grounds that the former value honesty (i.e., have an intrinsic aversion to deceit, etc.), while the latter do not. Excusable Dodger is distinguished from Pliable Dodger in the obvious way: Excusable Dodger's aversion to
deceit, etc., meets or exceeds the standard level of intensity, while Pliable Dodger’s does not. Neither Machiavellian nor Libertine Dodger have an aversion to deceit, etc. The difference is that Libertine Dodger actually has an intrinsic desire for deceit, etc., that exceeds a certain threshold of intensity. Brandt’s theoretical apparatus seems quite adequate to draw the distinctions noted by Hudson, and then some.

This account of the distinctions makes no appeal to a hierarchical structure of desires and aversions. Recent scholarship addressing the problem of freedom of the will has found it useful to distinguish first-order desires from second-order (and even higher-order) desires. Second-order desires have first-order desires as their objects. On some such accounts, a person’s will is to be identified with second-order desires that certain first-order desires be effective in conduct. The operation of second-order desires may require some level of reflection on, and perhaps rational assessment of, the content of one’s first-order desires. On this view, strength of will is a function of the intensity of one’s second-order desires. Applying this type of account to the case of Pliable Dodger, one might say that Pliable (like nearly everyone else) sometimes has conflicting first-order desires. Among the set of first-order desires is an aversion to deceit, etc. On some occasions (typically cases where first-order desires come into conflict), Pliable is forced to reflect on his first-order desires. Often such reflection activates a second-order desire that his first-order aversion to deceit, etc., be effective. Unfortunately, this second-order desire is not always of sufficient intensity to carry the day. In such cases, if the combined force of his first-order
aversion to deceit and his second-order desire that this aversion be effective in conduct is not of standard intensity, we say that Pliable's conduct is born of "weakness of will."

There is no reason why one attracted to Brandt's program of identifying character traits with desire/aversion structures could not incorporate a hierarchical structure of desires/aversions. Brandt himself sometimes speaks of second-order desires. Moreover, he advocates a process, which he calls "cognitive psychotherapy," as a way of arriving at a maximally rational set of first-order intrinsic desires. It may be that the recommendation that one engage in cognitive psychotherapy has force precisely because humans by and large have a second-order desire to achieve rationality at the level of first-order desires. At any rate, whether one identifies the will with first-order desires and strength of will with intensity of those desires, or one introduces a hierarchical structure to provide a more sophisticated account of the will and its operation, there seems no reason to abandon Brandt's identification of character traits with dispositions of the want and aversion kind.

Two considerations favor a hierarchical concept of the will. Introducing second-order desires allows one to provide an especially interesting account of conscience. Conscience, on this view, may be identified with a standing second-order aversion to failure of a person's first-order intrinsic moral motives to be effective in conduct. Moral blame might be explained as an emotional response to the frustration of this second-order aversion. This way of looking at conscience places the character trait of conscientiousness in a privileged place vis-a-vis other
standing moral motives. The effect of a moral conscience is to provide additional valence to the resultant valence of the various first-order moral motives operating in a given situation, and thereby to increase one's action-tendency to do the right thing.

In addition to providing an account of conscience, a hierarchical concept of will may allow us to provide a more satisfactory account of weakness of will than is possible with a single-tiered concept. Although one can provide an account of the weakness of will involved in Pliable Dodger’s failure to do the honest thing without introducing second-order desires, doing so seems to require that we identify certain first-order desires as morally appropriate and assign an intensity level that is morally acceptable. We judge Pliable Dodger’s failure to do the honest thing despite his being motivated to some extent by an aversion to deceit, etc., as "weakness" of will precisely because we think that Pliable should be more strongly motivated to avoid deceit, etc. Such an account of weakness of will requires an external normative evaluation of the agent’s motivation. This would seem to preclude treating Pliable Abe’s failure to do the dishonest thing, despite his commitment to principles of dishonesty, as a case of weakness of will. But surely there is a sense in which both Pliable Dodger and Pliable Abe suffer from weakness of will.

A hierarchical concept of the will may be able to provide an account of the weakness of Pliable Abe’s will. Suppose we define weakness of will relative to an individual’s own personal moral code. We identify the constituent act-types and intrinsic motivations of this code in large part by
observing when the person morally blames himself for acting or failing to act in certain ways. That is, we identify the constituents of a personal moral code by reference to the operation of a person's conscience. Weakness of will can then be defined as a failure to act in accordance with the dictates of conscience as a result of an insufficient level of first-order intrinsic moral motivation. Thus, by observing when Pliable Abe feels guilty, blames himself, etc., we may discover that he ascribes to the values of dishonesty. But sometimes his motivation to do the dishonest thing is not sufficiently strong to be effective in action. Judging the level of motivation to be of insufficient intensity does seem to require appeal to an external norm, for we must decide how strongly motivated a person committed to the principles of dishonesty should be. But I think that we can make such judgments without imposing our values on the person being judged. One way would be to assign the same level of intensity that our social moral code assigns to contrary principles. Thus, an adequate level of intensity for one committed to principles of dishonesty should match the level assigned by our social moral code to principles of honesty. At least this allows us to provide some content to the belief that both Pliable Dodger and Pliable Abe suffer from a similar weakness of will despite their not sharing the same values.

Introducing the operation of second-order moral motivation further complicates the task of making reliable estimates of the state of a given person's moral motivation. Not only do we have first-order moral motivations competing with first-order non-moral motivations (e.g., honesty versus greed), and first-order moral motivations competing with other
contrary first-order moral motivations (e.g., honesty versus compassion), but an additional force is now operating to strengthen the resultant first-order moral motivation, whatever that may be. Perhaps it is our recognition of the difficulty of making reliable estimates of moral motivation that helps account for the apparent common sense restriction of moral blame to instances of the overt manifestation of motivational faults. This may be attributable to a realization that conduct is usually the best evidence available of the state of a person's moral motivation. Even the person whose moral motivation is in question cannot always make reliable judgments in this area. One may not know just how strong one's moral motivation is until after the deed is done. This is especially true in light of the fact that a strong conscience may compensate for many weaknesses at the level of first-order moral motivation. Some may be less strongly inclined to help others in distress than others, but because of a particularly strong aversion to failing to conform to the demands of duty, help others in distress nonetheless. Perhaps, as good Kantians, they perform many such tasks because they think that duty requires it. Given the difficulty of making reliable inferences to overall strength of moral motivation in the absence of conduct that manifests that motivation, the restriction of blame to instances where substandard motivation is manifested in conduct seems a wise one.
4.4. Moral Blameworthiness and Voluntariness

One of the issues that appears to separate Bayles from Brandt is the necessity of requiring voluntariness for blameworthiness. This difference may be more apparent than real, if, as was suggested, Bayles is employing a nonmoral (generic) sense of "blameworthy." Be that as it may, it is worth noting that if one accepts the account of moral blameworthiness presented above, there is a straightforward sense in which moral blameworthiness requires voluntariness.

One who performs a harmful act, for example, is morally blameworthy only if performance of the harmful act can be attributed to a substandard level of intrinsic moral motivation. But if we follow the earlier suggestion and identify the will with a person’s intrinsic motivation for or against abstract objects (e.g., act-types), then one is morally blameworthy for the harmful act only if the act is attributable to a "defect" of will. Thus, operation of the will is required for moral blameworthiness. We could summarize this by saying that moral blameworthiness requires voluntariness. This voluntariness requirement may not satisfy some free will advocates, but it does distinguish moral blameworthiness from the concept of generic blameworthiness attributed to Bayles. And it would seem to preclude moral blame for harm caused by epileptic seizures and severe mental disorders, as well as that caused by stupidity, fear, and other states or conditions that do not involve substandard moral motivation.
4.5 Moral Blameworthiness and Criminal Punishment

We come at long last to the question of whether justified criminal punishment requires moral blameworthiness. In order to address this question, we must say something about the general justifying aims of punishment. Criminal punishment is probably the most severe form of social coercion brought to bear on individual members of society. Surely its use places upon those who would wield it a heavy burden of justification. Unless it is a necessary and reasonably effective way of addressing some serious social evil, it ought not to be employed. What evil, then, is contemplated?

If ours were a society composed entirely of persons with sufficiently high levels of moral motivation, criminal punishment would be unnecessary. We could count on intrinsic motivations against acts of murder, assault, theft, etc., to prevent such conduct from occurring. But of course ours is not such a society. Social conditions and the variation in techniques of moral education being what they are, we find a wide range of moral motivation toward such harmful acts. Some social misfits seem almost entirely to lack moral motivation to avoid harming others.

Just as important as these obstacles to effective moral education is the realization that it may not even be socially desirable to attempt to instill the level of moral motivation required for a crime-free society. Achieving and maintaining such a high level of motivation has its own social costs. The process of moral education becomes more laborious and time-consuming. The greater the degree of motivation to avoid certain conduct, the greater the psychological discomfort or pain associated with failures. In addition,
there is a danger that the increased aversion to a given act-type may be over-generalized and decrease the tendency to perform conduct that ought not to be discouraged (e.g., one's aversion to causing harm becomes so immobilizing that one refuses to drive a car or engage in many useful activities that pose even the most remote risk of harm). There is also a related problem of over "entrenchment"—where the motivation to perform certain act-types becomes so strong as not to yield to contrary motivations. In such cases the sense of obligation to avoid the act-types in question becomes virtually absolute. This can lead to undesirable results (consider the person so averse to lying as to inform the inquiring Nazi of the location of Jews-in-hiding). It can also lead to an incapacitating psychological dissonance if two such firmly entrenched dispositions were ever to come into conflict (consider the person with equally high motivations to avoid lying and to avoid harm to innocents, confronted with the inquiring Nazi). Given such considerations, it seems likely that a socially optimum level of moral motivation will fall considerably short of the maximum level of motivation that is theoretically attainable.

These observations support the conclusion that the level of moral motivation reasonably attainable for the vast majority of persons in our society (or any real society) will not (and perhaps should not) guarantee a crime-free society. If ordinary reasonably attainable moral motivation is insufficient to keep the incidence of seriously harmful conduct within acceptable limits, it may be necessary to resort to an institution of criminal punishment as a way of supplementing moral motivation. If the social costs
of significantly increasing the level of moral motivation across the board are sufficiently high (or the prospects of success sufficiently dim), it may actually be less costly to implement such a supplemental motivational system. This is because, if properly constructed, a system that strives to deter harmful conduct via threats of punishment need not be a significant intrusion into the lives of the majority who already possess a level of moral motivation generally sufficient to prevent them from engaging in criminally proscribed types of conduct. In that event, only those with a substandard level of moral motivation will experience the additional coercive influence of threats of criminal punishment should they exemplify a criminally forbidden act-type.

