Justifications, Excuses, and Affirmative Defenses

Murat C. Mungan*

October 30, 2018

Abstract

A defendant who admits to having committed an offense may nevertheless be acquitted if he can provide a legally cognizable justification or excuse for his actions by raising an affirmative defense. This article explains how affirmative defenses generate social benefits in the form of avoided unnecessary punishment. It then asks what kind of evidentiary standards must be used in order to balance these benefits against potential social costs arising from frivolous defense claims. It thereby provides an economic rationale for the uniformity across US jurisdictions in allocating the burden on the prosecution to prove the commission of the offense, as well as the variation across states in the standards of proof they use in determining the validity of affirmative defenses. The analysis also explains why mere assertions of undeterrability should not be considered as affirmative defenses.

Keywords: Justifications, excuses, defenses, affirmative defenses, judicial error, deterrence, punishment costs.

JEL classification: K0, K14, K40, K41, K42.

1 Introduction

In the United States, in order to secure a conviction, the prosecution must prove all elements of the offense. However, even if the prosecutor succeeds in doing this, the defendant may still avoid conviction if he can successfully raise an affirmative defense. These defenses are distinct from what are sometimes called *failure of proof defenses* or *negating defenses*, which consist of demonstrating that one of the elements of the offense has not been legally proven by the prosecution. Affirmative defenses, instead, typically protect a defendant who admits to having committed the offense, but claims that his acts were either justifiable or excusable. Common affirmative defenses include self-defense, defense of property, defense of others, necessity, duress, entrapment, insanity and intoxication.

*Professor, Antonin Scalia Law School at George Mason University. e-mail: mmungan@gmu.edu. I thank Florian Baumann, Bryan McCannon, Daniel Pi, and participants of the 35th Annual Meeting of the European Association of Law and Economics and the West Virginia University Department of Economics Seminar.

There is no variation, or room for variation, across states as to what burden the prosecutor must meet in order to legally prove that the defendant committed the offense. In 1970, *In Re Winship* unequivocally stated that the prosecution must prove all elements of an offense beyond a reasonable doubt. Interestingly, because the lack of an affirmative defense is not interpreted as being an element of the offense, states have some freedom in specifying the burden necessary to prove or disprove an affirmative defense.[1] This asymmetry has led to variation across states in what they consider to be affirmative defenses.[2] Perhaps more importantly, state laws also differ with respect to who carries what kind of burden of proof on issues relating to affirmative defenses. Thus, one state may require the prosecution to disprove an affirmative defense raised by the defendant beyond a reasonable doubt, while another state may require the defendant to prove the defense by preponderance of the evidence, while yet another state may allocate the burden on the defendant and require clear and convincing evidence.[3] Despite these variations, there is a commonality across states: an affirmative defense becomes an issue only if it is raised by the defendant who must provide some evidence to support it. Otherwise, it would be impracticable to allocate the burden of proof on the prosecution, since that would imply that the prosecutor must prove the absence of all possible affirmative defenses.

The presence of such large variation across states suggests the lack of a form of conventional wisdom, or common rationale, that explains the reasons for preferring one legal design over another. Interestingly, there exists, to the best of my knowledge, no economic analysis of burdens of proof in a context that involves issues related to both offenses and defenses.[4] This stands in sharp contrast to the sizeable economics literature analyzing the effects of various burden allocations in determining whether the actor has committed the harmful act.[5] A focal point of a strand in this literature is to explain the economic rationale behind having higher standards of proof in criminal trials than in civil trials. The issue of affirmative defenses, however, has not been analyzed as of

[1] *See, e.g.,* Gray (2017), discussing the rulings of Mullaney v. Wilbur, 421 U.S. 684 (1975) and Patterson v. New York, 432 U.S. 197 (1977).

[2] A state may, for instance, define murder as an intentional killing, and may allow a partial affirmative defense for acting under extreme emotional disturbance; or, it may define murder as an intentional killing which is not committed under extreme emotional disturbance. In the latter case, the prosecution would have to prove the absence of extreme emotional distress beyond a reasonable doubt. In the former case, the state may assign the burden of proving extreme emotional distress to the defendant.

