9

When the Rule Swallows the Exception

CLAIRE OAKES FINKELSTEIN*

I. The Logic of Exceptions

THERE is an intriguing problem in the criminal law about the relation between the elements of an offence and the affirmative defences that apply to it. The problem arises from the fact that there are two quite different ways of structuring any defence: the defence can be incorporated into an offence definition by making it what we might call a "negative offence element", or it can be set out in a separate provision and treated as a so-called "affirmative defence". For example, the permission to kill in self-defence can be incorporated into an offence like murder by including a requirement in the offence definition that the defendant *not* have killed in self-defence. Alternatively, the defence can be articulated in a general, abstract provision that applies to all relevant offences, while no mention of the privilege appears in any of the definitions of those offences. An important question, then, is when to treat a particular defence provision as a negative offence element and when to treat it as an affirmative defence.

The foregoing difficulty in criminal law raises the more general jurisprudential puzzle of what we might call the *logic of exceptions*[1], namely the question of the relation between rules of prohibition and the exceptions that qualify them. The importance of the larger topic stems from the fact that even the most stringent of prohibitions appears to be subject to myriad exceptional circumstances in which the prohibited conducted is permissible. One is not allowed to break promises, permanently deprive someone of his property

* Acting Professor of Law, University of California at Berkeley (Boalt Hall). I am grateful to Brian Bix, Scott Shapiro, and Benjamin Zipursky for conversations on the topic of this essay, to Peter Detre and Sanford Kadish for comments on various drafts, and to Corby Sturges for his assistance with research and for his comments on an earlier draft.

[1] The expression is taken from the title of an article by Glanville Williams. Glanville Williams, "The Logic of 'Exceptions'", (1988) 47 *Camb. L. J.* 261.

without his permission, unjustifiably deprive someone of his life, and so on, but there are circumstances under which one may do all these things.[2] One question about this interesting feature of our moral lives is why rules of prohibition remain rule-like in the face of an abundance of exceptions to them. Is there anything special about the sorts of qualifications that can apply to rules of prohibition without eroding the normative force of those rules? Is there anything special about the sorts of rules that can count as prohibitions in the face of numerous exceptions to their authoritative force?

Given the significance of rules of prohibition to legal reasoning, and the significance of exceptions to rules of prohibition, it is striking that the topic of exceptions has largely been ignored by legal philosophers. The criminal law literature itself contains only a small handful of fairly technical articles on the topic, and these are quite specific to the relation between offences and defences.[3] The larger jurisprudential literature consists mostly of stray remarks in longer works devoted to other subjects.[4] As Fred Schauer says, the exception is "an invisible topic in legal theory . . .", one about which it has been thought that "no interesting generalizations are to be derived", despite the fact that exceptions "surface almost everywhere throughout most legal systems".[5]

While Schauer does not himself attempt a diagnosis of the paucity of attention to exceptions,[6] two hypotheses might have occurred to him had he done so. The first is that it may seem natural to think that whether or not a qualifying condition constitutes an exception to a rule is purely a stylistic matter. It is a question of whether the qualification happens to be included in the statement of the rule, or whether it is left to the decision-maker to articulate on her own. If this is so, exceptions are best thought of as a superficial conceptual category. Whether a qualification to a rule is an exception is a matter of whether it is feasible to include the qualification in the statement of the rule. Since any rule is subject to many circumstances in which the rule will fail to apply, every rule will have a number of unarticulated

[2] Some theologians and even some philosophers have maintained that certain things a person can do are *so* bad that there are *never* any circumstances under which he may do them. See generally G.E.M. Anscombe, "Modern Moral Philosophy", in *Ethics, Religion and Politics, Collected Papers*, (1981), Vol. III, p. 26 . But this view appears to have secured few adherents, even among those who subscribe to a broadly non-consequentialist ethics.

[3] Williams, *supra* n.1; Glanville Williams, "Offences and Defences", (1982) 2 *J. Leg. Stud.* 233; Paul Robinson, "Criminal Law Defences: A Systematic Analysis", (1982) 82 *Colum. L. Rev.* 199, 204–29.

[4] Joseph Raz, *Practical Reason and Norms*, 2nd edn. (1990), pp. 187–88; Ronald Dworkin, "The Model of Rules I", in *Taking Rights Seriously* (1977), pp. 14, 24–5; Ronald Dworkin, "The Model of Rules II", in *Taking Rights Seriously* (1977), pp. 46, 71–78; Joseph Raz, "Legal Principles and the Limits of Law", (1972) 81 *Yale L. J.* 823, 829–34 ; David Lyons, *Forms and Limits of Utilitarianism* (1965), p. 122; H.L.A. Hart, *The Concept of Law* (1961), pp. 130, 136.

[5] Frederick Schauer, "Exceptions", (1991) 58 *U. Chi. L. Rev.* 871, 872 .

[6] He does mysteriously say that it rests "on a confused notion of the logical status of an exception", without explaining how the confusion has produced this result. *Ibid.*

conditions that constitute exceptions to it.

The second is connected with the first, although it stands in some apparent tension with it. This is the idea that the myriad qualifications that apply to a rule without being included in a statement of it do not in fact fall outside the scope of the rule. They are an *implicit* part of the rule, despite their formal exclusion from its statement. For the ones excluded lurk within the rule anyway; one has only to read between the lines to see them. This view seems connected with a general tendency to exaggerate the importance of interpretation in rule-following, and to suppose that *any* problem about the applicability of a legal rule can be solved by interpreting it correctly.[7] On this way of looking at the matter, the rule has swallowed the exception, for if all of a rule's qualifications are an implicit part of the rule itself, then there is no qualification outside a rule that could count as a true exception to it.[8] So once again, the topic of exceptions seems a confusion. Exceptions may exist as a superficial linguistic matter, but they disappear once the true structure of rules of prohibition is understood.

I shall argue that to think of exceptions as superficial products of a linguistic or stylistic decision, or to eliminate them by thinking of them as *already* implicit in the rules they qualify, is to misconceive the nature of rules of prohibition. For I shall argue that the logic of prohibition itself suggests the treatment of certain sorts of conditions as *external* to the rules they qualify. Thus even if it were possible to include in the statement of a rule all the conditions that could make the rule fail to be dispositive in a given case, it would not be desirable to do so. It is not just that it is less cumbersome to formulate rules succinctly, without a long series of "unless" clauses appended to the end. I wish to claim that it is in fact more *accurate* to omit certain kinds of conditions from the statement of a rule of prohibition. For omitting such conditions underscores their identity as exceptions, an identity that is not itself *determined* by whether they happen to be included in the statement of the rule.

I shall argue for a view of the relation between rules of prohibition and the conditions which qualify them that, contrary to two other views we will consider, preserves the structural independence of certain qualifying conditions from the rules of prohibition they defeat. Having suggested a way of thinking about exceptions generally, I shall then turn to the criminal law problem to see whether the suggested position in the larger jurisprudential debate illuminates the relation between offences and defences. Most criminal law scholars who have considered the question regard defences as implicit negative elements of the offences they qualify, and hence, like their jurisprudential counterparts, the criminal law scholars are inclined to regard the

[7] See Richard A. Posner, *The Problems of Jurisprudence* (1990), p. 299 (suggesting translating interpretive questions "into questions about consequences").

[8] Dworkin is the most prominent exponent of this position. See, e.g., Dworkin, "Model I", *supra* n. 4, at pp. 24–5.

question of how to structure offence definitions as largely a stylistic matter.[9] The Supreme Court has also endorsed this view of the line between offences and defences, by allowing the definition of an offence to be determined solely by state legislative pronouncement.[10] This positivistic approach has the effect, among other things, of rendering the constitutional presumption of innocence largely vacuous. Our jurisprudential discussion of exceptions will suggest a non-arbitrary way of drawing the line between offences and defences, and hence a way of preserving constitutional guarantees organized around the notion of a criminal offence.

II. Three Views of Exceptions

Let us begin by clarifying what we mean when we speak of an exception to a rule. An exception is a qualification of a rule that stands in a certain relation to it, namely it stands *outside* the rule it qualifies. Thus a qualification included in a statement of the rule is not properly speaking an exception to it. Consider, for example, the various qualifications that apply to a sign that reads "Do not enter unless authorized personnel". One of the rule's qualifications appears within the scope of the rule, that expressed in the "unless" clause. The qualification regarding "authorized personnel" is thus not properly speaking an exception to the rule.[11] By contrast, qualifications like "unless someone is having a heart attack inside the door and you are a doctor", or "unless someone who is herself 'authorized personnel' has invited you to enter" fall outside the rule, and the term "exception" is correctly applied in their case.