The requirement that the system of criminal punishment be properly constructed translates into a requirement that the system punish only those for whom there is sufficient evidence of a relevant motivational defect, and this in turn requires the recognition of criminal law excuses in those cases where the performance of a criminally harmful act does not warrant an inference to a substandard level of moral motivation. A system that did not recognize excuses in such cases would require a substantial sacrifice of liberty and security interests. As Brandt puts it: "Such a system could well be a nightmare, intolerable from the point of view of the average [adequately motivated] law-abiding citizen. What would life be like if one must anticipate a year in prison for accidentally running down a child, with no fault whatsoever on one's own part?" 103

Not only would such a "strict liability" system be oppressive, but it offers little, if any, additional deterrent impact over a system that recognizes
excuses:

Let us ask: what class of crimes would be deterred by a strict liability system but not by a system with excuses . . . . Not those committed by persons who do not know the law, so we must limit ourselves to persons relatively informed on the system of criminal justice. Suppose we think of an informed rational person, who for some reason wants his wife out of the way and is deliberating whether to make an attempt on her life. How will the fact of the system of excuses affect his thinking? Perhaps he can manage to have a bullet ricochet? Perhaps he can convince a jury that he mistook his wife for a burglar, or that he was acting under hypnotic suggestion, or that someone threatened to kill him if he didn’t, or that the episode was the result of a delusion, of schizophrenia? I think the rational person will do better to spend his time thinking how to commit the crime so that the jury will not be convinced that he actually did it. So how much deterrence will be lost, for rational informed persons, from knowledge of the system of excuses? The rational person will see that it is going to be very difficult for him to escape punishment by the excuse route. I think we may agree that people may be encouraged to commit crimes by knowledge that persons who commit crimes are mostly not punished. Perhaps they know that most murderers are never brought to justice. How many of these escape via the excuse route? Perhaps two percent? Will this two percent have a detectible effect on deterrence, given the general situation? Suppose we think, as may be true, that the deterrent effect of the criminal law comes through vicariously attaching negative affect, by conditioning, to the thought of a given offense. Is there any reason to think this conditioning process will be significantly affected by the knowledge that a very small proportion of persons escape punishment by the excuse route?

We might add that in order to insure a minimum impact on deterrence, it would be wise to make known the grounds of acquittal in cases where non-conviction rests upon a valid excuse. This could be accomplished by the use of special verdicts.
In addition to supplementing moral motivation to avoid identified types of conduct, criminal punishment actually reinforces that motivation. The imposition of punishment serves as a social vehicle for expressing moral condemnation (blame) of those who would voluntarily perform proscribed acts. As an emphatic denunciation of the underlying lack of moral motivation revealed by the act, it serves to reinforce whatever conscience may be at work in the heart of the criminal. It also provides reinforcement of the moral motivation of the typical law-abiding citizen.

This condemnatory function of criminal punishment is warranted only in cases where the violation of law was the result of a relevant motivational defect. Absent such a defect, a person is not morally blameworthy. Hence, moral condemnation is unjustified. Thus, the condemnatory function of criminal punishment also supports a system that recognizes excuses in cases where the presence of certain conditions (e.g., duress, involuntary intoxication, severe cognitive disability, etc.) defeat a normally warranted inference from criminally prohibited conduct to a substandard level of moral motivation.

So far our discussion has assumed that the availability of a valid excuse turns upon the level of intrinsic moral motivation of the person accused of an offense. The previous argument has been that the institution of criminal punishment is justified as a necessary supplement to (and reinforcement of) a system of moral motivation we have called a "social moral code." From this one might be tempted to conclude that a legally valid excuse must defeat an inference to substandard moral motivation.
This is clearly the way a properly constructed criminal justice system should operate. But in fact there is no guarantee that every criminal law will operate in the way it is supposed to operate. Morally bad criminal laws are not unheard of. What constitutes a valid criminal law excuse in a case where the motivation required by the criminal law conflicts with that embedded in the moral code? Suppose, for example, that a statute makes it a criminal offense for blacks to drink at a "whites only" water fountain?

Here it seems to me that in order to avoid confusion we ought to say that the relevant motivation, for purposes of determining whether a valid criminal law excuse is available, is that implied in the criminal statute that provides the basis of the criminal charge. It is important to keep in mind that excuses are a particular kind of response to a criminal charge with their own distinctive logic. One who raises an excuse is saying that some condition or other defeats a normally warranted inference from conduct to substandard motivation. A black who violates the statute because of duress, involuntary intoxication, or severe mental disease, has a valid criminal law excuse. One who deliberately violates the statute as a matter of principle does not. In saying this I do not mean to deny the iniquity of such a statute, nor to detract from the moral courage of the black who deliberately violates it. The violation may be morally justified. But pointing out that a statute is evil and is not doing the job criminal statutes ought to be doing is a different kind of claim from that of saying that one’s violation of the statute should be excused. Excuses are not designed to remedy all the ways a criminal code or its application may go wrong. They are designed to guard against just one: the unnecessary imposition of punishment (and blame) on those whose
conduct, despite violating the express provisions of a statute, does not
evidence a substandard level of motivation to conform to the requirements
of the statute.

The answer, then, to whether morally justified punishment requires
moral blameworthiness is that it does. Moreover, excuses, if part of a
properly constructed criminal justice system, will defeat an inference from
statutorily proscribed conduct to substandard intrinsic moral motivation, and
hence to moral blameworthiness. But if the statutes themselves are morally
corrupt, there is no guarantee that excuses will serve the purpose for which
they were designed.
ENDNOTES

1 P. Robinson, Analysis 209, 220.

2H.L.A. Hart, Punishment and Responsibility 219 (1968). So understood, it is a theory of mens rea in the broad sense discussed by Hart at Id.

3Bayles’ theory of responsibility is presented in M. Bayles, Character, Purpose, and Criminal Responsibility, 1 Law and Philosophy 5 (1982) [hereinafter cited as Character].

4Bayles, Character 8.

5Bayles, Character 7.

6In the context of discussing the mistake of law defense, Bayles says of a minister who relied upon the incorrect advice of an attorney general’s office that using a sign to aid in soliciting couples for marriage was lawful: "In this case, as in many others, both the predominant view and [my view] imply that the unfortunate minister should have been excused. On the predominant view, he lacked a fair opportunity to discover the law. On [my] view, his conduct exhibited a desirable character trait, namely, of being disposed to discover the law and conform to it." Character 15-16.

7See Character 19, where Bayles says (with apparent approval) that the criminal law is "not ordinarily concerned with motives."

8Character 7 (emphasis added).

9Character 13. For a similar statement in the context of his claim that the criminal law (rightly) is not ordinarily concerned with motives, see Character 19.
See, for example, his discussion of mental abnormality at Character 17-19.

Character 7-8.


Even if we restrict the type of character traits required for morally justified punishment to dangerous traits, in the manner suggested above, the type of blame of which the epileptic or mentally impaired is worthy is not moral blame.

It is also worth mentioning at this point that Bayles’ rejection of a voluntariness requirement for generic blameworthiness does not commit him to rejecting voluntariness as a requirement for morally justified punishment. As the next paragraph in the text suggests, it is perfectly consistent to maintain that excuses, which block an inference from conduct to generic blameworthiness, do not require voluntariness, while morally justified punishment does. If it should turn out, for example, that only those undesirable dispositions with manifestations subject to voluntary control can be effectively altered by criminal punishment (or the threat thereof), then it may make sense to limit punishment (and the threats) to such dispositions.

Character 18-19.

Actually, as we shall see infra at 171-72, Bayles may not reject moral blameworthiness as a requirement for justified punishment.

Character 7.

Character 12 (emphasis added).

Character 18 (emphasis added).

The underlying assumption is that punishment is *prima facie* undesirable and requires justification.

See discussion of Paul Robinson’s views on this *supra* at 116-17.
See the discussion of moral blameworthiness and voluntariness infra 204.

Actually, he does not categorically rule out punishment in cases where there is no overt manifestation of an underlying undesirable character trait. He says that granting the State the power to punish for mere undesirable character traits "unreasonably risks freedom for security." Character 19. This leaves open the possibility that if scientific technology were to develop to the point where we could reliably detect strong dispositions to commit criminal offenses independent of overt manifestations, coercive intervention, perhaps even punishment if that were an effective way to alter such dispositions, might be justified.

Character 7-8 (emphasis added).

Character 19.

In saying that for Bayles blameworthiness is not necessarily connected with conduct, I do not mean to imply that he thinks the concept of blameworthiness can be analyzed without mentioning conduct. According to Bayles, one is blameworthy for having undesirable character traits, which are undesirable dispositions; presumably, dispositions to act in certain ways in specified conditions. The point is rather that one is blameworthy for having such dispositions, whether or not one ever in fact has occasion to manifest them in conduct. Moreover, if there were some way of identifying dispositions other than through their conduct manifestations (e.g., through analysis of brain waves or whatever), one would be justified in blaming a person for having undesirable dispositions so identified. Overt manifestation of the disposition is not required.


Whether the appropriate way to conceive of the relationship between conduct and blameworthy character traits is evidentiary, as Bayles and Hume suggest, or a more "intimate" connection ("manifestation" or "exhibition") is required, is an issue that merits further discussion. See infra 187.

As noted earlier, Bayles thinks that there are important policy reasons for an act requirement. Supra 160, note 22.
29. Bayles does not make the suggested inference path from MPC mens rea to character trait entirely clear. One possibility is that the inference goes from mens rea to attitude, and then from attitude plus harmful behavior to character trait. Another possibility is that the inference goes from mens rea plus harmful behavior to attitude, and then from attitude to character trait.


31. See infra 237.

32. Bayles expresses reservations about the propriety of grounding criminal liability on negligence. Not only is the character trait in question much less undesirable than those normally indicated by harmful behavior that is purposeful, knowing, or reckless, but the inference from negligent harmful behavior to character trait seems "less valid." Character 10-11.


35. Character 15.

36. We may also doubt that the child has the conceptual apparatus to fully appreciate the nature of the harmful act. Does the child understand the concept of death? After all, cartoon characters consistently get shot or clobbered by huge rocks, shake themselves off, and continue their crazy antics. Why should a child think it any different for humans? Thus, another way of looking at the child’s act is to say that even if the child did "purposely" kill another human being, the mental element involved was not embedded in the kind of conceptual context necessary for it to have its normal criminal significance of indicating a disposition to cause death to humans.

Another possibility is that Bayles may simply decline to treat infancy as an excuse, preferring instead to treat the child defendant in a manner similar to that recommended for an adult who is "permanently" insane. The very malleability alluded to may move us to look for less drastic responses than criminal punishment. It seems unlikely that a child of three will comprehend the purpose of criminal punishment. Such harsh treatment is more likely to harden than to mold the child’s character in beneficial ways.
Thus, considerations of "public policy" may dictate that alternative forms of education or treatment be employed to correct the undesirable character trait, if we decide that the child has such a trait.

\[37\] Actually, Bayles does not describe the operation of excuses in quite this way. He apparently would include conditions that negate statutory \textit{mens rea} as "excuses" as well. So what I am suggesting here is a slight modification of the Baylesian theory to reconcile it with the earlier taxonomy of defenses sketched in Chapter I.

\[38\] "It is an affirmative defense that the actor engaged in the conduct charged to constitute an offense because he was coerced to do so by the use of, or a threat to use, unlawful force against his person or the person of another, that a person of reasonable firmness in this situation would have been unable to resist." MPC § 2.09(1).

\[39\] MPC § 220.1(1).

\[40\] MPC § 2.09(1).

\[41\] MPC § 2.08.

\[42\] "Criminal homicide constitutes manslaughter when: (a) it is committed recklessly . . ." MPC § 210.3. "Criminal homicide constitutes negligent homicide when it is committed negligently." MPC § 210.4.

\[43\] Suppose, for example, the defendant had a history of violent attacks while intoxicated. The Model Penal Code excludes evidence that a defendant was unaware, during the period of intoxication, of a substantial risk of harm as a result of the intoxication (offered to negate the element of recklessness), if the risk was one "of which [the defendant] would have been aware had he been sober." MPC § 2.08(2). Apparently, this represents a judgment on the part of the code drafters that \emph{anytime} one voluntarily becomes intoxicated, one is consciously disregarding a substantial risk of harming others. This seems untenable. Absent some past history of engaging in threatening behavior while intoxicated, there is no reason to suspect that I am endangering others by becoming intoxicated, let alone reason to think that at the time I choose to imbibe I must be consciously disregarding a substantial risk of harming others. Furthermore, is not the type of harm anticipated relevant here? I may realize that I risk saying something offensive to someone while intoxicated, but not think that I risk causing their death. If I do happen to cause their death because I fail
to comprehend a risk that I would have comprehended had I been sober, should I be guilty of manslaughter? A character trait model of the type defended by Bayles would suggest not.