[3] A few examples are useful to demonstrate this point. Colorado's criminal code, for instance, allocates the burden of proof to the prosecution and requires it to disprove the affirmative defense raised by the defendant beyond a reasonable doubt. On the other hand, New Jersey allows involuntary intoxication as an affirmative defense, but requires the defendant to prove the elements of the defense by clear and convincing evidence. A third variation is demonstrated by Michigan's criminal code, which requires a defendant who raises an insanity defense to prove his defense by a preponderance of the evidence.

[4] But, see, Hay and Spier (1997) briefly discussing a rationale, in the civil context, on allocating the burden of *production* to the defendant who raises an affirmative defense.

[5] *See, e.g.,* Demougin and Fluet (2005) and (2006), Lando (2009), Kaplow (2011), Mungan (2011), Rizzolli and Saraceno (2013), Obidzinski and Oytana (2017), Fluet and Mungan (2017), and Garoupa (2018).

yet.

In this article, I present an economic analysis of a very large class of affirmative defenses, namely those which are meant to protect acts that are either justified or are excusable. An act is justifiable if it enhances welfare. Thus, in the frequently studied hypothetical where a person breaks into a cabin to survive through a blizzard, the actor has a necessity defense based on a justification. Other affirmative defenses that are typically based on justifications are self-defense, defense of others, and defense of property. An act is excusable, if it is committed under circumstances where the actor could not be expected to react to legal incentives.[6] Insanity, duress, involuntary intoxication, and other diminished responsibility defenses are examples of defenses that are typically based on the defendant's actions being excusable.[7]

The objective of this article is to study three interrelated issues that pertain to these broad categories of defenses. First, as a preliminary matter, I ask what social benefits may be served by allowing defenses? Second, given that there are some social benefits from allowing defenses, what kind of evidentiary standards must be used in order to balance these benefits against potential social costs, and how do these standards compare to the analogous standards that must be used in issues pertaining to whether the defendant committed the act? In other words, how do the optimal standards of proof in determining whether the defendant committed an offense, and whether he has a valid defense, compare? Third, what reason is there to have structured affirmative defenses which require the presence of various elements, instead of having very broadly defined defenses? Is there not a good reason, for instance, to replace all defenses that the analysis pertains to, with one affirmative defense which states that the defendant ought to be acquitted whenever he has an excuse or justification, or alternatively, whenever doing so would be socially desirable?

To study the first question, I consider a model where courts can observe, without error, whether a person has committed an offense, and, whether he satisfies the requirements of a defense. In a simple Beckerian setting, where the monetary sanction is set optimally, there are no social gains generated by most defenses: if a person is deterred in the Beckerian setting, his offense causes harms that off-set his private benefits, and hence, there is typically no reason to change the person's behavior; or the optimal sanction fails to deter the person,

[6] In fact, this deterrence based explanation appears to be the primary rationale offered -at least a century ago- by some scholars for excuses, such as duress. *See, e.g.,* Hitchler (1917).

One can of course envision a person taking precautions to reduce the likelihood with which he will be involuntarily intoxicated or placed under duress. Similarly, one can envision the law having an effect on the behavior of even minors or people who have mental disabilities. However, the law generally defines these categories of defenses to be available only to a subgroup of individuals who are involuntarily intoxicated, placed under duress, are infants, or have mental disabilities. More importantly, the law places these limitations in a way that the defense is available only to the individuals who are likely to be the least responsive to incentives.