What I shall call the "interpretivist" view of exceptions suggests that all exceptions to a rule can be determined by interpreting the rule correctly. This view is most prominently espoused by Dworkin in his early essay "The Model of Rules I". In that essay, Dworkin famously argues that rules are applicable "in an all-or-nothing fashion".[12] If a rule applies, and if it is valid, the rule dictates the answer to the case. If the rule applies but turns out *not* to be dispositive of the outcome, the rule is not valid, in which case Dworkin says the rule "contributes nothing to the decision".[13] The notion of an *exception* to a rule is precluded by the understanding of rules Dworkin suggests. For if a

[9] See generally Williams, "Logic of Exceptions", *supra* n. 1.

[10] *Patterson* v. *New York*, 432 U.S. 197 (1977).

[11] This point is consistently overlooked in the small corpus of writings about exceptions. H.L.A. Hart, for example, says: "It does not follow from the fact that such rules have exceptions incapable of exhaustive statement, that in every situation we are left to our discretion and are never bound to keep a promise. A rule that ends with the word 'unless. . .' is still a rule". Hart, *supra* n. 4, at p. 136.

[12] Dworkin, "Model I", *supra* n. 4, at p. 24.

[13] *Ibid.*

rule could have exceptions, then the rule could be valid and could apply to a case at the same time that it would fail to dictate its outcome. In order to block this possibility, Dworkin unsurprisingly says that "an accurate statement of the rule would take this exception into account, and any that did not would be incomplete".[14] And although it would often be "too clumsy to repeat [the exceptions] each time the rule is cited",[15] Dworkin thinks there is no conceptual objection to stating a rule in a way that includes all of the conditions that would make the rule non-dispositive. The more of a rule's qualifications included in the statement of the rule, the more accurate that statement is.

It is famously Dworkin's point to argue that matters are otherwise where principles are concerned. For unlike with rules, there are many instances in which valid principles will not dictate the outcome of cases to which they apply. Presumably we should not describe these as cases in which an *exception* defeats the principle, since Dworkin says that the conditions that qualify a principle "are not, even in theory, subject to enumeration".[16] It is reasonable to suppose that a qualifying condition can only constitute an exception if the thing it qualifies would be dispositive of the outcome in the absence of the qualification in a case to which it applies.

On Dworkin's account, then, there is no room for talk of exceptions anywhere in a legal system. For there is no such thing as an exception to a rule, since any articulated exception is only a more precise statement of the rule. The point is that a rule cannot *fail* to dictate the outcome of a case where it applies and still be a *rule*. A principle, on the other hand, might fail to dispose of a given case to which it applies, but we cannot speak of an "exception" here either, since a principle might fail to dispose of a case even in the absence of the qualifying condition. There is no such thing as an exception to something that fails to dictate the outcome of particular cases.

Let us now turn to an apparently quite different position on exceptions, that presented by Fred Schauer.[17] For Schauer, a case in which an exception prevails is one in which a decision-maker recognizes that applying the rule would not further the purpose the rule itself was originally designed to promote. He therefore thinks there is no logical or conceptual difference between saying that something falls within an exception to a rule and saying that the "central principle" that stands behind a rule does not warrant extending the rule to the instant case. If a judge were to admit an exception to a rule which by its terms applied, the judge would be *altering* the rule in light of the rule's background justification:

> "[I]f arguments about exceptions are in reality arguments about the rule itself, then in many other contexts it is important to resist the idea that exceptions exist apart

[14] *Ibid.* at p. 25. [15] *Ibid.* [16] *Ibid.*

[17] Schauer, *supra* n. 9; see also Frederick Schauer, *Playing By the Rules: A Philosophical Examination of Rule-Based Decision-Making in Law and in Life* (1991), pp. 115–16.

> from rules, and, consequently, that *adding an exception is anything other than changing the rule* . . . Now that we know that exceptions are continuous with the rules they are exceptions to, however contingent that continuity may be, we can see that there is no difference between adding an exception to a rule and simply changing it".[18]

To allow an exception to a rule is thus to *change* the rule and to fashion a new rule which includes the relevant qualifying condition within its statement.[19] For Schauer, the only legitimate reason to change a rule stems from the rule's own background justification. Not every case in which the original purpose of the rule would fail to be furthered by applying it, however, is one in which there is sufficient warrant for changing the rule. Otherwise, Schauer correctly suggests, the rule would lose all normative force, since it would only apply in those cases in which the result reached by applying the rule could be reached directly from the rule's background justification. Schauer concludes that rules ought to be modified only when their application in a particular case would produce a result that is "so far out of bounds, so absurd, so preposterous that it is analogous to an abuse of discretion and would therefore be reversed".[20] Short of such extreme results, the presumption in favour of following the (original) rule should lead to adherence to its letter. The fact that a rule's purpose would not be furthered in a particular case is not by itself sufficient warrant for modifying the rule on a case-by-case basis.

Schauer also suggests, however, that we should distinguish a case in which a decision-maker admits a *new* exception to a rule from one in which she merely identifies a qualification of some articulated rule that has previously been accepted as part of the rule. He says:

> "[T]he issue is not whether rules may have exceptions and still be rules, for of course they may. It is whether rules may be subject to exceptions added at the moment of application in light of the full range of otherwise applicable factors and still be rules, and the answer to that question is "no".[21]

Schauer's thought seems to be that *adding* an exception at the moment of application is tantamount to changing a rule, but that *recognizing* a previously

[18] Schauer, "Exceptions", *supra* n. 5, at 893 (italics added).

[19] Dworkin sometimes sounds as though to recognize an exception to a rule by way of interpretation is in fact to change the rule. For example, he says that a judge may be licenced to change an existing rule of law when he finds that it would advance a binding principle to do so: Dworkin, "Model I", *supra* n. 4, at p. 37. But it makes little sense to say that a certain conclusion is reached *both* by drawing out the meaning of a rule through interpretation *and* by changing that same rule. The former presupposes that the rule already contains the conclusion within it, when the rule is applied to the given case. The latter suggests that the conclusion can only be reached by diverging from the meaning of the rule as it presently stands. In light of Dworkin's overall commitment to interpretation, we should probably ignore his remarks about changing a rule in this context.

[20] Frederick Schauer, "Formalism", (1988) 97 *Yale L. J.* 509, 547.

[21] Schauer, *Playing by the Rules*, *supra* n. 17, at p. 116.

articulated exception is not. But how can a judge recognize an existing exception to a rule on Schauer's account without changing the rule? To recognize an exception is to reach *outside* the rule to dispose of a case. This, however, is exactly what Schauer thinks constitutes changing a rule. The distinction Schauer *means* to draw is that between recognizing an existing qualification and adding a new one. The former suggests that the qualification is already a part of the rule itself, and if this is the case, it is *not* the recognition of an exception. The latter *is* potentially the recognition of an exception, but only if it does *not* involve changing the rule. For to recognize an exception to a rule is to preserve the rule, not to change it.

Schauer, therefore, has no basis for distinguishing the recognition of an existing exception from admitting a new one. Both require allowing a qualification that falls outside the scope of a rule to dispose of a case, even though there is a rule which applies to the case and which would dispose of it differently. For Schauer, both constitute changing an existing rule. What Schauer *can* distinguish is recognizing an existing qualification from adding a new one. But neither is a case of admitting an exception. The former is to treat an "exception" as an implicit part of a rule, as Dworkin does, and so it is no exception at all. The latter is simply to change the rule, and thus again, no exception is admitted. This is why, despite the differences between their positions, Schauer can be charged with Dworkin's mistake of squeezing exceptions out of existence.[22]

Now consider a third position on the nature of exceptions. If we reject the view of rules as all-or-nothing, we can think of rules as sharing certain characteristics with principles. In particular, we might think of rules as having *weight*, and thus allow for the possibility that rules, like principles, can conflict. Joseph Raz argues along these lines in his attack on Dworkin's distinction between rules and principles.[23] Raz suggests that rules, like principles, can be balanced by a decision-maker and treated as dispositive of a case only when they outweigh other rules with which they conflict.[24]

Strikingly, Raz attempts to demonstrate his point by offering an example drawn from the criminal law: the relation between the prohibition on assault and the qualification that allows for self-defence. Without articulating the point as such, Raz effectively argues that a statement of the law of assault should be thought of as complete without including the self-defence qualification within the scope of the prohibition. He suggests that we think of the rule

[22] Curiously, Schauer himself embraces this result, despite the focus on the topic of exceptions his work introduces: "Probing [the logical status of an exception]", he writes, "prompts the realization that there is no logical distinction between exceptions and what they are exceptions to, their occurrence resulting from the often fortuitous circumstance that the language available to circumscribe a legal rule or principle is broader than the regulatory goals the rule or principle is designed to further". *Ibid.*

[23] Raz, *Legal Principles*, *supra* note 4, at 830–31.