Query whether a defendant could be guilty of murder if there was a conscious decision to ingest drugs in order to "steel one's nerves" with the hope and intention of taking the life of another during the intoxicated state. The intoxication results and the rest of the story is as above. At the time of letting the dagger fly the defendant did not intend that its target be killed. It seems to me that the character trait model would permit a conviction for murder in this case. The revealing act occurs at the time of ingestion, when the required mens rea is present.

The Model Penal Code test is whether "by reason of such intoxication the actor at the time of his conduct lacks substantial capacity either to appreciate its criminality [wrongfulness] or to conform his conduct to the requirements of law." MPC § 2.08(4).

Similar comments are in order even if the mens rea in question is MPC knowledge or purpose. Of course the degree of impairment via intoxication has to be quite severe, comparable in effect to temporary insanity, if it is to excuse. See MPC § 2.08(4).

As with intoxication, ignorance or mistake of law may also function as a denial if the ignorance or mistake is incompatible with the statutory mens rea of the charged offense. Statutes defining the crime of income tax evasion frequently require that the defendant know that the conduct violates the law. See e.g. United States v. Murdock, 290 U.S. 389 (1933); Hargrove v. United States, 67 F.2d 820 (5th Cir. 1933) (crime of knowingly attempting to evade payment of income taxes required proof that defendant was aware of duty to pay taxes).

Bayles does not, in his discussion of the mistake of law excuse, distinguish crimes malum in se from those malum prohibita.

What constitutes a sufficiently reasonable attempt is a matter of some debate. The relevant Model Penal Code provision is MPC § 2.04(3).
A number of spurious arguments have been offered for why the law should not recognize ignorance or mistake of law as an excuse. It has been said that such an excuse would encourage ignorance of the law. People v. O’Brien, 96 Cal. 171, 31 P. 45 (1892). Another argument is that it would become a shield for the guilty because of the difficulty of determining (a) whether the defendant was ignorant of the law, and (b) assuming the defendant was ignorant, whether the ignorance was faultless. Id. A corollary of this fear is that the defense will encourage collusion between attorney and client in an attempt to excuse the client’s wrongdoing. Hunter v. State, 158 Tenn. 63, 61, 12 S.W.2d 361, 363 (1928). Public interest in promoting the educational function of the criminal law is said to require that individuals not be acquitted on the grounds of unawareness of the law. Such acquittals, it is said, would contribute to public uncertainty and confusion as to what conduct has been made criminal. G. Williams, Criminal Law: The General Part 289 (2d ed. 1961). Finally, it is argued that allowing such an excuse conflicts with the basic principle of legality upon which our legal system is based: "[T]here is a basic incompatibility between asserting that the law is what certain officials declare it to be after a prescribed analysis, and asserting, also, that those officials must declare it to be . . . what defendants or their lawyers believed it to be. A legal order implies the rejection of such contradiction. It opposes objectivity to subjectivity, judicial process to individual opinion, official to lay, and authoritative to nonauthoritative declarations of what the law is." J. Hall, General Principles of Criminal Law 383 (2d ed. 1960).

Despite such objections, the American Law Institute has recommended two limited exceptions to the general prohibition of the use of mistake of law as an excuse:

A belief that conduct does not legally constitute an offense is a defense to a prosecution for that offense based upon such conduct, when: (a) the statute or other enactment defining the offense is not known to the actor and has not been published or otherwise reasonably made available prior to the conduct alleged; or (b) he acts in reasonable reliance upon an official statement of the law, afterward determined to be invalid or erroneous, contained in (i) a statute or other enactment; (ii) a judicial decision, opinion or judgment; (iii) an administrative order or grant of permission; or (iv) an official interpretation of the public officer or body charged by law with responsibility for the interpretation, administration or enforcement of the law defining the offense. MPC § 2.04(3).
We may call these the "no publication" and "reasonable reliance" excuses. Let us examine to what extent this proposal avoids the above objections.

As for the first objection, ignorance of the law will not be encouraged if the defense is allowed only in cases where (1) the law in question is not known and has not been made reasonably available, or (2) there was reasonable reliance upon a statement of the law issued by one charged by law with the task of making, interpreting, or enforcing the law. In neither case could one have eliminated the ignorance or avoided a mistake by employing any means reasonably available.

In response to the first part of the second objection, it may be questioned whether proving the state of mind (or lack thereof) required for this defense is any more difficult than other issues that are commonly raised in a criminal trial. Williams 291. The problems of proof associated with demonstrating that the defendant's ignorance was faultless are mitigated somewhat by restricting the defense to cases of no publication and reasonable reliance upon official statements of the law. The obtainment vel non of these circumstances should present no special problem for adjudication. See commentary to MPC § 2.04(3). Moreover, since the defense proposed is an affirmative defense, with the defendant bearing the burden of proof, difficulty of proof can hardly be a good reason for taking away the defense.

The problem of collusion is virtually eliminated by restricting the reliance defense to official statements of law, thereby excluding reliance upon a private attorney's advice from the scope of the defense. Id.

It hardly seems to advance the educative function of the criminal law not to allow citizens to rely upon official statements of the law. In fact, failure to allow those who seek out an official statement of the law, to rely upon that statement, would seem to penalize those who take the educative function most seriously.

Finally, one of the primary functions of the principle of legality is the prevention of unnecessary governmental interference in the private lives of individuals, by making the law available so that individuals may plan their lives in a manner allowing them to remain within the law. Failure to allow reliance upon official statements of the law as a defense introduces an element of arbitrariness and unpredictability that the principle of legality is designed to eliminate.
Bayles’ own example (borrowed from Richard Brandt) is of a psychopath who is fascinated by the gurgling sounds people make as they take their last breaths while being strangled to death. Character 17. The question raised in the text can be avoided in such a case because it may be that the psychopath intends to kill the victims, so that the mens rea for murder is present. But suppose instead that the psychopathic killer is simply fascinated by the gurgling sounds and does not even realize that people are being killed in order to satisfy a rather bizarre urge. No doubt Bayles would insist that the state is justified in apprehending and subjecting such a person to coercive treatment of some sort.

J. Feinberg, Deserving 265-71.

I am not suggesting that moral blameworthiness is sufficient to make someone justifiably punishable.

Whether moral blameworthiness requires an overt manifestation of the character trait (through conduct, for example) is a question yet to be addressed. In many of the examples that follow, I shall assume that there has been an overt manifestation, and shall sometimes speak as if the blame is for that manifestation. On a character trait theory of the type defended by Bayles, this language will have to be interpreted as elliptical for a statement about a blaming attitude towards the blamed person on account of a character trait that the overt manifestation reveals.

R. Brandt, Blameworthiness and Obligation, in A. I. Meldon (ed.), Essays in Moral Philosophy 3, 25 (1958) [hereinafter cited as Blameworthiness].


Of course I may blame my parents, but that is a different matter.

R. Brandt, A Theory of the Good and the Right 169-70 (1979) [hereinafter cited as Good and Right].
Some who recognize that they have "higher standards" may not even expect others to live up to the standards of their own personal code. If so, then other-directed blame may not occur in situations where manifestation of the same level of moral motivation would evoke self-directed blame.

Kant would probably shudder at the thought, for such a suggestion seems antithetical to the ideal of humans as autonomous self-legislators. But it is, I think, an ideal that we cannot realistically expect the vast majority of human beings to live up to.

There may be a question here of to what extent the person being blamed must share the social moral code (or even be a member of the same society) as the person blaming. Obviously, simply having a substandard level of moral motivation does not preclude blame, or no one would be blameworthy. This may be a matter of degree. A complete alien whose "moral code" is radically different from ours hardly seems an appropriate object of blame, but what of the person who shares all of the features of our social moral code, save one: the person is not averse to lying or deceiving people? I do not have an answer to this troublesome question.


The intensity required is "at least the standard level," where the standard level is to be determined by what the typical person in a language community would regard as a minimally sufficient level to warrant ascribing the relevant trait to a person. *Traits* 36-37.

To say that a disposition is intrinsic is to say that its characteristic cognitive, emotional, and behavioral manifestations are not wholly derivative from some other disposition. *Traits* 30.

Actually, Brandt prefers the psychological term "valence" over "want" and "aversion." An event or state of affairs that is wanted is "positively valenced," while an event or state of affairs to which one is averse is "negatively valenced." Good and Right 24.

*Traits* 26.
73*Traits* 31.

74*Traits* 31.

75See Brandt's discussion of the trait, generosity, at *Traits* 33.

76*Traits* 34.

77*Supra* 161.

78*Traits* 24.

79*Traits* 25, 29.

80*Traits* 29-30.

81*Traits* 27. Thus, the motivational theory fits comfortably within the account of human action sketched in Chapter II.

82S. Hudson, *Character Traits and Desires*, 90 Ethics 539 (1980).

83Hudson, *Character Traits and Desires* 542.

84One might wonder if this is true of (3) in the proposed analysis of "X is sympathetic." After all, (3) is explicitly about X’s motivation. But the motivation in question appears to be what Brandt elsewhere calls "action-tendency." Action-tendency is a function of "expectancy" and "valence," the latter being roughly equivalent to "desire." Brandt, Good and Right 25-26, 48. An increase in action-tendency is only one characteristic result of an increase in valence (holding everything else, e.g., expectancy, constant). Thus, (3) does not differ fundamentally from the other components in Brandt’s "analysis." (3) describes only one among many characteristic manifestations of a certain valence (desire) structure with at least a "standard" level of intensity.

85*Character Traits and Desires* 540.

86*Character Traits and Desires* 543.

87*Character Traits and Desires* 543.

88*Character Traits and Desires* 544-45.
Character Traits and Desires 547.

Character Traits and Desires 549 (quoting Brandt, Traits 33) (emphasis added).

Character Traits and Desires 549.

It is not clear whether he is attributing this view to Brandt, or whether this is his own view on desires. See Character Traits and Desires 543.

More formally, "an outcome O is occurrently intrinsically valenced [desired] for a person at a time t if and only if his central nervous system at t is such that if he judged (thought with belief) that a certain act by him at t would tend to bring about O, then, even in the absence of any further judgements about O (such as its probable further effects) not contained in the concept of O, there would be an increase in his tendency to perform that act." Brandt, Good and Right 32.

In Traits Brandt says: "[t]he wants/aversions which are identical with traits of character must be 'intrinsic.' What [is] meant by this is that the truth of the 'if . . . then . . . ' statements which constitute the explanation of the trait-name is not derivative, or at least not wholly derivative, from some other want/aversion." Traits 30. Moreover, the examples he discusses bear witness to both the abstract nature of the objects and the intrinsic nature of the desires/aversions that are character traits. Considerateness is a "concern for discomforts, embarrassments, etc." Honesty is "aversion to appropriation of the property of others, deceit, etc." Traits 32-33. Conscientiousness is "aversion to failure to do one's duty." Unselfishness is "a relatively high interest in the welfare of other persons." Kindness is "aversion to causing any kind of distress in others. Truthfulness is "aversion to deviation from the truth, or to the kind of interpersonal relationship which results when one party indulges in deception." Traits 34. Finally, of courage Brandt says: "There is something in Aristotle's suggestion that courage requires a desire to maintain one's honor, or an aversion to disgrace; the point is that there are some values which have priority over personal safety in a courageous man, whereas a cowardly man will back down on any issue when a serious personal threat arises." Traits 35 (emphasis altered).