[7] Naturally, the line between justifications and excuses can become blurry in many cases. For example, it can be unclear whether a store clerk who is forced to hand over the cash in the store's registry at gun point is justified or excused, since, presumably she is unresponsive to legal incentives at gunpoint and her act enhances welfare.

which implies that the existence of a defense would not change his behavior. Thus, a rationale for defenses in the Beckerian setting is limited to exceptional cases where the commission of the act generates positive externalities to third parties other than the party being harmed as a result of the offense (e.g. defense of others). Hence, for most cases in which the actor has a justification or an excuse, there is no need for a defense, even when courts have perfect information and, therefore, can implement defenses without generating error costs. However, when punishment is non-transferable (e.g. imprisonment), then, under perfect information, there are always benefits to refraining from punishing individuals who are either unresponsive (because by definition their behavior cannot be changed by the presence or absence of punishment, hence, punishment is wasteful), or whose acts lead to benefits greater than social harms (because we want them to commit these acts, and thus, punishment generates costs, without any benefits). To formalize this idea, the model considers non-monetary punishment. However, it should be noted that similar conclusions can be reached by considering indirect costs, such as increases in efforts by offenders to avoid punishment.[8]

The perfect information case highlights potential gains from allowing defenses, in the form of a reduction in the harms inflicted through punishment. However, when there is imperfect information, the availability of defenses leads to wrongful acquittals, and, thereby leads to losses in deterrence, which translates into increased criminal harm. This is certainly a valid concern, but, similar deterrence costs are also implicated by the presence of errors in the determination of whether the person actually committed the offense. Thus, optimal burdens and standards of proof, whether for establishing affirmative defenses or offenses, trade-off harms from punishment versus losses in deterrence. Two key insights that pertain to this trade-off allow a comparison between the optimal burdens of proof for affirmative defenses versus offenses, and thereby address the second question studied.

The first insight is that a court that incorrectly finds that the defendant committed the offense incentivizes people -who do not face circumstances that would justify or excuse the commission of the act- to commit the offense by reducing the opportunity cost of committing the offense.[9] The same is not true for denying the defense to a person who commits the offense under justifiable or excusable circumstances, because the person in such circumstances is either unresponsive to incentives (e.g. when he is involuntarily intoxicated), already has the incentives to commit the act (e.g. in the case of self-defense, or in a case where the necessity defense would be applicable and the law is enforced via optimal sanctions), or it would be beneficial to incentivize him to commit the act (e.g. in the case of defense of others, where the internalized part of the justified act does not off-set the expected cost of conviction, and, thus, the

[8] *See, e.g.,* Malik (1990), Langlais (2008), and Sanchirico (2010).

[9] The exact impact of these types of type-1 errors on deterrence is unclear (*see* Lando and Mungan (2018)). The analysis presented here relies on type-1 errors reducing deterrence only to some degree, and is not affected by type-1 errors having a smaller effect on deterrence than type-2 errors due to reasons explored in Lando and Mungan (2018).

person who would be entitled to an affirmative defense is over-deterred). The second insight is that, for typical crimes, the population which does not commit the offense is much larger than the population of individuals who commit the act under justifiable circumstances. Thus, an increase in the frequency of wrongfully convicting people in the former category causes a much larger increase in the population of convicts than a comparable increase in the rate at which people in the latter category are punished. These two reasons, combined, reveal a strong rationale for having a stricter standard for offenses than for defenses, and thereby explain why it is optimal to allocate the burden of proof on the prosecution for matters that pertain to offenses. The same observations also explain that it may be socially desirable to allocate the burden of proof on the defendant for many affirmative defenses where (i) the defense is based on the actor having a justification or excuse, and (ii) the offense rate is small. However, the optimality of assigning the burden onto the defendant depends on specific conditions that pertain to the gains from deterrence versus costs of punishment, on the margin. The comparison between these gains and costs depends on the characteristics of the population in question, as well as punishment costs, and, thus, the optimal burden allocation may differ from one society to another.

These observations are useful in answering the third question: why have structured affirmative defenses at all? A more specific question allows us to consider this issue more discretely: why not allow a defense for a person who merely asserts that he had such large benefits from committing the crime that he would have been undeterred even if a defense was not available? Punishing such an individual is socially wasteful, and, therefore, in the perfect information case, it would be optimal to refrain from punishing him.[10] Thus, the answer to this question relies on the court's ability to distinguish between people who have valid justifications and excuses and those who do not. Structured affirmative defenses typically require proof of some concrete element which is, statistically, a good proxy for the person's likelihood of being unresponsiveness to enforcement schemes (e.g. that the person was suffering from a specific mental defect). The existence of these elements make the evidence generating process pertaining to the affirmative defense *informative*. This means that there are pieces of evidence which only people with actual defenses can provide with a high likelihood (e.g. a doctor's report pertaining to the mental defect of the defendant). Claims that refer to the unobservable benefit that one derived by committing a crime do not satisfy this property. Hence, many claims pertaining to the undeterrability of a particular defendant will not be associated with evidence generating processes that are very informative. Therefore, even when such defenses are allowed, it would be very difficult for a particular defendant to provide evidence sufficient to meet the optimal standard of proof. This observation suggests that it may not be welfare enhancing to allow defendants to raise such defenses, given the