[24] *Ibid.*

prohibiting assault and that permitting self-defence as conflicting or "interacting", and that it is particularly because rules can "conflict or 'interact' . . . that they can modify and qualify one another".[25]

Raz is on the right track in understanding rules as capable of conflicting, and in explaining the relation between at least some rules of prohibition and some defences in these terms. His therefore strikes me as the correct framework for developing a general logic of exceptions. I shall nevertheless suggest that Raz misunderstands the nature of his own insight in several important respects. First, as we shall see, the particular question of the relation between the criminal law's rules of prohibition and their associated defences is a far more complicated one than Raz supposes. For it turns out that sometimes an offence should be understood as related to its associated defence in terms of a conflict of rules and sometimes not. We shall see why this is so in our exploration of the criminal law problem in the next part.

Second, Raz's particular interest is in the problem of the individuation of *laws*. His concern is to argue that the law prohibiting assault and the law permitting self-defence should be thought of as two laws rather than one.[26] The reason he offers for this position is, as he says, that "we should adopt a doctrine of individuation which keeps laws to a manageable size, avoids repetition, minimizes the need to refer to a great variety of statutes and cases as the sources of a single law, and does not deviate unnecessarily from the (admittedly hazy) common sense notion of a law".[27] And he further stresses that the approach which classifies rules of prohibition and their defences as separate laws "is closer to the way lawyers ordinarily think about the law".[28] It is not necessary, however, for Raz to make any claims about the nature of *law* in this context. For the real debate between Dworkin and Raz is about whether *rules* can conflict, and there may or may not be a one-to-one relation between a rule and a law. It might indeed turn out that the correct theory of law makes assault and self-defence two parts of a single law. But that would not resolve the question of whether their relation should be thought of in terms of a conflict of rules, and thus whether self-defence should be thought of as an *exception* to the rule prohibiting assault.[29]

[25] *Ibid.* at 832.

[26] Raz also thinks the example demonstrates a certain difference between the way rules conflict and the way principles conflict. He writes: "Conflicts between rules are determined solely by their relative importance; conflicts between principles are determined by assessing their relative importance together with the consequences for their goals of various courses of action". We will not attend to this suggestion here, however. *Ibid.* at 833.

[27] *Ibid.* at 832.

[28] *Ibid.*

[29] It is puzzling that Raz chooses to put his point against Dworkin in this way, especially in view of the reasons he offers for his position on the individuation of laws. For Dworkin could have responded to Raz's claim that it would be better from a practical standpoint to individuate laws finely, rather than coarsely, by saying that manifestly it is *not* better to think of laws that way, for it is by doing so that we encounter the problem of conflict of laws. And to the point that lawyers ordinarily tend to think of laws this way, Dworkin could have responded that this is surely irrelevant, since ordinary lawyers are not by and large theoreticians, and hence their intuitions are not a reliable guide to the jurisprudential foundations of the legal vocabulary they use. Dworkin, however, does not give these answers in his remarks responding to Raz Dworkin, "Model II", *supra* n. 4, at pp. 71–78.

Let us stick, then, to Raz's suggestion that *rules* can conflict, and leave his ideas about the individuation of laws to one side. Not only, it would seem, can rules conflict with one another, but a rule can conflict with a principle, and in particular, with a principle other than the one that constitutes the rule's own background justification. Once we accept this understanding of rules, we have a natural way of thinking of exceptions, one that preserves the exception as structurally independent from the rules of prohibition it qualifies. In light of the possibility of conflicts of the aforementioned sorts, a valid rule which applies to a given case might stand in need of no revision, and still not be dispositive of the outcome in that case. Only then would the condition that qualifies the rule constitute a true exception to it. For an exception to a rule arises when an applicable rule cannot dispose of a case, because the outcome of that case must be determined in accordance with some *other* rule or principle instead.

We are now in a better position to understand where Schauer's account goes wrong. Schauer incorrectly assimilates recognizing an exception to modifying a rule in light of its own background justification. The reason he thinks recognizing an exception to a rule is changing the rule is that he thinks a rule can only fail to be dispositive if applying it in a particular case would defeat the rule's own background justification. He is focused on cases in which the failure of the rule is what he calls "internal".[30] He is certainly correct that *if* a rule's failure were of the internal variety, recognizing an exception to it for that reason *would* be tantamount to changing the rule. For in that case one would be returning to the justification in lieu of following the rule, and this constitutes a recognition that the rule was not correctly formulated in the first place and that it stands in need of modification. To allow that a rule could fail to be dispositive in a case to which it otherwise applies, when the reason for the failure is that following the rule would not promote the purpose behind the rule's initial adoption, suggests that the rule's purpose would have been better served by a slightly different rule all along.

The logic of exceptions, however, is more correctly understood in terms of what Schauer calls "external" failure, namely conflict between a rule and something other than the rule's own background justification. If we focus on external, rather than internal, failure, we can see why recognizing an exception is importantly different from altering a rule. An exception arises when there is an implicit conflict of principles: the principle which finds expression in the first rule conflicts with a second principle which is entirely separate from the rule. The conflict can manifest itself as one between two rules or as

[30] Schauer, *Playing By The Rules*, *supra* n. 17, at p. 117.

one between a rule and a principle. We might agree with Dworkin that the notion of an exception has no place where a conflict of principles is at issue, for the reason he gives, namely that principles do not aspire to conclusive application. And so we can say that an exception arises in a case involving a conflict between a rule, on the one hand, and something else, on the other, where that something else can be either another rule or a principle.

Schauer is certainly correct that a rule would do no work if every time the application of the rule failed to further the rule's own background justification the decision-maker decided in accordance with the justification rather than with the rule. So the rule cannot be respected *qua* rule if it is to be "followed" only where doing so would promote some background purpose. But it is consistent with Schauer's claims about respect for rules to recognize a sharp distinction between internal and external failure. For unlike internal failure, where the failure is external, the decision-maker can decide not to follow the rule and still regard herself as bound by it. This is because her commitment to follow rules, as well as principles, can place her in a position in which to follow one rule would be to violate another rule or to reject the force of a principle by which she is bound. Ironically, then, *following* a rule may require a decision-maker to ignore the rule in cases in which it applies. Deciding in accordance with an exception rather than with an applicable rule sometimes reflects a recognition of the weight or importance of a contrary rule or principle; it need not be a rejection of the rule to which it is an exception.

Once we take the fact that rules can conflict into account, along with the fact that different rules can have different purposes, Schauer's claim that to fashion an exception to a rule is to change the rule becomes quite difficult to defend. It would involve not only the plausible requirement that rules have a high degree of resistance with respect to their own background purposes, but the rather less plausible requirement that they have a high degree of resistance with respect to *another rule's purposes* as well. Schauer is partially aware of the distinction, for he passingly mentions the importance of external resistance: "[F]or a rule to be a reason for action it must also . . . have some degree of resistance to external defeasibility. A reason with no resistance to any other reason is no reason at all".[31] But while the suggestion would rescue his account of exceptions, it seems incorrect. A rule that supplies a conclusive reason for acting just in case there is no *other* reason that weighs against it is still a rule. While such a rule would be a *weak* one, it seems wrong to say it would not be a rule at all. It would determine the outcome of a case conclusively where no contrary reason weighed against it. Since the assumption Schauer would need to defend his account of exceptions thus appears to be flawed, his account of exceptions is unacceptable.

Granted, this criticism of Schauer does depend on one important and

[31] Schauer, "Exceptions", *supra* n. 5, at 118.

perhaps controversial claim: the idea that different rules *can* reflect different background purposes. As Schauer himself notes, the possibility of external failure disappears in utilitarian or other "single-valued" systems.[32] For where there is only one purpose that any legitimate rule can have, a rule could not conflict with any purpose other than its own. All cases of failure would be internal. The notion of an *exception* thus depends on the existence of multiple values in a system, since it requires a conflict among values. It should not be surprising, then, to find Schauer extolling the virtues of the commensurability of values in another context, despite the fact that he purports to subscribe to "the ontological correctness of incommensurability".[33] The enthusiasm for commensurability is consistent with the focus on internal failure and the tendency to ignore external failure. But given Schauer's apparent willingness to admit the "ontological correctness" of multiple values, he should also embrace the possibility of external failure, and this would allow him to accept the understanding of exceptions I have proposed.

It should be noted, however, that one need not defend incommensurability down to the ground in order to recognize the possibility of multiple values. Even a single-valued system can have multiple values for practical purposes, since epistemic limitations may make it impossible to plot each separate value on the metric of value that underlies the system. While the theory of exceptions I have suggested depends on at least the practical importance of multiple values in a system of rules of prohibition, this does not seem to constitute a weakness, even if all apparent systems of value are ultimately commensurable with one another.