See R. Brandt, Good and Right 126-29.
Recall that there is an intermediate case between Machiavellian and Libertine Dodger: Pliable Abe. Abe embraces the values of dishonesty, but not with sufficient intensity to make his doing the dishonest thing something one can count on.


See his discussion of Bishop Butler's concept of "self-love," which Brandt treats as a second-order desire. Good and Right 84-86.

There is an issue of causal over-determination here that needs to be addressed. Suppose the harmful act would have occurred without a motivational defect (e.g., because of stupidity), but the person also had a motivational defect that would have caused the act as well. One suggestion is that the person has still willed the harmful act so long as the relevant motivational defect is a necessary component of a set of causally sufficient conditions for performance of the act. This allows for the possibility of there being more than one set of causally sufficient conditions. It may be that stupidity is a necessary component of some other such set of causally sufficient conditions for performance of the act. In such a case, I would say that the person is morally blameworthy.

Actually, this may depend on the nature of the disorder. If it is cognitive, then substandard motivation is not implicated. But what if the impairment is a motivational one?

104 Id. 192-93.

105 Supra 116, note 89.

106 This is Brandt's view as well. R. Brandt, A Motivational Theory of Excuses in the Criminal Law 174.
5.1. A Motivational Theory of Justification

Against the foregoing background theory of criminal responsibility, how are we to understand criminal law justification defenses? The position that will be defended here is that in a properly constructed criminal justice system justification defenses, like excuses, defeat a normally warranted inference from proscribed conduct to substandard moral motivation (i.e., a motivational fault). Thus, justifications are properly classified with excuses as "exculpatory defenses."

According to the account presented in the preceding chapter, the criminal law is a motivational supplement to a social moral code. It performs this task by threatening (and delivering) punishment to those who violate important prohibitions of the moral code and thereby manifest a motivational fault. The fear of punishment will (it is hoped) deter those who might otherwise perform tokens of act-types against which the moral code would assign a strong aversion. Actual punishment, in addition to showing that the threats made are genuine, provides an opportunity for the community at large to condemn the criminal, thereby (hopefully) strengthening relevant moral motivation in the criminal and reinforcing the already adequate moral motivation of the average law-abiding citizen.
Certain types of conduct evidence a motivational fault. Those who purposely, knowingly, or recklessly harm others typically manifest varying degrees of indifference (lack of aversion) to causing the relevant type of harm. One who *purposely* harms another performs an act with the conscious object of harming that person, and so actually wants (on balance) to harm the person. One who *knowingly* harms another performs an act with an awareness (belief?) that it is practically certain that harm will result. Such an agent may not want (on balance) to harm another, but the conduct performed does evidence a substantial indifference (lack of aversion) to harming another. One who *recklessly* harms another performs an act despite an awareness (belief?) that the act poses a substantial risk of harming another. In such a case the act does not evidence the level of indifference to harm evidenced by purposely or knowingly harming another, but it does evidence a substantial indifference to causing harm.

Although harmful conduct that includes one of the above "culpability elements" is "incriminating," in the sense of evidencing a motivational fault, the presence of certain conditions, call them "excusing conditions," defeats the normally warranted inference. Such conditions as involuntary intoxication, infancy, extreme cognitive impairment, duress, etc., defeat the inference by indicating that some set of conditions not including a motivational fault may have caused the performance of the conduct. The claim is that the presence of such excusing conditions raises a reasonable doubt that the harmful conduct (including the statutory *mens rea*) would not have occurred without a relevant motivational fault. This is the "logic" of excuses.
Another way to defeat the inference from incriminating conduct to motivational fault is to argue that a reasonable person without a motivational fault would have been motivated to perform a token of the same (incriminating) statute-violating act-type if confronted with the same circumstances. To defeat the inference in this way is to offer a justification defense. As a first approximation, we may suggest the following analysis of the validity conditions of a criminal law justification defense:

\[ d \text{ is justified in performing a statute-violating (incriminating) act-token } a \text{ at } t \text{ in circumstances } c \text{ if and only if a reasonable person with at least the standard level of moral motivation assigned by the relevant social moral code, including that attached to violating the statute, would be motivated to perform } a \text{ at } t \text{ in } c \text{ by a belief that doing so would level-generate tokens of act-types } A_1 \ldots A_n, \text{ which act-types are also constituents of the social moral code.} \]

For reasons that will become apparent later, I propose to say that one who meets the above conditions has a \textit{prima facie} valid justification defense.

Let us apply the analysis to some paradigm cases, assuming that the relevant social moral code is our own. The druggist who administers a prescription drug to a heart attack victim under circumstances in which a reasonable person would conclude that calling a physician to request a prescription would greatly reduce the chance of saving the victim’s life, has a justification defense on the above analysis. The act-type, saving the life of a fellow human being, is strongly valenced in our moral code, much more strongly than an aversion to violating a law designed to prevent distribution of potentially harmful drugs without proper expert advice and other forms of drug abuse. A person with an acceptable level of aversion to violating
the law, who also has a strong desire to save human lives, would be motivated to violate the law if doing so was believed necessary to save a life.

The driver of a car who leaves the scene of an accident in order to rush a victim to a hospital emergency room, thereby violating a hit-and-run statute, also has a justification defense to a criminal charge based upon violation of the statute. At least this is true if a reasonable person in the circumstances would have concluded that failure to leave the scene immediately would pose a serious threat to the life or health of the accident victim. Again, the moral motivation to save a life or preserve the health of another is stronger than the motivation assigned by the social moral code towards conforming to reporting requirements designed to facilitate the fair adjudication of personal injury claims, or whatever other policies the hit-and-run statute may be designed to promote.

Arnold, who deliberately blows up several houses at the perimeter of a town, when doing so would appear to a reasonable person to be necessary to stop the advance of a raging forest fire, has a justification defense on the proposed analysis. This is true even if the wind shifts after he has blown up the houses, making their destruction unnecessary. A person with adequate moral motivation who believed that blowing up the houses was necessary to stop the fire would be motivated to do so, despite being adequately motivated to avoid destroying buildings owned by others.

The fact that it yields intuitively correct results in such cases is some evidence in support of the proposed analysis. It is doubtful, however, that
the analysis can stand without qualification. Consider the earlier example of the terrorist arsonist who just happens to blow up precisely the houses needed to save the town from an oncoming forest fire. If we have sufficient evidence of the terrorist’s motives, we may conclude that blowing up the houses does manifest a relevant motivational fault after all, viz., insufficient aversion to destroying the property of others, and perhaps insufficient aversion to causing serious physical injury or death to others as well. Given reliable evidence of terrorist motives, it hardly seems reasonable on a motivational account to afford the terrorist a justification defense.

How should we deal with examples like that of the terrorist arsonist? My suggestion is that we regard justification defenses as defeasible, allowing certain conditions to "defeat" the defense. On this view, the conditions stated above are to be taken as necessary conditions of a prima facie valid justification defense only. What conditions should be permitted to defeat an otherwise valid justification? In order to explore this question, let us consider the following cases: (1) A the terrorist, ignorant of an approaching fire, burns Arch’s field in an effort to send a political message to Arch (the town mayor) and other officials. As luck would have it, burning the field saves the town. (2) B, aware of the approaching fire, sees an opportunity to "get even" with Arch, whom B suspects of sleeping with his wife. B gleefully burns Arch’s field, thereby saving the town. (3) C is unaware of the approaching fire, has no particular grudge against Arch, but joins in burning the field because C enjoys watching fires. (4) D helps burn the field because D’s agents are at that moment engaged in a million dollar drug transaction in town. (5) E the arsonist, is on the way to burn Arch’s
field when E sees the approaching fire. E sets fire to Arch's field as planned.

Assuming that each of the above defendants can successfully establish a *prima facie* valid justification for burning the field, should any be denied reliance on the defense? If we examine the cases, we find that they divide into two classes. In the case of (2) and (4), even though the defendant may have acted partly from morally dubious motives, the motive that would move a reasonable well-motivated person to burn the field was also operative. In particular, both B and D burned the field in part because they believed that doing so would save the town. Admittedly, B's being motivated by this belief was due in part because of a belief that saving the town would be a good defense to burning Arch's field. And D had an ulterior motive as well (financial gain). But the fact remains that precisely the motivation that one would want to be effective in moving people to violate the arson statute was operative in both cases. In neither case would the defendant have burned the field absent a belief that doing so would save the town.

This is not true of the other three cases. The belief that burning the field would save the town played absolutely no motivational role in A or C's burning the field, for both were unaware of the approaching fire. E, by hypothesis, would have burned the field even if doing so would not have saved the town.

If we think of justification defenses as vehicles for encouraging the violation of statutes in situations where such violation will generate tokens
of act-types to which the social moral code assigns stronger moral motivation than it does to conformity to the statute, then the rationale for justification defenses would seem to dictate affording defendants B and D a defense. But since A, C, and E were not influenced by the relevant contrary moral motivation, a justification defense is not appropriate in their cases. This assessment of the cases is reinforced by the observation that denying B and D a justification defense might discourage such defendants from acting to save the town. Surely this is not a desirable result. But denying A, C, and E such a defense would have no effect on the tendency of such defendants to perform the act favored by the social moral code.

Examination of the above cases suggests the following justification defeating condition:

\[ (DC1) \quad d \text{ does not have an otherwise valid justification defense for performing a statute-violating (incriminating) act-token } a \text{ at } t \text{ in } c \text{ if } d \text{ was not motivated to perform } a \text{ at } t \text{ in } c \text{ by the "justifying belief" that } a \text{ would level-generate tokens of act-types } A_1 \ldots A_n. \]

There are two ways this can happen. The defendant lacks the relevant justifying belief altogether (A and C); or, even though the defendant has the belief, it plays no motivational role in the defendant's violation of the statute (E).

Are there any other justification defeating conditions? One suggestion is that perhaps demonstrating that d's belief that a will level-generate tokens of A_1 \ldots A_n was unreasonable should defeat an otherwise valid justification. After all, the \textit{prima facie} case for a justification requires that the belief be reasonable. If unreasonableness of the "justifying belief"
is to be a defeating condition, two limitations seem in order. First, the type of unreasonableness must be such as to evidence a relevant underlying motivational fault. Suppose, for example, that Arnold correctly believes that blowing up houses at the perimeter of town will save the town and that is why he blows up the houses. But now suppose we discover that Arnold was merely operating on a hunch, or that he consulted tea leaves, or that he failed to check the direction of the wind. Such carelessness on his part might indicate that Arnold is not really all that concerned about preventing harm to others. If he had cared enough, perhaps he would have been more careful in his calculations. He would not have relied on a hunch or consulted tea leaves when so much was at stake. He would have taken the time to check the direction of the wind. If so, then we might very well conclude that Arnold's motivation level is not acceptable after all.

On the other hand, if Arnold is simply very stupid, he may arrive at the conclusion that blowing up the houses is necessary, when a more intelligent, better informed person would not. But unless we have some reason for thinking that Arnold's unsupportable inference is ultimately the product of an insufficient aversion to causing harm to others, its being unreasonable ought not defeat an otherwise valid justification defense.