[10] One response, not formally analyzed here, is that society obtains retributive benefits from the punishment of this hypothetical person, but, it does not obtain similar benefits from punishing a person with an excuse (e.g. an insane person, a child, a person under duress, or, more generally, a person with diminished capacity or responsibility). Thus, retributive benefits may supply a further rationale for denying some defenses.

costs a court would need to incur to simply entertain such claims. Therefore, one may conceive of structured defenses as a way of economizing on litigation costs by directing defendants towards only producing evidence which is informative.

I formalize these points in sections 2 -4. In sections 2 and 3, I construct a law enforcement model wherein defenses are possible under perfect, and imperfect information, respectively. In section 4, I analyze how the nature of the evidence pertaining to the affirmative defense affects the social value of allowing the defense. Section 5 suggests avenues for future research and concludes. Appendixes in the end contain various proofs, derivations, and supplementary analyses.

2 Affirmative Defenses with Perfect Information regarding Justifications and Excuses

Society consists of a continuum of risk-neutral individuals, who are clustered into three groups, and have opportunities to commit an offense which causes harm h. The first group consists of individuals, who have personal benefits (b) from committing crime, which varies among individuals within the group. The size of this group is normalized to 1, and the benefits from offending among this group are distributed with the cumulative distribution function (CDF) M, with $m = M'$. The support for these functions is $[\underline{b}, \overline{b}]$ with $\overline{b} < h$ and m is continuous for all $b > 0$, which implies that $m(\overline{b}) = 0$. The second group, of size φ_J, consists of individuals who similarly encounter opportunities to commit the offense, but have justifications for committing it: by committing the offense they obtain a private benefit of b, and potentially, confer an external benefit of v to third parties, such that $b + v > h$ for all individuals within the group.[11] Thus, the government has the option of making a justification defense available to individuals within this group who commit the harmful act. To ease notation, it is assumed that v is constant. The distribution of benefits from crime are captured by the CDF N, with $n = N'$ and both of these functions have support $[\underline{b}, \infty)$ where $\underline{b} > h - v$. The third group, of size φ_E, consists of individuals who are unresponsive to enforcement policies, either because they face extreme costs associated with not committing the crime (e.g. in the case of duress), or, they are simply unable to weigh costs and benefits. Thus, the government has the option of making a excuse defense available to individuals within this group who commit the harmful act. The average social benefit generated through the commission of the offense by individuals in this group is denoted $\widehat{b}$.[12]

The government can perfectly observe individuals' groups, but not their benefits. To deter crimes, the government punishes individuals who have committed them, unless they have a valid defense. To focus the analysis on the optimality

[11] The case where $v > 0$ captures defenses of others and some necessity defenses. $v = 0$ captures self-defense cases as well necessity defenses where the entirety of the harm to be prevented by the commission of the act are targeted towards the actor.

[12] A defense is not allowed for people in the first group who have benefits which exceed the sanction, because the government is unable to observe individuals' benefits. A rationale for this approach, is provided in section 4.

of defenses, the severity of the punishment is assumed to be exogenously given, and the punishment cost incurred by a convict is normalized to 1. In addition to the cost to the convict, punishment results in social costs of $(\sigma - 1)$ (e.g. due to costs of administration, maintenance, etc), such that the total cost of punishment is σ per-convict. For similar expositional reasons, the probability of auditing people's behavior is also normalized to 1. It is assumed that $\overline{b} > 1$ such that there is under-deterrence among the first group. A brief analysis in Appendix A shows that the standard Beckerian result holds when the audit probability and the punishment are determined by the government: it is optimal to impose the maximal sanction along with a low probability of audit which leads to under-deterrence. Thus, the inclusion of these variables as endogenous choices for the government does change any of the results that are derived based on the assumptions listed above.