III. Offences and Defences

Let us now consider the implications of the account of exceptions suggested in the previous part for the problem of the relation between offences and defences. Raz's idea that rules can conflict gives us a way of formulating the criminal law problem. Thinking of offence definitions and defence provisions as separate rules leads naturally to thinking of defences in the way we thought of exceptions in the previous part, namely as qualifications that stand outside the rules of prohibition they qualify. On this view, the criminal law's offences would be complete without inclusion of their associated defence provisions, and the latter would be external to the offence definition. The suggestion that rules can conflict thus gives us a way of capturing the criminal law notion of an *affirmative defence*. On the other hand, thinking of

[32] *Ibid.* at n. 8.

[33] Frederick Schauer, "Commensurability and Its Constitutional Consequences", (1994) 45 *Hastings L.J.* 785, 806 ("fostering a public belief in commensurability . . . , despite the ontological correctness of incommensurability, might also produce in the aggregate better . . . results".).

the relation between offence definitions and defence provisions as part of one large rule would lead naturally to thinking of defences as integral to the rules they qualify. On this way of looking at the matter, a complete statement of an offence would require mention of the absence of its various defences. This would in turn be a way of capturing the notion of a *negative offence element*. The question of whether we should think of a rule and its associated qualifications in terms of a conflict of rules thus maps nicely onto the question of whether to think of defence provisions as affirmative defences or as negative offence elements.

Raz suggests we think of the relation between assault and self-defence in terms of a conflict of rules, and hence, in effect, that we think of self-defence as an affirmative defence. Indeed, he suggests that we might think of *any* qualification of a criminal prohibition in terms of a conflict of rules. He thinks, for example, that the defence of mistake of fact would be appropriately analysed this way.[34] Raz's implicit suggestion is thus that we regard all of the criminal law's defences as affirmative defences. Glanville Williams, by contrast, who represents the prevailing view among criminal law theorists, favors the second way of conceiving the relation between offences and defences. Williams writes: "Rationally regarded, an 'exception' merely states the limits of an offence. A person whose act falls within the exception does not commit the offence".[35] Williams thinks, in other words, that a defence to a criminal provision is just an implicit part of the offence definition itself. Whether set forth in a separate rule-like statement or not, each defence should be thought of as part of the definition of each offence to which it applies. Williams thus appears to think of all of the criminal law's defences as negative offence elements, whether or not explicitly included in the statement of the defence.

What is curious about the above disagreement is that the criminal law does not choose up sides in the way that Raz and Williams seem to think. Although self-defence *is* normally articulated as a separate, affirmative defence, a defence like mistake of fact is generally understood as a negative offence element, since it defeats a criminal offence only if it 'negatives' an element of the offence definition.[36] That is, a defence like mistake can usually be derived from the various offence definitions to which it applies, since a defendant will have a defence of mistake whenever his mistake results in his failure to satisfy the *mens rea* required for the offence. For example, the crime of theft might be defined as "tak[ing], or exercis[ing]

[34] Raz, "Legal Principles", *supra* n. 4, at 832.

[35] Williams, "Logic", *supra* n. 1, at 262.

[36] In particular, it negatives the mental element. See Model Penal Code, section 2.04(1). As the drafters of the Model Penal Code realized, mistake of law can also function in this way, that is, a mistake of law sometimes negatives an element of the offence. This could arise, for example, when the offence requires that the defendant have done something "unlawfully", and the *mens rea* of 'knowingly' is taken to apply to that element. See *Ratzlaf* v. *United States*, 510 U.S. 135 (1994).

unlawful control over . . . movable property of another with purpose to deprive him thereof".[37] It follows that one is not guilty of theft if one takes movable property under a mistaken view of its ownership, thinking it is one's own when it is not. The existence of a defence of mistake is an unavoidable product of the fact that the right kind of mistake negatives the *mens rea* requirement for a person taking property to be guilty of the offence. One need supply no principle external to the offence definition for the defence to become available.

Granted, mistake is still sometimes set forth in a separate defence provision.[38] And on the other side, a defence like self-defence is *not* systematically articulated in a separate provision.[39] But the vicissitudes of state legislative efforts do not alter the fact that it is *unnecessary* to provide separately for a mistake defence, for the defence is already implicit in any offence definition that contains a mental state requirement.[40] Conversely, we must understand jurisdictions that lack a statutory self-defence provision as operating on the basis of a separate, judge-made self-defence rule nevertheless, since the permissibility of killing or assaulting in self-defence cannot normally be derived from the offence definitions that the permission to act in self-defence qualifies.[41] It is not surprising that legislative distinctions fail to track the conceptual distinctions of criminal prohibition, since legislatures operate in the absence of any clear theory of the relation between offences and defences. For this reason, we cannot derive much guidance from existing drafting decisions, in our efforts to construct a theory of offence definition.

The question is thus not whether, *pace* Raz, we should think of all defences as affirmative defences, or whether, *pace* Williams, we should think of them as negative offence elements. The question, rather, is *when* we should think of a defence one way and *when* we should think of it the other. For sometimes it is correct to think of the relation between offences and defences in terms of a conflict of rules and sometimes not. Sometimes a qualification stands outside a rule, constituting an "exception" to it, and sometimes it is part of the rule it qualifies, either implicitly or explicitly. Can we identify a general test for determining when we should think of the relation between rule and qualification in terms of affirmative defence and when we should think of it in terms of negative offence element?

[37] MPC, section 223.2(1).

[38] See MPC, section 2.04.

[39] California, for example, has no independent self-defence provision. The permissibility of acting in self-defence must be inferred from the substantive provisions that would apply were self-defence impermissible, in particular from the basic homicide provision. Cal. Crim. Code, section 187.

[40] Not all mistakes of law, however, is based on lack of *mens rea*. Sometimes, as set out in the MPC, section 2.04(3), the defence may be based on inadequate publication of the prohibition, rather than on negativing an element of the offence definition.

[41] A possible exception to this is the sort of homicide statute that includes the term "unlawfully" in the offence definition. I discuss this below.

Let us begin by considering what turns on the choice between these two approaches. As we shall see, under prevailing practices there are three characteristics commonly associated with each way of structuring the offence/defence relation. None of these characteristics is a *necessary* concomitant of the choice between affirmative defence and negative offence element. But they are typical features of each approach, and it will be convenient for us to think of the choice between affirmative defence and negative offence element in terms of them.

The first and most important implication of the choice between these two ways of relating offences and defences stems from the constitutional requirement that the prosecution prove every *element of an offence* beyond a reasonable doubt.[42] This means that if a defence provision is a negative offence element, the state's constitutional requirement will extend to it, and the prosecution must prove it, along with the other definitional elements of the offence, beyond a reasonable doubt. If, on the other hand, a qualification counts as an affirmative defence, the state is constitutionally permitted to require the defendant to bear both the burden of production and the burden of proof with respect to that defence. In support of this apparently harsh possibility, it is sometimes suggested that states are under no constitutional obligation to extend affirmative defences to potential defendants anyway, and if a state can abolish a defence altogether, it can surely place a heavy burden on the defendant to prove he fell within the privilege permitted.[43] It is for this reason that the designation of a qualification as an affirmative defence has significant constitutional consequences with respect to burdens of proof and the presumption of innocence.

A common hybrid approach shies away from this extreme result, allocating only the burden of production for defenses to the defendant, and leaving the burden of proof on the prosecution.[44] The result is that the burden of disproving a defence does not arise until the defendant has introduced some initial evidence regarding the defence he wishes to assert. The Model Penal Code presents a somewhat unusual version of this approach. In typical fashion, it puts the burden of production for affirmative defences on the defendant. But it makes the *negative* of every justification and excuse an element of every offence, thus adopting a negative offence element approach to all such defences.[45] This has the result that the prosecution bears the burden of proof with respect to the *absence* of any justification or excuse for

[42] This may apply both as to the burden of production, *Jackson* v. *Virginia*, 443 U.S. 307, 314–20 (1979), and as to the burden of proof, *In re Winship*, 397 U.S. 358 (1970).

[43] See Joshua Dressler, *Understanding Criminal Law* (1995), section 7.03[B]. It is an interesting question whether a state really could abolish a defence like self-defence entirely. The reasoning in *Patterson* suggests it probably could.

[44] See *Ibid. at* section 7.03.