How unreasonable must the mistake be? A plausible suggestion is that it must be sufficiently unreasonable to amount to criminal negligence or recklessness. If this is the standard, then a second limitation seems to follow: unreasonableness amounting to criminal negligence or recklessness in arriving at the "justifying belief" may operate as a justification defeating condition only with respect to crimes of negligence or recklessness.
Showing that a defendant accused of an intentional crime was grossly negligent or reckless in arriving at a justifying belief does not establish the degree of motivational fault necessary for the primary offense. It should not, therefore, suffice to negate an otherwise valid justification. Otherwise, through the agency of the justification defeating condition a lower standard of motivational fault will have replaced the higher statutory standard. With these observations in mind, a second defeating condition may be proposed:

(FC2) \( d \) does not have an otherwise valid justification defense for performing a statute-violating (incriminating) act-token \( a \) at \( t \) in \( c \) if \( d \)'s "justifying belief" that \( a \) would level-generate tokens of act-types \( A_{1} \ldots A_{n} \) is sufficiently unreasonable as to amount to criminal negligence or recklessness in accepting that belief; such unreasonableness shall constitute a justification defeating condition only for criminal offenses for which negligence or recklessness, as the case may be, suffices to establish culpability.

This concludes the presentation of the theory of justification to be defended. The theory is a hybrid one. In order to raise a *prima facie* case that the defendant's incriminating conduct is permissible, the defendant must produce evidence that a reasonable adequately-motivated person would have been motivated to perform the act. This is an objective test. But since the *prima facie* case can be defeated by evidence of the defendant's actual motivation, mental states are also relevant to the ultimate validity of the defense.

One virtue of the motivational theory is that it provides a rationale for why the actual beliefs and motives of the defendant are relevant to whether a justification defense should be available. Recall that Fletcher provides
two arguments against a purely objective theory of justification and for the inclusion of a "justificatory intent" requirement: the privilege and exception arguments. Neither seems persuasive. If we view the criminal law as a motivational supplement to the social moral code, and we regard justifications as exculpatory defenses that defeat an inference from incriminating conduct to motivational fault, it becomes clear why proof that the defendant did not "advert to the relevant [justifying] circumstances" in acting deprives the defendant of a justification. If we discover that the defendant was unaware of the relevant circumstances, the fact that a reasonable adequately-motivated person would have been motivated by such an awareness becomes irrelevant to our assessment of the defendant's motivational state. The normally warranted inference from incriminating conduct to motivational fault is reinstated. Barring an excuse, we are entitled to conclude that the defendant is "guilty."

The proposed analysis relies upon a distinction between "incrimination" and "justification." Justification is a way of defeating the incriminating force of certain conduct. It does so by revealing the defendant's moral motivation for performing an incriminating act and arguing that the motivation thus revealed is not faulty. The distinction between incrimination and justification warrants further investigation, for it is not always clear how the distinction is to be drawn.

5.2 Incrimination versus Justification

In the taxonomy provisionally offered in Chapter I, it was said that both denials and offense modification defenses maintain that what a
defendant did is not within the intended scope of a criminal statute that forms the basis of the criminal charge. In the case of a denial, the claim is tied to specific language of the statute. In the case of an offense modification, it is not. Both defenses, however, are grounded on interpretations of legislative intent, and basically claim that the defendant’s conduct did not produce "statutory harm or evil." By contrast, a defendant who asserts a justification defense does not deny engaging in conduct that produced statutory harm or evil. The defense is that even if the conduct violated the terms of a statute and did produce statutory harm or evil, doing so was a necessary and effective means of avoiding an even greater evil, or was in some other way morally justified.

After considerable discussion we have arrived at a motivational interpretation of the distinction between denials and offense modifications, on the one hand, and justifications and excuses, on the other. If we assume that we are dealing with a properly constructed criminal justice system, i.e., one that operates as an effective, but not overly oppressive, motivational supplement to the social moral code, then the performance of an act that violates the express terms of a criminal statute will evidence a motivational fault on the part of the violator. Both denials and offense modifications maintain (for slightly different reasons) that such an "incriminating" act was not performed. Justifications and excuses do not deny performance of an incriminating act, but claim that the normally warranted inference to motivational fault is blocked. In the case of excuses this is done by pointing to some other condition (often a non-motivational cognitive fault)
that raises doubt about the causal role of the defendant's moral motivation. In the case of justifications the inference is blocked by alleging that even if the defendant's moral motivation was a cause of the conduct, the motivation in question is not faulty. This account of the logic of justification and excuse depends crucially upon a distinction between *incrimination* and *exculpation*.

The distinction can be a slippery one. Take self-defense. Is the absence of self-defense a necessary element of the incriminating use of force against another, or is the use of such force sufficient to incriminate, so that self-defense is better thought of as a justification? Do we, for example, define murder as "the intentional killing of another human being, not in self-defense," in which case we assign absence of self-defense to the incriminating dimension; or do we relegate self-defense to the exculpatory dimension of justification and define murder simply as "the intentional killing of another human being"? Is there any way to decide which issues properly belong to incrimination and which to exculpation? This is not an easy question to answer. One prominent criminal law scholar has recently identified the task of "working out the basis of the incriminating dimension of crime and relating this incriminating dimension to the exculpatory dimension of justification and excuse" as "the unmet challenge" of criminal theory.7

No doubt the best attempt to meet this challenge to date comes from the work of George Fletcher. Fletcher does not distinguish denials from offense modifications. He would, I think, regard both as negating what he calls the incriminating dimension, or "Definition" of crime. He thinks that it
is very important for criminal theory to distinguish the dimension of incrimination from that of exculpation. In the following passage he attempts to explain the significance of preserving the distinction between Definition and justification:

The basic idea behind the distinction between Definition and justification is that in our social and moral life we sense a difference between conduct that is routine and accepted and conduct that may be right, but that is rendered right only by providing good reasons. It is the difference between punching a ball and punching someone in the jaw. There is no need to justify punching a ball. It seems odd even to ask someone punching a ball to justify what he is doing. . . . Punching a person is different; this is conduct that is typically suspect. Yet in some cases it might be rendered proper and acceptable--say by self-defense, consent to a boxing match, perhaps even by a disciplinary privilege. . . .

Taking the controversy over abortion as our guide, we can see that there is a radical difference between the two types of cases in which we conclude that there is no "legal harm." To say that the fetus is not an interest worthy of independent protection is to classify the fetus along with organs of the mother's body. To appeal instead to a theory of justification is to acknowledge that the fetus is different from an appendix, but that other interests sometimes require the sacrifice of a concededly worthy interest. Though the result might be the same, the way we conceptualize the issue expresses our sense of the importance of the fetus as an independent interest.

Collapsing the distinction between Definition and justification generates a view of justified harm as though it were the same as consequences that fell beyond the scope of the criminal law. It is to treat aborting a fetus in the same way one treats killing a fly.

So one reason for preserving a distinction between Definition and justification is to provide some mechanism by which the criminal law can
recognize the moral difference between such cases as harming a fetus or an adult human being for good reasons (e.g., to save one's own life) and such cases as harming a fly for little or no reason. What is the moral difference between these two types of cases? Fletcher suggests that the difference lies in the importance of the interests being violated. The life of a fly is of no consequence and so one may "routinely" kill a fly without thereby violating an important interest. Killing a fly is not an "incriminating" act. There is no need to justify killing a fly. The life of a fetus or an adult human being is of great consequence and so one may not "routinely" (but only "exceptionally") kill a fetus or an adult human being. Killing a fetus or an adult human being is an incriminating act. Doing so requires justification.

No doubt there is something to this account of the difference between the non-incriminating act of killing a fly and the incriminating but justified act of killing another human being in self-defense. But an exclusive focus upon the nature of the interest being violated does not provide an entirely satisfactory account of incrimination. It does not, for example, explain why the typical criminal offense definition includes one of the Model Penal Code "culpability elements," or *mens rea.* Why, for example, define "murder" so that it consists of "purposely or knowingly causing the death of another human being"? An exclusive focus upon the nature of the interest being violated seems unable to provide an answer. For if an interest is harmed, it is equally harmed, whether the harmful conduct is accidental or intentional.

The motivational theory of criminal responsibility does, however, offer a plausible answer to this question. For according to this theory,
conduct that violates an important interest can be a reliable indicator of the state of an actor’s moral motivation. But it is so only if a psychological element is included. The reason for including *mens rea* in the criminal offense definition is to facilitate an inference from harmful behavior to motivational fault, and to the degree of motivational fault. Although Fletcher is right to distinguish killing a fly from killing a human being in self-defense, focusing exclusively upon the nature of the interests being violated does not provide an entirely adequate explanation of the difference.

Suppose we grant that it is important to distinguish incriminating from non-incriminating but justified conduct, we still do not have a way to decide which elements belong to incrimination and which to justification. Again, we look to Fletcher for some initial guidance. In the course of suggesting an answer to this question, he also provides a more detailed account of the dimension of incrimination. According to Fletcher, we identify the elements of incrimination, the Definition, "by finding [1] the minimal set of criteria that, [2] in the given society, [3] conveys a morally significant prohibition."\(^{11}\) Conditions of criminal wrongdoing not contained in this minimal set are to be assigned to the exculpatory dimension of justification.\(^{12}\) To illustrate, if we were to list out the possible sets of criteria to be included in the definition of murder, we might come up with a list of act-types something like the following:

1. killing
2. killing another human being
3. intentionally killing another human being
4. intentionally, but non-defensively, killing another human being
Now suppose we ask which of these act-types, if prohibited, would yield a morally significant prohibition. The answer, of course, depends on what is meant by "morally significant prohibition." In the following passage Fletcher sheds some light on this issue by commenting on the closely related concept of a "morally coherent imperative":

The minimal demand on the Definition is that it state an imperative that is morally coherent under given social conditions. What makes the imperative coherent is our comprehending the core cases of prohibited conduct. Whether there is a consensus on these core cases would seem to turn on several factors. One factor is the statistical relationship between routine and extraordinary cases. As cases of justified conduct become more numerous, it becomes increasingly difficult to think of the Definition as prescribing the normal, and the justification as providing the exception. In addition to statistical regularity, one needs a moral consensus supporting the wrongfulness of conduct in these core cases. As the moral consensus breaks down, more and more exceptions are urged, until one encounters the proverbial phenomenon of the exceptions swallowing the rule.13

Let us assume that what makes the prohibition of a given act-type morally significant is its constituting a morally coherent imperative. According to Fletcher, a morally coherent imperative is one that in some way expresses a moral consensus on a set of "core" acts. Although Fletcher claims that this moral consensus depends on several factors, at bottom it seems to be a statistical test that he has in mind. Suppose we take as our population the set of acts containing tokens of the prohibited act-type (e.g., killing another human being). Let us further suppose that within this population some of the acts are regarded by most members of society as on balance wrongful. Call these the "core" acts of the imperative. As I
understand Fletcher's proposal, there must be such a set of core acts if the imperative is to be morally coherent. But in addition, the relative frequency (within the population) of core acts must be fairly high. If either of these conditions fails, for example, there is insufficient consensus on the wrongfulness of putative core acts, or the relative frequency of core acts within the population with respect to which a consensus exists is too low, the imperative is not "morally coherent."

Looking at the above list of act-types, the act-type of killing would not seem to qualify as the basis of a morally coherent imperative. There are simply too many instances of killing that would not (in our society at least) be condemned as wrongful. Think of all the mosquitoes and flies killed in the course of a single summer. Killing another human being is certainly in a different category from killing flies, but it is not clear that a sufficient percentage of such acts would by general consensus be deemed wrongful. In addition to the killings of human beings that occur in self-defense, in the course of the proper exercise of force by law enforcement officials, and in circumstances of "necessity," many such killings are the result of non-negligent accidents. An imperative that covered all killings of other human beings might, therefore, be too inclusive to be morally coherent. If we add, however, that the killing is intentional, we seem to arrive at an act-type that meets Fletcher's test for moral coherence. More often than not, one who intentionally kills another human being has done something that by general consensus will be deemed wrongful. We might add that such a person typically is morally blameworthy for doing so. The performance of such an act is, therefore, "incriminating."
As the above account suggests, it is possible for a "morally coherent imperative," the violation of which is incriminating, to cease being so. Suppose that at one time abortions to save the life of the mother, although by moral consensus accepted as permissible, are relatively rare occurrences in the population of abortions. In such a case, having (or performing) an abortion might be an incriminating act that could be justified if performed to save the life of the mother. Now suppose that over time the relative frequency of life-saving abortions increases dramatically, while the frequency of other abortions decreases to the point where the typical abortion is a life-saving one. At some point the performance of an abortion per se will cease being incriminating. The imperative proscribing abortions will cease being "morally coherent," on the above account. It will be replaced by something like: "Do not have (or perform) non-life-saving abortions." If this happens, only non-life-saving abortions will be incriminating.

There is a temptation to assume that the "morally coherent imperatives" selected by Fletcher’s statistical criterion are meant to be identified with norms of a social moral code. But any such identification must be resisted. Suppose we lived in a society where people never (or rarely) killed other human beings unless there was good reason to do so (euthanasia, stopping insane person from causing great harm to others, defense from invasion by outsiders who threaten death or great bodily harm, etc.). On the above statistical interpretation, intentionally killing another human would not be typically wrongful in this society. From this it would
follow that a prohibition against intentional killing would not be "morally coherent." But surely members of the society may place a very high value on human life. In fact their doing so may explain the relative infrequency of unwarranted killings. In that case the intentional killing of another may be regarded by them as *prima facie* wrongful, in the sense that the social moral code assigns a strong aversion to such act-types. Instead of having a weak or nonexistent moral imperative forbidding the intentional killing of another, our imaginary society seems to possess an unusually forceful (and coherent) one.

If we are to specify the content of criminal law offense definitions by reference to the concept of incrimination, we must resist the assumption that these definitions will simply mirror *prima facie* moral rules of the social moral code. Nor should we assume that the conditions under which one may be called upon to provide a defense to a criminal charge will be the same as those calling for a defense to a comparable moral charge. Given that one who is suspected of committing a criminal offense may be subjected to a very oppressive liberty-invading process, even if the person is innocent, we may wish to set the threshold of incrimination fairly high, so as to minimize unnecessary invasions of individual liberty. Thus, in a society where abortions are routinely performed (and accepted) as a life-saving measure, it may indeed make sense to decide that the threshold of incrimination has been reached only if an abortion is not a life-saving one.

Suppose we grant that every offense definition ought to express a "morally coherent imperative." Clearly, not every morally coherent imperative should be expressed in an offense definition. According to
Fletcher, only imperatives that contain the *minimal* set of criteria required to state a morally coherent imperative should be expressed in an offense definition. So even though a prohibition of the fourth act-type listed above (intentionally, but non-defensively, killing another human being) would be a morally coherent imperative, it is not a definition-informing one because it contains criteria (properties) beyond those required to state a morally coherent imperative. To put the point in a slightly different way, in order to reach the threshold of incrimination, it is not necessary to add that the intentional killing of another human being is not in self-defense. Consequently, absence of self-defense should not be included in the incriminating dimension (the definition) of the crime of murder. The element of self-defense belongs to the exculpatory dimension: it is a justification.15

Why stop at the minimum, one might ask. Douglas Husak offers the following possible response:

But why stop at the minimum? Since Fletcher does not directly address this question, an answer will have to be attributed to him. Stopping here, I suppose, guarantees that components of liability that exceed this minimum threshold are categorized as (the absences of) justifications. If offenses were comprised of the *maximum* set of components that incriminate the defendant, the class of justifications would be empty. Since the line must be drawn somewhere (if the distinction between denials and justifications is to be preserved), it is best to stop at the minimum. If a more plausible reconstruction of Fletcher’s reasoning is forthcoming, it will have to be provided by Fletcher himself.16
I think that one can do a little better than the slippery slope argument offered by Husak to support drawing the line where Fletcher does. Even if, because of an interest in protecting liberty, we set the threshold of incrimination fairly high, once that threshold is reached the competing interest in public safety becomes paramount. Those who have engaged in incriminating conduct should be apprehended and run through the criminal process.

Even if it is possible to provide a theoretical defense of Fletcher’s criterion for distinguishing definition from justification, problems of application remain. Consider Fletcher’s own application of the test to the question of whether consent should operate as a justification for rape, or non-consent should be an element of the definition. He lists, "in order of ascending incrimination," the following act-types:

1. touching
2. [engaging in] sexual contact
3. [engaging in] forcible sexual contact
4. [engaging in] non-consensual, forcible sexual contact

Fletcher thinks that the threshold of incrimination is probably reached at the second level, but certainly is reached by the third:

It is difficult to argue that touching per se is incriminating. In some societies, all forms of human contact might be regarded as trespassing on the domain of another, but we regularly accept a gentle hand on the shoulder as both the price and the benefit of group living. Sexual contact is obviously different. Intimate touching of the genitals is hardly routine; the touching requires a good reason. The reason, or the justification, might be the consent of the person touched or it might be the necessity of performing an operation in an
emergency situation. This seems to me to be sufficient to regard the definition of rape as sexual penetration, with consent functioning as a ground for regarding the sexual act as a shared expression of love rather than as an invasion of bodily integrity.

The case of Morgan is even clearer, for the penetration was forcible. It is conceivable that a woman would enjoy being taken by force and that her consent would justify the forcible penetration. But it would be implausible to treat non-consent as well as force as necessary conditions for rendering the sexual act suspect.

It must be admitted that the line separating definition from justification is not always clear. Even Fletcher seems a bit unsure whether in the above case it is to be drawn at level two or three. Moreover, if we follow Fletcher’s intuitions and draw it at level two, does it not follow from the proposed test that the absence of force is a justification (or perhaps an excuse)? If so, a defendant who had sex with a drugged or unconscious victim could successfully plead absence of force to justify (or excuse) the act. Application of Fletcher’s test seems to misclassify an element that clearly belongs in the definition of rape.

In addition, application of the test requires generation of a list of elements (act-types) in order of increasing incrimination. But there does not seem to be a unique list corresponding to each criminal offense. Husak suggests the following alternative to Fletcher’s list for rape:

1. touching
2. forcible touching
3. forcible sexual touching
4. nonconsensual forcible sexual touching.
Here is another:

1. touching
2. nonconsensual touching
3. nonconsensual sexual touching
4. nonconsensual forcible sexual touching.

And another:

1. touching
2. nonconsensual touching
3. nonconsensual forcible touching
4. nonconsensual forcible sexual touching.

If we follow Fletcher's lead, the classification of a given element will depend on which of these lists one selects. But Fletcher offers no guidance on this selection.

Here we may benefit from the insight that the motivational theory of criminal responsibility brings to bear on the concepts of incrimination and justification. On this view, conduct is incriminating to the extent that it evidences a motivational fault in the actor. Justification consists in providing reasons for thinking that the actor's moral motivation is not faulty, despite the performance of incriminating conduct. Using this as our guide, it is clear why certain elements, such as the absence of force or the fact that the touching was not sexual, does not justify an otherwise incriminating act: such conditions do not show that the underlying moral motivation is not faulty.

The method of simply listing out the various possible act-types in order of increasing incrimination, identifying some properties as belonging to the Definition and allocating the rest to justification, will not work. We
may be able in this way to identify the minimum content of incrimination, if
we could agree upon a single correct list. But this seems most unlikely.
And even if a single list were universally agreed upon for a given crime, and
we were to identify the minimum content of incrimination, we cannot just
allocate all of the remaining elements to the exculpatory dimension of
justification. We need a theory of justification, like that provided by the
motivational theory, by which to judge the justificatory relevance of the
various elements.

One final comment is in order before we leave the topic of
incrimination versus justification. The fact that one can construct a number
of different lists of increasingly incriminating act-types, some of which
overlap or converge, does not jeopardize the account of criminal law
justification offered here. However one defines a given type of
incriminating conduct, the logic of justification is the same. Incriminating
conduct, if it is genuinely incriminating, evidences a motivational fault. A
justification defense is one that defeats the inference to motivational fault
by maintaining that under the circumstances in which the incriminating act
was performed, there was nothing faulty about the actor's motivation.  

5.3. The Serial View of Defenses

Some scholars endorse what Douglas Husak has called the "serial
view" of criminal law defenses. According to this view, the various
categories of criminal law defenses identified in the taxonomy of defenses
offered in Chapter I may be located along a continuum: "Each successive
type of defense need be considered only if those preceding it are
unavailable.' If a defendant *denies* the commission of a criminal offense, issues pertaining to other defenses cannot arise, and so on along the defense spectrum: 'Justification and excuse are serial rather than alternative determinations. One asks whether an actor should be *excused* only after he has determined that the act was not *justified*. If the act was justified, there would of course be nothing to excuse.' . . . According to this conception, a defendant who alleges a given defense should be understood to concede the unavailability of each defense prior to it along the spectrum."

Husak rejects this serial view of defenses, based on examples like the following:

Suppose that Sue is attacked without provocation in her home by a thug wielding a knife. She retreats to her bedroom with her assailant in hot pursuit. The thug provides every indication of his intent to kill her. Sue is an extreme pacifist who would rather lose her own life than kill another. Still, she hopes to scare her attacker away by brandishing a hunting rifle she reasonably believes to be unloaded. Much to her horror, the rifle is loaded; she accidentally fires it, and kills her assailant.

Several valid defenses are available to Sue in the unlikely event that she is charged with murder. She can *deny* that her killing was intentional; she can *justify* her conduct as involving self-defense; or she can *excuse* her killing as mistaken or accidental. Given her pacifist convictions and the remorse she would be expected to feel about her conduct, it would not be surprising if her inclination were to excuse her killing. Use of this defense, however, should not be taken as an admission that no denial or justification is available. But the serial view of defenses entails these very concessions. According to this view, it is incoherent for Sue to defend by alleging a denial or justification in the event that her excuse is disbelieved or for some reason not recognized as valid."

--

23

24
It seems to me that Husak’s proposed counter-example to the serial view is not persuasive because it fails to satisfactorily distinguish the issue of whether defenses in more than one defense category can be simultaneously *pleaded* from the issue of whether a defendant may in fact simultaneously *have* valid defenses in more than one category. In addition, the example gains some apparent plausibility by misclassifying some of the defenses discussed and by failing to take proper account of whether all of the valid defenses available are responses to the *same charge*.