[45] MPC, section 1.13 (9)(iii)(c). The provision applies only to "justifications" and "excuses", leaving other possible defences outside the offence definition.

which the defendant has supplied sufficient initial evidence.[46] Unlike the Model Penal Code, some states leave the burden of proof for affirmative defences on the defendant, but in this case they almost always require that the defendant prove his defence by the lower standard of preponderance of the evidence.[47]

What is important for our purposes is to see how arbitrary the resulting constitutional doctrine is under these various schemes. If a qualification to a criminal offence falls under the designation of "affirmative defence", the state is entitled to allocate both the burden of production and the burden of proof to the defendant, and to do so at the higher "proof beyond a reasonable doubt" standard. The solution of returning the burden of proof to the prosecution once sufficient evidence has been produced by the defendant is thus not constitutionally mandated. Notice, however, that the MPC's negative offence element approach to justifications and excuses suggests that prosecutors operating under the Code would be in a somewhat different position: In theory, they *would* have an obligation to disprove justifications and excuses once entered into evidence, and thus the hybrid solution would be constitutionally mandated in virtue of the Code's negative offence element provision.[48] The Code's own constitutional burden would be alleviated by eliminating this provision and signing on to the usual treatment of justifications and excuses as affirmative defences. And this seems to make it rather too easy for a state legislature to slip out from under its constitutional burden with respect to a given defence.

Obviously what is needed if the constitutional guarantee is to have any consistent content is a theory of the line between offences and defences on which to base judicial interpretation of criminal provisions as well as legislative drafting decisions. But the Supreme Court has enthusiastically embraced the absence of any such general theory, making the choice between affirmative defence and negative offence element a positivistic product of state legislative activity.[49] It accordingly accepts that the requirement of proof beyond a reasonable doubt simply does not apply to any provision the legislature chooses to draft as an affirmative defence. The Model Penal Code is nominally in agreement, insofar as it provides that whether a defence is affirmative depends on whether the Code or other statute so designates it.[50]

[46] There are exceptions to this structure, such as the defence of entrapment, which the defendant must prove by a preponderance of the evidence. MPC, section 2.13. It should also be noted that the Code does not say how *much* initial evidence is required to shift the burden to the prosecution to disprove the defence.

[47] Dressler, *supra* n. 43, at section 7.03[D]2.

[48] MPC, section 1.13(9)(c).

[49] *Patterson* v. *New York*, 432 U.S. 197 (1977) (upholding New York law shifting burden of proof by preponderance of the evidence to defendant for defence of extreme emotional disturbance).

[50] MPC, section 1.12(3). But see *infra* n. 52.

Now there is a test for identifying affirmative defences courts have traditionally used, namely the "peculiar knowledge" rule, which makes something an affirmative defence if it lies "peculiarly within the knowledge of the defendant".[51] While it is important for courts to name a non-legislative approach to the problem of affirmative defences, the peculiar knowledge rule is not the right place to look for a definition.[52] As Glanville Williams has argued, the rule is at best a confusion and at worst pernicious, since one might as well say that the defendant's mental state at the time the offence was committed is a matter peculiarly within the defendant's knowledge, thus removing the burden of proving *mens rea* from the prosecution altogether.[53] Better than the peculiar knowledge rule, at any rate, would be a substantive rule of constitutional stature. For the real problem with the purely legislative approach to identifying affirmative defences is that it makes the "protection" afforded by the constitutional presumption of innocence a thin one, since a state legislature can remove the burden of proof from its prosecutors with respect to any of the elements of an offence simply by engaging in an exercise of creative drafting.

While the most important implication of the distinction between offences and defences is that having to do with burdens of proof, there are other implications of the choice that deserve mention. One stems from the rule that the *mens rea* requirement of an offence applies to every material element of the offence. A defence provision incorporated into an offence as a negative offence element will almost always count as a material element.[54] Thus the choice of whether to treat a defence as an affirmative defence or as a negative offence element will have implications for the mental state required for the defence provision: If the defence is treated as an affirmative defence, the *mens rea* for the offence will not govern it, and the mental state for the affirmative defence will have to be separately established. But if the defence is treated as a negative offence element, the *mens rea* for the offence

[51] See, e.g., MPC, section 1.12(3)(c).

[52] The designation "affirmative defence", however, has less significance under the MPC than it would in another code, given the negative offence element approach to justifications and excuses in section 1.13(9)(c). What calling something an "affirmative defence" does under the Code is to place the defence within section 1.12(2)(a), which says that the prosecution need not disprove affirmative defences unless "there is evidence supporting such defence". This means simply that the prosecution is under no burden to disprove an affirmative defence with respect to which the defence has not met its burden of production. It is a question whether under prevailing Supreme Court precedent, it is acceptable for the Code to place the burden of production with respect to a negative offence element on the defendant, which is what the MPC has effectively done. There would, of course, be no difficulty in the absence of section 1.13(9)(c).

[53] Williams, "Logic", *supra* n. 1, at 267 *et seq.*

[54] That is, it will be a material element as long as it is connected with the harm or evil it was the purpose of the offence to prevent. This will exclude provisions having to do with statutes of limitation and the like. See MPC, section 1.13.

will *normally* provide the mental state for the defence as well.[55] For example, if a legislature chooses to treat self-defence as a negative element of an assault provision, the *mens rea* for the crime would normally apply to the negative element "not acting in self-defence". If the statute defined assault as "intentionally or knowingly attacking a person not in self-defence", the mental state requirement of "intentionally or knowingly" would apply to the negative element "not in self-defence", with the result that the prosecution would have to prove the defendant knew he was not acting in self-defence. The hybrid solution to burdens of proof aside, the prosecution would thus have to prove this as part of the *prima facie* case. Of course this just means that the defendant would not be guilty of assault if he thought he *was* acting in self-defence. If, on the other hand, self-defence were treated as an affirmative defence, the prosecution would not have to show that the defendant knew he was not acting in self-defence. In that case, if the defendant thought he was acting in self-defence, the prosecution would still be able to make out its *prima facie* case just by showing that the defendant had knowingly attacked another.

A final point on which the distinction between affirmative defence and negative offence element may matter stems from the traditional requirement that the defendant have had a reasonable belief in the existence of the attendant circumstances where an affirmative defence is concerned. While the Model Penal Code largely does away with this requirement,[56] it has been the prevailing approach both in the United States and in Britain whenever a defence is set forth in a separate provision. As Glanville Williams says, it was an accepted part of the common law that "a mistake in relation to a defence element had to be reasonable, whereas this requirement was not generally imposed upon a mistake as to a definitional element".[57] For example, a defendant who claims she acted in self-defence will normally have to prove not only that she believed it was necessary for her to do so, but also that her belief was a reasonable one.[58] But if self-defence were treated as a negative offence

[55] I say normally because the legislature can always establish a separate *mens rea* for any element of an offence definition. So it is in the absence of a special *mens rea* for the negative offence element that including a defence in the offence definition will result in the prosecution's having to show not only that the defensive circumstances did not obtain, but also that the defendant had whatever mental state the statute generally requires with respect to those circumstances *not obtaining*.

[56] The MPC eliminates the traditional reasonableness requirement from self-defence, making the availability of the defence turn on the defendant's honest belief in the need to use defensive force alone: MPC, section 3.04. However, it re-introduces the rough equivalent of that requirement in a separate provision by removing the defence if the belief is negligently formed for any crime whose *mens rea* requirement is negligence or higher, as well as if the belief is recklessly formed for any crime whose mens rea requirement is recklessness or higher: MPC, section 3.09(1).

[57] Williams, "Logic", *supra* n. 1, at 269.

[58] See, e.g., *People* v. *Goetz*, 497 N.E. 2d 41 (N.Y. 1986); *State* v. *Norman*, 366 S.E. 2d 586 (N.C. 1988).

element, the independent requirement of reasonableness would normally not be imported into the defence, since the mental state for the defence would then be determined by the *mens rea* for the offence.

Let us first consider the reasonableness requirement. If a defence provision is an affirmative defence, there is no impediment to attaching a reasonableness condition to it, as the traditional approach to self-defence indicates. But if a defence is a negative offence element, attaching a reasonableness requirement produces confusion. This became apparent when the House of Lords delved into the problem of mistake as a defence to rape. Courts had previously thought that a defendant's mistake about whether a victim was consenting had to be reasonable if the defendant was to claim mistake as a defence to rape. This was so, despite the fact that the *mens rea* traditionally required for rape was awareness or knowledge, suggesting that the defendant had to *know* the victim was non-consenting if he was to be guilty of rape, since non-consent is typically an element of rape. The inconsistency of these rules finally came to light in the case of *R.* v. *Morgan*,[59] where the Lords noticed that if the defendant honestly thought, no matter how unreasonably, that his victim was consenting, he could not have the *mens rea* of knowledge required for rape. It followed "inexorably", as the Lords said, that a defendant who formed an unreasonable belief that the victim was consenting must be able to claim a defence. This permits the following two observations: first, that a defence saddled with a reasonableness requirement should probably be an affirmative defence,[60] and second, that if a defence provision is structured as a negative offence element, it must not be interpreted in a way that imposes a reasonableness requirement on the defendant in claiming it, unless the statute explicitly establishes such a requirement.

Now turn to the belief component of the reasonable belief requirement for affirmative defences. If a court attaches a belief requirement to the designation of a defence as affirmative, then the choice of whether to treat a defence as an affirmative defence or as a negative offence element can be thought of as having these further implications: a defendant who was *unaware* he was acting in defence of his life would not be able to claim the defence if self-defence were treated as an affirmative defence, but he might still be able to claim it if it were treated as a negative offence element. The defence must fail in the former case if the defendant is unaware he is defending his life when he attacks another. In the latter case, the requirement of "not acting in self-defence" would be an element of the offence, and the defendant could not be convicted if he had been acting in self-defence, whether he knew it or not.[61] If treated as a negative offence element, then, self-defence would relate to an

[59] [1976] A.C. 182.

[60] That is, it is an affirmative defence unless the offence definition itself contains a reasonableness requirement that attaches to the negative offence element.

[61] At least this is the case unless we wish to say that a person cannot be acting "in" self-defence if he is not aware he is doing it. But we shall leave this grammatical complication to one side.

offence the way non-consent relates to rape, namely that if the victim was consenting, there is no rape, regardless of whether or not the defendant knew she was consenting. A defendant who attacks another and thereby unwittingly saves his own life would be like a person who mistakenly thinks a consenting woman is non-consenting. In both cases, there might be attempt liability, but there could be no liability for the substantive offence.

The negative offence element approach appears to allow a defendant to claim the benefit of a defence *ex post*, that is, even if he had not been aware of the need to defend his life at the time he attacked. While the result does indeed make sense in the case of rape—if the victim was consenting, there is no rape, regardless of what the defendant thought—it makes rather less sense in the context of self-defence. A person who attacks or kills maliciously ought not to be able to claim the benefit of a justification for killing if he only later finds out that he *would* have been justified had he acted on other motives under the circumstances.

One commentator strongly disagrees. Paul Robinson thinks a defendant unaware of an available justification should be able to claim it nonetheless.[62] In support of this way of looking at the matter, he argues that criminal prohibitions are designed to identify particular social harms or evils, and that the relevant evil is not described in the absence of the qualifying condition. He suggests that when a person acts in a way that *would* be justified were he aware of the circumstances surrounding his act, the harm or evil that the relevant rule of prohibition was meant to prevent has not in fact occurred, even if the defendant thought it was occurring, and even if the defendant intended that it occur.[63] Robinson, then, ought to think it a *benefit* of the view that folds defences into offences as negative elements that it implies that a defendant need not be aware of a justificatory condition in order to claim it as a defence.[64]

But this argument we have constructed for Robinson strikes me as a *reductio* of the position that the negative of a justification should be thought an element of an offence. For I think it obvious that a defendant should not be able to claim self-defence if defending himself was not his *reason* for violating the criminal prohibition.[65] So if treating justifications as negative elements of the offences they qualify *entails* that a defendant can claim the benefit of a

[62] Robinson, *supra* n. 3.

[63] *Ibid.* at 45.

[64] Curiously, Robinson does not. He insists on taking an affirmative defence approach despite his interest in obtaining this result, one most easily reached on the negative offence element approach: *Ibid.*

[65] I am here ignoring an intermediate position, namely that a defendant must be aware of a justificatory condition in order to claim it by way of defence, but that awareness will suffice. Perhaps it need not have been his reason for acting, as long as he was aware of it. See George Fletcher, "The Right Deed for the Wrong Reason: A Reply to Mr. Robinson", (1975) 23 *U.C.L.A. L. Rev.* 293.

justification of which he was unaware, that ought to provide a basis for thinking of justifications as affirmative defences instead.[66]

We have explored three features that can be associated with the choice between treating a defence provision as an affirmative defence and treating it as a negative offence element. When we consider the various combinations of offences and defences that are prominent in the criminal law, we should now be able to identify clear intuitions about each as to whether the defence should be regarded as an affirmative defence or as a negative offence element. What we find, for example, is that rape and its associated qualification of consent is structurally quite different from assault or murder with their associated qualification of self-defence: There are compelling jurisprudential grounds for treating consent as a negative offence element in relation to rape, and for treating self-defence as an affirmative defence in relation to assault and homicide. The matter is certainly not arbitrary or stylistic, and thus should not be subject to the pragmatic determinations of judicial interpretation or state legislative activity.

In the case of self-defence and assault, it seems natural to treat the defence provision as an affirmative defence. To put the matter intuitively, this is because we think the prosecution's *prima facie* case should consist of having to show that the defendant intentionally attacked the victim. Such an attack is an event of significance to the criminal law. Moreover, the prosecution should not have to prove that the defendant *knew* he was *not* acting in self-defence, at least not in the first instance. For in theory, the prosecution would have to show not only that the defendant knew he was not acting in self-defence, but that he knew he was not acting out of necessity, in defence of others, for purposes of law-enforcement, etc.[67] Or consider the typical murder statute, which makes it a crime intentionally or recklessly to cause the death of another human being.[68] The permissibility of killing in self-defence is normally not incorporated into the offence definition, and this again seems the correct approach. The prosecution can therefore make out a *prima facie* case merely by showing that the defendant caused the death of a human being with the required mental state. The defence must then introduce evidence that the *prima facie* case does not entail guilt

[66] If I am wrong and it turns out that treating justifications as negative elements has no such entailment, we would need to decide whether to treat them this way on other grounds. As it turns out, there *are* other grounds for rejecting the negative element approach to justification, and thus we need not in fact determine whether the negative element approach entails that defendants can claim justifications retrospectively.

[67] Granted, the hybrid solution we considered above goes some distance to solving this difficulty as a practical matter. For on that solution, the prosecution has no duty to disprove an affirmative defence until the defendant has introduced sufficient initial evidence of the defence.

[68] See New York Penal Law section, 125.25 (McKinney 1988). The required recklessness condition is in fact a heightened recklessness requirement. The defendant must have committed the act "under circumstances evincing a depraved indifference to the value of human life". *Ibid.* at section 125.25(2).

because the defendant had an excuse or justification for his conduct.[69] It seems right in this instance that the prosecution should carry no burden with respect to the excuse or justification at the level of the *prima facie* case.

Granted, one type of homicide provision leaves matters less than crystal clear, namely where the term "unlawfully" is included in the offence definition. Does a statute that defines homicide as "unlawfully causing the death of another human being" treat self-defence as a negative offence element or as an affirmative defence? On the one hand, the term seems to incorporate self-defence along with the other justifications into the offence definition, since if a defendant killed in self-defence she did not kill unlawfully, and so the "unlawfully" element of the offence definition is negatived. On the other hand, it is possible that the term should not be taken too seriously in this context, because we must still look to the wording of the separate self-defence provision before we have a basis for exculpation. The Supreme Court *could* have used the former interpretation of such a term to distinguish the statute in *Patterson* v. *New York* from that in the earlier case of *Mullaney* v. *Wilbur*.[70] The homicide statute in *Mullaney* contained an "unlawfully" provision which the *Patterson* homicide statute lacked. One might, then, argue that the heat of passion defence functioned as a negative offence element in *Mullaney*, but that it remained an affirmative defence in *Patterson*, on the grounds that the term "unlawfully" incorporates the defence into the offence definition.[71] But it seems a mistake to allow such a large conceptual question to turn on the possibly uninformed drafting choices of state legislatures. Instead, the term "unlawfully" is probably best understood as simply a reminder that there may be a separate rule or principle that can come into conflict with the offence definition. It does not alter the fact that the *prima facie* case is still established without the prosecution's having addressed the various justifications and excuses that might fall under the term. The inclusion of the term "unlawfully" in the typical homicide statute thus does not appear to falsify the claim that self-defence is most naturally related to homicide and to other offences as an affirmative defence and not as a negative offence element.[72]

Consider, by contrast, the relation between rape and consent. Here treating the qualification as an affirmative defence would lead to curious results: the offence definition would simply be *intercourse with a woman*, meaning that

[69] MPC, section 1.12 makes justifications and excuses affirmative defences, and thus the prosecution does not have to disprove them until the defendant has produced initial evidence of the defence. But because their negation is also part of the definition of each offence, the burden to disprove them must return to the prosecution.

[70] 421 U.S. 684 (1975).

[71] At least one commentator has tried to distinguish the cases on these grounds. See Dressler, *supra* note 43, §7.03[a].