As I understand it, the serial view does not preclude a kind of serial conditional *pleading* of defenses. That is, there is nothing in the view that prevents Sue from denying that she intentionally killed her assailant; while also asserting that if that claim is disbelieved, an intentional killing was justified in the circumstances; and at the same time maintaining that if the claimed justification is not accepted, the circumstances were such that she has a valid excuse for the unjustified killing. Moreover, there is nothing in the serial view to preclude Sue from forgoing the first two pleas (as Husak suggests she might want to do) and resting her case on the excuse defense. What the serial view does (apparently) preclude is the possibility of multiple defenses in different categories being simultaneously valid. Thus, it would seem to follow from the serial view that

(1) If $d$ has a valid denial to a criminal charge, it is false that $d$ has either a valid justification or a valid excuse to the charge.

(2) If $d$ has a valid justification to a criminal charge, it is false that $d$ has a valid excuse to the charge.

And it follows from (1) and (2) that
(1') If $d$ has either a valid justification or a valid excuse to a criminal charge, it is false that $d$ has a valid denial to the charge.

(2') If $d$ has a valid excuse to a criminal charge, it is false that $d$ has a valid justification to the charge.

It is (1') and (2') that suggest the consequence that a defendant who pleads a defense in the series concedes the unavailability of defenses "prior" to it in the series. This troubles Husak because he would like Sue to be able to plead an excuse without forgoing the option of pleading denial or justification in the event that the excuse fails. Strictly speaking, the serial view does not bar Sue from doing so. If in the course of making the case for a valid excuse, Sue (or her attorney) decides that another tack is called for, she could plead either of the other two types of defense. In doing so she would, on the serial view, have to abandon the excuse defense. Moreover, it is quite possible that in attempting to make the case for an excuse she will have hurt her chances of success on either of the other two defenses. After all, the logic of the excuse argument does, according to the serial view, assume the unavailability of denials and justifications. There may be, therefore, good strategy reasons for not making an excuse one's first line of defense.

It would appear that the motivational theory of this dissertation entails a modified serial view of defenses. On the motivational account, a valid denial precludes both valid justifications and valid excuses. Unless conduct is incriminating, there is nothing to justify or excuse. Thus, the motivational theory accepts (1) and (1') above as implications of the logic of defenses. But it need not accept (2) and (2'). It is inconsistent both to
deny that one violated a statute and at the same time claim that one’s violation of the statute was justified or excused. But there would appear to be no inconsistency in asserting both that a reasonable adequately-motivated person would have been morally motivated to violate the statute in the circumstances (justification), while at the same time maintaining that conditions were present that block an inference from statutory violation to a motivational fault (excuse).

If we apply this analysis to Sue’s case, it follows that if she has a valid denial based upon her claim that she did not intend to kill her assailant, she cannot at the same time have either a valid justification or excuse. Husak’s discussion of Sue’s case is misleading, for it classifies mistake and accident as excuses, when in fact these defenses are functioning to negate statutory intent. Consequently, they qualify as denials to a murder charge, not as genuine alternatives to a denial.

If we stick to the facts of the hypothetical described by Husak, it would seem that any valid excuse that Sue might have that would be consistent with her also having a valid denial to the murder charge would not itself be a defense to the charge of murder. The gist of her excuse claim is that under the circumstances she was so emotionally distraught as not to be thinking clearly when she brandished the weapon. In a calmer moment she would never point a gun at another human being, even if she had reason to think that it was unloaded. Although, if she were in a normal emotional state, we might be able to infer from her conduct that she has a degree of moral motivational fault that is typically indicated by an act of negligent or
reckless homicide, her disturbed condition blocks this inference in the present circumstances. Her excuse, then, is a defense to negligent or reckless homicide, not murder (an intentional homicide).

It may also be possible for Sue to have a valid justification defense, while at the same time having a valid denial to the murder charge. But such a justification would be offered, not as a defense to murder, but as a defense to some other charge (negligent or reckless homicide). The claim would be that the act of brandishing the rifle in what would in normal circumstances amount to criminal negligence or recklessness, was in the particular circumstances of repelling an assault, an act of justified self-defense. A reasonable person with adequate moral motivation would have been motivated to act as she did under the circumstances. Note that the excuse and justification that are consistent with Sue's denial of the murder charge are consistent with each other. Sue may consistently claim that a reasonable adequately-motivated person would have brandished the rifle in the way she did, and that her emotional state was such as to block an inference from violation of a criminal law statute (proscribing negligent or reckless homicide) to motivational fault. Husak's proposed counter-example is persuasive against the modified serial view advocated by the motivational theory. It appears plausible because it fails to adequately distinguish conditional pleading of multiple defenses from possessing multiple valid defenses, it misclassifies some of the defenses discussed, and it fails to take note of the fact that the multiple defenses available may not all be responses to the same criminal charge.
5.4. Justification and Universalization

Many theorists maintain that criminal law justification defenses are capable of being universalized. In fact, this feature of justifications is often cited as a basis for distinguishing justifications from excuses. George Fletcher, puts the point this way:

Claims of justification lend themselves to universalization. That the doing is objectively right (or at least not wrongful) means that anyone is licensed to do it. . . . Excuses, in contrast, are always personal to the actor; one person's compulsion carries no implications about whether third parties will be excused if they act on behalf of the endangered defendant.

An illustration may help lend some *prima facie* plausibility to this basis for distinguishing justifications from excuses. A (once again) decides to set fire to B's field in order to create a firebreak that will stop a raging forest fire from engulfing an entire community. According to the universalizability of justification thesis, if A is justified in torching B's field by virtue of that act's saving the community, then any other person would also be justified in torching the field if doing so would save the community. But now suppose that A, while in a severe psychotic state, suffers from the delusion that a raging forest fire threatens the community. A torches B's field under the mistaken belief that doing so is necessary to save the community. A's act is not justified. Further suppose that we decide that the psychotic state provides A with a valid excuse. According to the non-universalizability of excuse thesis, it does not follow that any other person would be excused for torching B's field. Non-psychotic Ned, who torches the field even though he does not think it necessary for some greater good, may have no excuse.
Although examples like the above lend some intuitive plausibility to the universalization criterion, one must be careful to insure that the claimed universalizability of justifications is properly understood. We may grant that if A burns B’s field in circumstances in which a reasonable adequately-motivated person would be morally motivated to do likewise, barring some defeating condition, A is justified in torching the field. Moreover, every person who burns B’s field under the circumstances has the same prima facie justification for doing so. But if the examples appealed to in making the case for the defeasibility of justifications are persuasive, not everyone, regardless of their beliefs or motivation, may be (on balance) justified in torching the same field. So what is universalizable is the prima facie case only.

It is also worth noting that, as our previous discussion of alleged implications of the justification versus excuse classification in the area of accomplice liability suggests, one can also universalize excuses if one includes reference to the psychology of agents in the principle to be universalized. Consider once again the example of psychotic A who torches the field because of an insane delusion that there is an approaching fire that threatens the community, etc. If some other person suffers from the very same type of psychological condition, that person also has an excuse for torching the field on precisely the same grounds as A. One must be careful, therefore, not to overestimate the differences between justifications and excuses vis-a-vis universalizability. Still, the fact that the prima facie case for a justification is universalizable does provide a mechanism for
avoiding certain problems that might otherwise arise from a failure to preserve a sufficiently sharp distinction between justifications and excuses.

5.5. Justification or Excuse: Sending the Wrong Message?

In our earlier discussion of Robinson's objective harm theory we noted that the failure to preserve a sharp distinction between justifications and excuses may adversely affect the deterrent function of the criminal law and foster judicial inconsistency as judges fearful of "sending the wrong message" to the public refuse to make available in practice excuses that the law recognizes in theory.29 I expressed some doubt then about having to adopt a purely objective theory of justification in order to preserve a sufficiently sharp distinction between justification and excuse to avoid these potential problems.

It should be clear that the proposed hybrid theory of justification does not threaten to blur the distinction between the two categories of exculpatory defenses to the point of undermining deterrence or encouraging judicial inconsistency. In order to prevail on a justification plea, the defendant must be able to show that a reasonable adequately-motivated person would have been morally motivated in the same circumstances to act as did the defendant. What the defendant did was not criminally wrongful. Such conduct is to be tolerated, even encouraged. There is no danger that a court that says that the conduct was "justified" will be sending the wrong message.

Similarly, a defendant who pleads an excuse, on the grounds of mental disability, involuntary intoxication, etc., is making a different kind
of claim. No attempt is made to condone the conduct. It may not have been conduct that a reasonable well-motivated person would have performed under the circumstances. It certainly was "incriminating." But we are not entitled to conclude from this that it manifested a motivational fault sufficient to sustain criminal liability. So long as the bases for finding a defendant "not guilty" are made clear, there should be no threat to deterrence, nor reason for judicial temerity.

5.6. Self-defense

In the context of our earlier critique of Robinson's objective harm theory of justification, it was suggested that a more satisfactory understanding of self-defense might be arrived at if we avoid focusing exclusively upon the conduct (or status) of aggressors and include within our purview characteristics of the defender as well. Against the background of the motivational theory of justification, we may now develop this suggestion a bit further.

A social moral code designed to be a practical guide to conduct should take note of such very strong natural human motives as the motivation towards self-preservation. A code that sought to build contrary altruistic moral motivation sufficient to overcome the motive to survive would surely have set itself a difficult, if not impossible task. Moreover, the social costs of instilling and maintaining such powerful moral motivation would seem prohibitive. And it is not even clear that it would be socially desirable to achieve the level of altruistic motivation necessary to overcome
the natural motivation to survive, if indeed that could be accomplished. There is something to be said for encouraging strong self-interested motives, for doing so may yield a society of strong self-reliant individuals.

Add to these observations that one of the objects of every existing social moral code is (and should be) the discouragement of unwarranted aggressive attacks that threaten the life or physical integrity of members of the moral community. Given this goal, it may make sense to actually encourage use of force in self-defense as a disincentive to such aggression. This can be translated into a moral "right" of self-defense against guilty aggressors. In order to make it very clear that innocent defenders are to be preferred over guilty aggressors, a moral code could reduce the motivation to avoid harming the interests of such aggressors. This can be cashed out by reference to a forfeiture of rights theory, or a view that the rights of defenders always override competing rights of guilty aggressors, regardless of the number of aggressors involved.

The issue of the validity of self-defense against multiple innocent aggressors is a more difficult one. But the motivational account suggests a possible rationale for recognizing self-defense as justified even in these cases. Of course, one cannot (for the reasons mentioned in our earlier discussion) rely on a forfeiture rationale. But perhaps the recognition that the motive to survive is a very powerful one, perhaps so powerful as to guarantee that any attempt to resist it with contrary moral motivation will fail or be so costly to maintain as not to be socially desirable, can suffice to support a moral right of self-defense even in these cases. I am not sure. At any rate, a criminal justice system operating to supplement the moral
motivation of a social moral code should reflect the contours of the moral right of self-defense, however those are to be drawn precisely.

5.7. Conclusion

The dominant theories of criminal law justification in the contemporary marketplace of ideas all tend to start with an assumption that such defenses, in sharp contrast to excuses, essentially involve an "objective" evaluation of conduct. Whereas an agent's motivation may be relevant to whether a valid excuse should be recognized in certain situations, motivation is either deemed irrelevant to the validity of justifications, or is tacked on as an afterthought to deal with potential counter-examples. The approach here has been to think of the logic of justifications as akin to that of excuses in that both defeat an otherwise warranted inference from conduct to motivational fault. Recognizing the common ground between justifications and excuses helps explain why they are so often grouped together as "exculpatory" defenses, while recognizing the intuitive force of the distinction between them. This final chapter has merely scratched the surface of the issues that could be discussed in connection with this account of criminal law justifications. But though much remains to be done, it is hoped that enough has been said to provide a prima facie case for recognizing the motivational account as a worthy competitor in the ongoing debate over how best to understand the "logic" of criminal law defenses.
ENDNOTES

1 To simplify initial presentation of the theory of justification being advocated, reference to MPC negligence has been omitted here. Model Penal Code negligence raises issues that require further clarification of the proposed analysis. In particular, it seems difficult to maintain a distinction between incrimination and justification for crimes of negligence. This issue and others arising in connection with the distinction between incrimination and justification are discussed infra at 249, note 21.