[72] There is another reason to think that statutes should not be drafted this way. Among other things, if the term "unlawfully" is included in the offence definition, *mens rea* will apply to it. Legislatures unwittingly end up writing a mistake of law defence into statutes when they draft them this way. See *Ratzlaf* v. *United States*, 510 U.S. 135 (1994).

the prosecution could establish its *prima facie* case by proving the defendant and the victim had intercourse. Moreover, the defendant would be *prima facie* guilty of rape even if he did not know he was having intercourse with a non-consenting woman (since the offence's *mens rea* would not apply to the defence). A further quite crucial feature is that a *prima facie* case could be established under this scenario even if the woman was *consenting*, since the element of *non-consent* would not be a part of the offence definition. One strong intuition about rape, however, is that there is no *prima facie* case made out from the fact of intercourse alone. Another is that there is no rape unless the defendant has some level of awareness that he is having intercourse with a non-consenting woman. A third is that there is no rape if the victim is in fact *consenting*, even if the defendant is firmly convinced she is not. Compare this last point with the relation between assault and self-defence: we are not normally inclined to say there is no assault if the defendant acted in self-defence. Rather, we tend to think there was an assault, but that it was justified. These three points taken together suggest that the relation between rape and consent is that of offence definition to negative offence element, unlike the relation between assault or homicide and self-defence, which differs on each of these points.

While we have clear intuitions about homicide and self-defence, on the one hand, and rape and consent, on the other, the explanation for these intuitions remains mysterious. What we would ideally like is to be able to identify an underlying logic in the law's approach to demarcating offences from defences. While it is possible that any deep logic we might identify is itself misguided, we would have reason to trust the tradition on this question if the historical pattern we uncover is also one that can be theoretically justified. In the next part I shall argue that the traditional approach to the relation between offences and defences can indeed be justified in terms of the theory of exceptions we developed in Section II.

IV. Exceptions as Conflicts of Principles

I have argued that an exception exists when an applicable rule fails to dispose of a case because another rule or principle that conflicts with it is dispositive instead. I have also argued that this sort of conflict is indirectly a conflict of principles. In the case of a conflict of rules, this is because the rules reflect different background principles. The clash between them thus expresses a clash of principles. Where a rule comes into conflict with a principle, the principle that supports the rule clashes directly with another principle. I have also suggested that it makes no sense to speak of exceptions where two principles conflict, since, as Dworkin has pointed out, it is not in the nature of principles to have dispositive application to *any* case.

The foregoing account suggests that we should treat a qualifying condition

as a part of the rule it qualifies when the principle the qualifying condition expresses is the *same* as that which gave rise to the rule of prohibition. And it suggests that we should treat a qualifying condition as an exception when the condition expresses some broader background principle, which principle is *different* from that which gave rise to the rule of prohibition. In terms of our criminal law problem, this means we should treat a defence provision as a negative offence element when the principle that gives rise to the defence is the *same* as the principle that gives rise to the offence, and we should treat it as an affirmative defence when the defence provision reflects a principle independent from that reflected in the offence definition. This way of looking at the matter imposes a requirement on the formulation of the defence as well as on the rule of prohibition itself, namely that both must be separately justifiable. An offence definition cannot stand unless it can be justified in terms of its own background principle. Nor should an affirmative defence be admitted unless it possesses the structural independence that having its own background justification supplies.

The above theory of offence definition suggests the following about how criminal prohibitions should be structured. A rule of prohibition must be articulated in a way that reveals its embodiment of some harm or evil it is the object of the criminal law to prevent. This requirement stems from several different sources. First, such rules place stringent limitations on the autonomy and freedom of individuals, suggesting a presumption against them in our political morality which must be overcome by the existence of a justification in favour of them. It seems plausible to think that a criminal offence which did not identify a harm or evil as the object of prohibition would not meet this requirement of justification. Second, there is an additional constitutional dimension to the requirement, at least in the case of serious offences, which is separate from that involved in questions about burden of proof. Those subject to the intrusion of a state criminal procedure must have notice of any rule whose violation would subject them to punishment.[73] The reason it is fair to presume general knowledge of the law, at least in the case of criminal prohibitions, is that citizens are placed on notice by the wrongful character of the forbidden act itself. The presumption of knowledge of the law would be unjust if serious crimes were formulated in a way that failed to expose the wrongfulness of the acts they prohibit.[74]

Where a defence is concerned, the background justification that supports the rule may be of various sorts, unlike with rules of prohibition. There are, nevertheless, some broad themes that run across the various defences, and in particular, there appear to be a few consistent principles that help to explain

[73] *Winters* v. *New York*, 333 U.S. 507 (1948).

[74] The criminal law gives substantial recognition to this point by allowing for a defence of mistake of law to regulatory offences which are not self-evidently wrongful, where the prohibition they enact has not been adequately publicized. See, e.g., MPC, section 2.04(3).

the existence of the justification defences as a group. It is plausible to suppose that many such defenses can be understood as reflections of a larger commitment to balancing the permissible interest an agent takes in her own well-being against the legitimate interest a state takes in the collective welfare. Self-defence, I have argued elsewhere, is particularly understandable as a reflection of the state's interest in giving political recognition to private agent-relative permissions.[75] This is a different principle from the one that supports the defence of law-enforcement, for example, even if both defences remain reflections of a single larger background justification.[76]

The requirement that each rule be justifiable in terms of its own background principle suggests *why* the asymmetry in the traditional formulation of a crime like rape with its associated qualification of consent, on the one hand, and a crime like murder, with its associated qualification of self-defence, on the other, is not an *ad hoc* feature of our criminal law, as Glanville Williams and others seem to think. For the harm that justifies the rule of prohibition in the case of rape is *not* one that reflects a harm or evil if non-consent is omitted from the offense definition.[77] *Intercourse with a woman* does not identify a harm or evil, the commission of which establishes a defendant's *prima facie* susceptibility to punishment. An extreme scepticism about the underlying purposes of criminal prohibition would be required to defend the position that an offence definition of this sort could constitute a *prima facie* harm as well as any other. The requirement that an offence definition target a normatively suspect activity implies that non-consent should be included among the definitional elements of rape.

This analysis is supported on the defence side of the equation. For the criminal law in most cases rejects the idea that a harm otherwise criminal could be justified when done with the consent of the victim.[78] Since consent, at least in the criminal law, does not usually constitute a value of its own, a rule extending a defence on the grounds that the victim was consenting would not be supported by any independent principle of polit-

[75] Claire Finkelstein, "On the Obligation of the State to Extend a Right of Self-Defence to its Citizens" (forthcoming, 1999) 147 *Univ. of Penn.L.Rev.*

[76] The same could be said for the various excuses. The defences of insanity and infancy are justified by quite different notions of impairment, and thus they are supported by quite different local principles. Nevertheless, both of the local principles which stand behind the rules for these defences are supported by a more general background justification having to do with the nature of responsibility and the way in which it constitutes a requirement for the application of any criminal sanction.

[77] A common alternative has been to treat force or threat of force as the element that is added to intercourse to constitute rape. But courts have interpreted such statutes as containing an implicit requirement of *non-consent*, thus underscoring the conceptual inevitability of the approach that treats non-consent as an element: *State in the Interest of M.T.S.*, 609 A. 2d 1266 (1922).

[78] This much is clear from the prosecutions for manslaughter of participants in a voluntary game of Russian Roulette for the death of one of the players. See, e.g., *Commonwealth* v. *Atencio*, 189 N.E. 2d 223 (Mass. 1963).

ical morality.[79] This is not the place to enter into a discussion of why this is so, and why the criminal law differs so markedly from other areas of law, like tort law, which place quite a different value on the notion of consent.[80] Suffice it to say that a fuller treatment of the question would explore the largely paternalistic nature of criminal prohibitions, and the overall place that the state occupies in our political system as vindicator of the worth of human life.

Matters are otherwise, however, when we consider the relation between the rule of prohibition that outlaws intentional killing and that which justifies it under the exceptional circumstances in which the killing takes place in defence of one's life. The harm the prohibition seeks to eliminate is clear, and the purpose of the prohibition would not be frustrated by applying it to those who kill in self-defence. Indeed, it is a well-known aspect of the history of murder laws that a person was formerly guilty of that offence even if he killed *se defendendo*.[81] Someone, such as a pacifist, might argue that it would *better* vindicate the worth of human life to punish all who shed human blood, even those who do so in self-defence. The rule of prohibition itself, then, can be meaningfully captured without including self-defence as a negative element. On the defence side, as I have argued, there are reasons for thinking of self-defence as justified by its own background principle of political morality. There are thus reasons on both the offence and the defence sides for distinguishing homicide from rape in this regard, and for treating non-consent as an element of the latter offence at the same time that we treat self-defence as an affirmative defence.