2 The definitions of "purposely," "knowingly," and "recklessly" being relied upon are roughly those recommended by the Model Penal Code. MPC § 2.02. One significant difference is that the Model Penal Code definition of "recklessly" requires that the substantial risk in question be "unjustifiable." No such qualification is built into the definitions of "purposely" and "knowingly." This qualification and a similar one appearing in the MPC definition of "negligence" are discussed infra at 249, note 21.

3 Thus, excusing conditions cast doubt on the claim that the "best explanation" of the incriminating conduct is one that includes a motivational fault as a necessary causal condition.

4 The defeasibility approach seems a natural way to incorporate a plausible evidentiary presumption that one who acts in the way a reasonable well-motivated person would act is doing so for the reasons that such a person would act. In other words, we are to presume innocence (acceptable motivation) unless there are very good reasons for thinking otherwise. It is reasonable to expect the State, if it is to subject a person to criminal punishment, to bear the burden of proof (production of evidence and risk of non-persuasion) that an incriminating act (violation of statute) has been performed. Assuming this is accomplished, the defendant might reasonably be expected to provide some defensive response (or face conviction). If the defendant expects to rely upon a justification, some evidence that the
incriminating act was morally justified must be produced. I am suggesting that it should suffice for this purpose for the defendant to argue (with supporting evidence) that a reasonable adequately-motivated person would have been motivated to violate the statute under the circumstances. Assuming this is done, the State has two options. It may argue (with supporting evidence) that a reasonable adequately-motivated person would \textit{not} have been motivated to violate the statute, or it may attempt to \textit{defeat} the justification by providing evidence (roughly) that the \textit{defendant} did not act for the reasons that a reasonable adequately-motivated person would have acted under the circumstances. In either case, I would require that the State bear the burden of persuasion beyond a reasonable doubt.

\textsuperscript{5}To some extent this might be mitigated by offering an excuse. But on what grounds would one base such an excuse? There would appear to be no disability or other condition available to block an inference to substandard motivation. In fact, we are assuming knowledge of their motives for acting. Otherwise, there would be no question about their having a valid justification defense.

\textsuperscript{6}\textit{Supra} 127.

\textsuperscript{7}Fletcher, \textit{The Unmet Challenge of Criminal Theory}, 33 Wayne L. Rev. 1439, 1443 (1987).

\textsuperscript{8}G. Fletcher, \textit{The Right Deed for the Wrong Reason: A Reply to Mr. Robinson}, 23 UCLA L. Rev. 293, 310-11 (1975) (emphasis added).

\textsuperscript{9}In addition to providing a mechanism for recognizing the moral difference between the "routine" and the incriminating but justified violation of an interest, Fletcher suggests four "areas of legal dispute" where the distinction between Definition and justification could have important implications: "First, it is of critical importance in deciding when external facts, standing alone, should have an exculpatory effect. Secondly, it might bear on the analysis of permissible vagueness in legal norms. Thirdly, it might bear on the allocation of power between the legislature and judiciary in the continuing development of the criminal law. And fourthly, it might be of importance in analyzing the exculpatory effect of mistakes." Rethinking 555. A fifth area of legal dispute might be added to this list: the appropriate allocation of the burden of persuasion on issues bearing on criminal liability. See Fletcher's discussion of this issue in Rethinking 516-52.
Evidently, Fletcher does not find this test entirely satisfactory, since it predates the *Unmet Challenge* article. This may be a good place to compare Fletcher's distinction between incrimination and exculpation to H.L.A. Hart's distinction between offense definition and Distribution. H.L.A. Hart, *Punishment and Responsibility* 6-12. Fletcher's line may be one to be drawn within Hart's category of offense definition. But more likely Fletcher would accuse Hart of over-simplifying the structure of the conditions of criminal liability. On this view, Hart's definition-distribution dichotomy fails to recognize an intermediate category of justification. Fletcher adopts a concept of criminal wrongdoing that includes Hart's definition component, but supplements it with a justification component. On Fletcher's view, the structure of criminal liability looks like this:

Wrongdoing

```
<table>
<thead>
<tr>
<th>Definition (Incrimination)</th>
<th>Justification (Excuse)</th>
<th>Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Exculpation</td>
</tr>
</tbody>
</table>
```


I omit acts of war. Presumably, any imperative that might form the basis of a criminal homicide charge would be understood to exclude such killings.

This assumes, of course, that self-defense is not properly classified as an excuse.


Rethinking 705.
Rethinking 705. In the *Morgan* case the House of Lords held that mistake as to consent, whether reasonable or not, is a valid defense to a rape charge. The case involved four defendants who had forcible intercourse with the victim. They had allegedly been told by the victim’s husband that resistance should be expected, but would be feigned. The defendants were convicted under instructions that in order to constitute a valid defense a belief in consent must be *reasonable*. On appeal they successfully argued that *any* belief in consent, whether or not reasonable, negates the definitional element of intent to rape, and so is a valid defense. In recognizing the merit of the defendants’ argument the House of Lords in effect treated non-consent as an element of the definition of rape. Although agreeing with the defendants’ legal argument, the Lords upheld the conviction on the grounds that the trial court’s instructions were not prejudicial. Regina v. Morgan, 2 W.L.R. 923 (1975).

Husak makes a similar point about the element of intent in the context of a murder charge. If we list out the possible elements in the manner suggested earlier, according to Husak we reach the threshold of incrimination with the act-type "killing another human being." If so, then we must assign absence of intent to the dimension of exculpation. Absence of intent to kill would then have to be treated as a justification or excuse for killing. But intent is almost universally regarded as an element of the definition of murder. D. Husak, *Philosophy* 201. I think that the reason intent is almost universally regarded as an element of the definition is that most people do not think that the threshold of incrimination has been reached at "killing another human being."

I would be remiss if I were not to comment on a complication that arises in the context of crimes of recklessness and negligence. The above account relies upon a distinction between incrimination and exculpation. The claim was made that it is reasonable to include Model Penal Code culpability elements (*mens rea*) in statutory definitions because such definitions should proscribe act-types that meet, but do not substantially exceed the threshold of incrimination. This line was recommended as representing a reasonable balance between the interests of individual liberty and security. The suggested assignment of the MPC *mens rea* elements seems plausible for intentional crimes (crimes of purpose or knowledge), since one can define "purposely" and "knowingly" without any mention of reasonableness or unreasonableness of any risk of harm that one’s conduct may pose to others. But it is less clear that one can preserve the purity of
the incrimination versus justification distinction in the case of crimes of recklessness and negligence. When we look at the MPC definition of "recklessly" we discover that one who recklessly harms another "consciously disregards a substantial and unjustifiable risk that [harm] will result from his conduct." And a person who negligently harms another does so without being aware of a "substantial and unjustifiable risk" of harm of which he should have been aware. MPC § 2.02.

The asymmetry between the definitions of "purposely" and "knowingly," on the one hand, and "recklessly" and "negligently," on the other, may be attributable to the drafters' belief that the ordinary concepts of recklessness and negligence entail that the (risky) conduct is unjustified, while those of purpose and knowledge do not. The claim that a person's reckless conduct is justified does sound a bit odd, whether or not it is self-contradictory. Nevertheless, it seems to me conceptually tidier to separate the issues of incrimination (prima facie wrongdoing) and justification (on balance wrongdoing) embedded in an accusation of recklessly harming another. Doing so without producing confusion might require use of a more neutral description of the relevant statutory mens rea, e.g., "riskfully" rather than "recklessly," or perhaps the more cumbersome expression, "consciously disregarding a substantial risk of harm" would do. If one were to adopt this proposal, then one might substitute statutes that proscribe harming others as a result of consciously disregarding a substantial risk of harm, for those proscribing recklessly harming others. This would enable one to separate out the issue of whether or not in a given case the conscious disregard of a substantial risk of harm was justified.

If we take a close look at the MPC definition of "recklessly," we find that the issues of incrimination and justification are in fact conceptually distinct: "A person acts recklessly with respect to [harming another] when [1] he consciously disregards a substantial [and unjustifiable] risk that [harm] will result from his conduct. [2] The risk must be of such a nature and degree that, considering the nature and purpose of the actor's conduct and the circumstances known to him, its disregard involves a gross deviation from the standard of conduct that a law-abiding person would observe in the actor's situation." MPC § 2.02(2)(c). The second sentence simply spells out what the code drafters take to constitute an unjustifiable risk. We could, then, delete the reference to the issue of justifiability included in [1] without loss of meaning, and take the remainder to articulate the conditions under which one's risky behavior is incriminating. If we do so, [1] and [2] will articulate conceptually distinct requirements of liability: [1] incrimination and [2] absence of justification.
The distinction between incrimination and justification becomes more difficult to maintain in crimes of negligence. According to the Model Penal Code definition, "a person acts negligently with respect to [harming another] when [1] he should be aware of a substantial [and unjustifiable] risk that [harm] will result from his conduct. [2] The risk must be of such a nature and degree that the actor's failure to perceive it, considering the nature and purpose of his conduct and the circumstances known to him, involves a gross deviation from the standard of care that a reasonable person would observe in the actor's situation." MPC § 2.01(2)(d). If we try the same move here as was suggested in the case of reckless conduct, we immediately run into problems. The second sentence does not (merely) spell out the conditions under which the risk of harm is unjustified, it provides the conditions under which the actor should be aware of the substantial risk of harm. Thus, deleting the explicit reference to unjustifiability in [1] does not result in two conceptually distinct requirements of liability, for (according to what remains of [1]) in order for (negligent) conduct to be incriminating it must be unjustifiably risky. If we neuter the normative expression, "should be aware," and replace it with "is not aware," it is arguable that [1] no longer is incriminating. No doubt there are many substantial risks that each of us poses to others in the course of a lifetime of which we are unaware. This in and of itself is no evidence of substandard moral motivation. It seems that in the case of negligence incrimination may require absence of justification. Perhaps we should just admit that negligence is a special case where incrimination and absence of justification merge. If we were to do so, then there could be no valid justification defense to a crime of negligence. That does not strike me as an implausible result. The claim, "I was justifiably negligent," may be self-contradictory.

22Philosophy 194.

23Philosophy 194. The first internal quote is from Robinson, Defenses 105. The second is from Robinson, Justification 266, 275.

24Philosophy 195.

25Recall that this does not mean that she cannot plead more than one type of defense conditionally.
26 Paul Robinson is such a theorist. See the discussion of his theory of justification, *supra*, Chapter III.

27 Rethinking 761-62.

28 *Supra* 121, 126.

29 *Supra* 106.

30 *Supra* 114.


Brandt, R. B. A Utilitarian Theory of Excuses, 78 Philosophical Review 337 (Jl 69).

Brandt, R. B. Fairness to Indirect Optimific Theories in Ethics, 98 Ethics 341 (Jan. 1988).


Hudson, S. *Character Traits and Desires*. 90 Ethics 539 (1980).