The requirement that offence and defence each be justifiable in terms of its own background principle is quite a general one. It admittedly falls short of the more precise theory a court or legislature would require in order to have

[79] There are of course some exceptions to this. Consent is the principle that allows us to distinguish boxing from assault, in criminal law as well as outside it. But the general position of consent in the criminal law is so markedly different from its position in other areas of the law, such as in torts and contracts, that it is difficult to conceive of it as a significant independent value in the former.

[80] See W. Page Keeton, *et. al.*, Prossner and Keeton on Torts 4th edn. (1984), section 18. But consent may even be better thought of as a negative element in tort law as well: *Ibid.*

[81] See Claire Finkelstein, "Self-Defence as a Rational Excuse", (1996) 57 *U. Pitt. L. Rev.* 621. A possible objection might occur to anyone familiar with the particular aspect of criminal law history to which I refer. For the same treatment was extended to killing *per infortunium*, that is to accidental killing, as to killing *se defendendo*. But it is currently part of the prohibitory norm itself that a defendant must have killed intentionally (or knowingly or recklessly). So by my own argument, it looks as though the requirement of *mens rea* should not itself be part of the formulation of the offence. But matters are quite complicated, and, I think, significantly different where *mens rea* elements are concerned. For imagine what would be required to make the absence of mens rea an affirmative defence. It is not clear the suggestion could be made to work. Moreover, there has arguably been evolution on the question of unintentional harms. They are no longer thought of as the harm or evil which the criminal law seeks to eliminate. It is rather intentional harm that is at issue.

guidance interpreting criminal provisions and drafting new ones. The schema nevertheless gives us a way of approaching such questions, and while the details of the theory remain to be filled in, we can begin to see how it would apply in particular cases.

Suppose a legislature wanted to draft a statute making it illegal to operate a vehicle without a licence. It might do this by defining the relevant offence as "purposely or knowingly operating a vehicle without a licence". The *absence of a licence* would then be a definitional element of the offence, and the presence of a licence would be a negative offence element. Alternatively, the legislature could define the offence as "purposefully or knowingly operating a vehicle", and then allow an affirmative defence in the case in which the defendant had a licence. Under the first approach, the prosecution has the burden of proving that the defendant operated a vehicle without a licence, and moreover that the defendant *knew* he was operating it without a licence. Under the second approach, the prosecution need only prove that the defendant knew he was operating a vehicle. It would then be up to the defendant to show, for example, that he thought he had a licence, and a court might well require him to prove that his belief was reasonable. Moreover, under the first approach, if, without knowing it, he actually *had* a licence, say, because he mistakenly thought his licence had expired, he would not be guilty of the offence because he would not satisfy the negative element of absence of a licence. Under the second approach, however, he would still be guilty of the offence, since the fact that he actually had a licence would not exonerate him if he did not know at the time that he had it.[82] Does the account offered here help to choose between these alternative ways of formulating a rule prohibiting the operation of a vehicle without a licence?

I think it does. On the offence side, the conduct of operating a vehicle does not appear to describe a harm or evil which it is the concern of the law to prevent. The analysis on the offence side thus suggests that it is necessary to include the absence of a licence as an element, since that qualification is required in order to make the description of the prohibited act wrongful. On the defence side, it is clear that having a licence does not reflect its own principle or value. Both offence and defence considerations thus suggest that we should treat the absence of a licence as an element of the offence.

A more difficult case is the sort of provision considered in *Mullaney* and *Patterson*, namely the defence of extreme emotional disturbance and its relation to the offence of murder. To recall, the question in those cases was whether a murder provision qualified by the reduction to manslaughter for extreme emotional disturbance should be thought of as relating offence and

[82] Again, this is assuming that the court is inclined to read a "reasonable belief" requirement into the affirmative defence of possession of a licence. It is always possible for a court or legislature to structure an affirmative defence so that the defendant need have had no belief with respect to it in order to claim it.

defence as negative offence element or as affirmative defence. If the former, both statutes would be invalid, on the grounds that they shift the burden to the defendant to prove the existence of the defence. If the latter, then the burden-shifting provisions are acceptable, and absent proof by the defendant that he was emotionally disturbed at the time of the killing, the establishment of the state's *prima facie* case would be sufficient to convict the defendant of murder. We have already suggested that one difference between the *Mullaney* and the *Patterson* statutes should not make a difference to our thinking on this question, namely whether the murder provision is drafted in a way that includes the term "unlawfully" in the offence definition. What is needed now is a deeper analysis of the relation between murder and extreme emotional disturbance in terms of the appeal to principle we have articulated.

From the standpoint of the theory of offence definition I have sketched, the Court was correct in *Mullaney* to think of a defence of extreme emotional disturbance as an implicit part of the murder provision it qualifies. On the offence side, the harm or evil that a murder provision targets is significantly reduced or entirely absent when the defendant's mental state is significantly impaired. Murder provisions are particularly designed to target intentional killing. It is only by taking too narrow a view of the notion of intentional action that we might be misled into thinking that extreme emotional disturbance does not impair the "intentionalness" of an agent's conduct. On the defence side, there is no *value* reflected in the allowance made for extreme emotional disturbance, as there is in the case of self-defence, necessity, or defence of others. Emotional disturbance is a psychological condition, not a normative stance to be promoted. Indeed, to the extent the law has anything normative to say about agents who experience such surges of emotion that they lack the self-restraint required to avoid criminal behaviour, it is that these defendants should *learn* greater self-control. Far from regarding emotional disturbance as a positive political principle requiring protection in a separate defence provision, such emotional states are part of what the criminal law would ultimately like to prevent and discourage.

The wisdom of treating the defence of extreme emotional disturbance along the lines of mistake and consent, rather than along the lines of self-defence and necessity, is reinforced when we consider the two other characteristics with which we have identified the choice between affirmative defence and negative offence element. On the question of *mens rea*, it seems reasonable to think of extreme emotional disturbance in roughly the same way we think of mistake. Just as a defendant who makes a mistake about a material element of an offence cannot be guilty of the offence because he lacks the required mental state with respect to that element, so too a defendant who commits a crime in a state of extreme emotional disturbance is likely to lack the *mens rea* if the crime requires intentional or knowing conduct. So like mistake, the defence of extreme emotional disturbance may

sometimes be derived from the offence definition, and this is why it should not be thought of as an affirmative defence.[83]

Finally, extreme emotional disturbance should also be thought of along the lines of mistake with respect to the requirement of reasonable belief that normally attaches to affirmative defences. Just as it was a confusion to apply a reasonableness requirement to the mistake defence, so it would be misguided to do so in the case of extreme emotional disturbance. While a reasonableness requirement is not a necessary concomitant of an affirmative defence, the fact that it would not even in theory be possible to attach such a requirement to a defence like extreme emotional disturbance indicates something about its status, namely that its absence should be thought an implicit part of the offences to which it applies.

The tendency of commentators and judges to regard the concept of a criminal offence as a product of legislative fiat reflects a deep positivistic orientation towards the criminal law. The dangers of this tendency are most strikingly revealed in a decision like *Patterson*, where the parameters of an important constitutional protection for individual liberty was made to depend upon the Court's willingness to articulate a meaningful demarcation of the line between offence and defence. The Court's abdication of this important task is only comprehensible against the background of a firm commitment to seeing the criminal law as largely devoid of normative content. But this stance is puzzling in light of the Court's apparent unwillingness to empty other constitutional guarantees of such content. It is surprising, then, that the Court should so readily embrace a morally vacuous reading of the presumption of innocence. The embrace, granted, has occurred by indirection: the meaningfulness of the presumption of innocence depends on the meaningfulness of the concept of an offence, and the vacuity of the constitutional concept is the product of having purged the criminal law concept of its natural understanding. The result, however, is the same as if the presumption of innocence were itself judged to be an empty and arbitrary principle.

This same "positivism" asserts itself at the more general jurisprudential level in the tendency to think of the parameters of a rule of prohibition as merely a matter of formulation, and to ignore the possibility of a deeper logic of prohibition and its overriding conditions. It ironically shows itself in the position that Dworkin and Schauer arrive at from different directions: the idea that there is no such thing as an exception to a rule, where an exception qualifies the rule from outside it without compromising the rule's validity. For to accept the collapse of exceptions into their associated rules of prohibition is to obscure the way in which general principles of justification call into play

[83] True, the language of "negative offence element" may sit uneasily. In the case of mistake, there is no single negative element that is a part of the offence definition for every crime to which mistake is a defence. The relevant offence element is simply anything that can be negatived by the existence of a mistake on the defendant's part. The same should probably be said about extreme emotional disturbance.

aspects of our political and moral commitments that are not reflected in the justification for the various rules of prohibition themselves. Which qualifications are, and which are not, exceptions to a rule depends on an understanding of the particular principles that stand behind both the rules and their associated qualifications.