In this section, the only choices of the government are whether or not to allow defenses based on justifications and excuses. Thus, welfare, can be expressed as

$$\widetilde{W}(\psi_J, \psi_E) = \int_1^{\overline{b}} (b - h - \sigma) m(b) db + \int_{b_J}^{\infty} (b + v - h - \psi_J \sigma) n(b) db + \varphi_E (\widehat{b} - h - \psi_E \sigma) \tag{1}$$

where

$$\begin{aligned} b_J &\equiv (1 - \psi_J)\underline{b} + \psi_J \max\{1, \underline{b}\} \qquad (2) \\ \psi_J &= \begin{array}{ll} 0 & \text{if justification defense} \\ 1 & \text{if no justification defense} \end{array} \text{; and} \\ \psi_E &= \begin{array}{ll} 0 & \text{if excuse defense} \\ 1 & \text{if no excuse defense} \end{array} \end{aligned}$$

An investigation of (1) reveals the following result.

Proposition 1 *(i) If punishment is socially costly (i.e. $\sigma > 0$), allowing defenses for justifications as well as excuses always enhances social welfare. (ii) If punishment is not socially costly (i.e. $\sigma = 0$), allowing defenses for excuses does not enhance welfare, and allowing defenses for justifications enhances welfare only if $\underline{b} < 1$.*

Proof. First, note that $\widetilde{W}(\psi_J, 0) - \widetilde{W}(\psi_J, 1) = \varphi_E \sigma$, thus, an excuse based defense enhances welfare if, and only if, $\sigma > 0$. Second, note that

$$\widetilde{W}(0, \psi_E) - \widetilde{W}(1, \psi_E) = \int_{\underline{b}}^{\infty} (b + v - h) n(b) db - \int_{\max\{1, \underline{b}\}}^{\infty} (b + v - h - \sigma) n(b) db \tag{3}$$

When $\underline{b} \geq 1$, (3) implies that $\widetilde{W}(0, \psi_E) - \widetilde{W}(1, \psi_E) = \varphi_J \sigma$, which implies that a justification based defense enhances welfare if, and only if, $\sigma > 0$. On the other hand, when $\underline{b} < 1$,

$$\begin{aligned} \widetilde{W}(0, \psi_E) - \widetilde{W}(1, \psi_E) &= \int_{\underline{b}}^{\infty} (b + v - h) n(b) db - \int_1^{\infty} (b + v - h - \sigma) n(b) db \\ &= \int_{\underline{b}}^{1} (b + v - h) n(b) db + \sigma[\varphi_J - N(1)] > 0 \qquad (4) \end{aligned}$$

which implies that justification based defenses enhance welfare whenever $\underline{b} < 1$. ■

The proof of proposition 1 simply demonstrates that compared to a regime where there is no defense, setting $\psi_E = 0$ enhances welfare by reducing the cost of imprisonment (by an amount equal to $\sigma\varphi_E$). The effect of a justification depends on whether sanctions have a deterrent effect on some people with justifications. When $1 \leq \underline{b}$, sanctions have no deterrence effect, and, thus, the only impact of a defense is to enhance welfare by eliminating punishment costs, which equal $\sigma\varphi_J$. On the other hand, when $1 > \underline{b}$, a justification defense causes an increase in the measure of justified offenses committed, from $N(1)$ to φ_J. Thus, allowing the justification defense causes a reduction of $\sigma[\varphi_J - N(1)]$ in imprisonment costs, while increasing social welfare due to justified offenses committed by $\int_{\underline{b}}^{1}(b + v - h)n(b)db$. It is worth noting that this latter social benefit is present only when some individuals with justifications do not have personal benefits that off-set the harm from crime, but, whose acts are nevertheless justifiable because they confer benefits to third parties such that $v > h - \underline{b}$. This may happen, for instance, in cases involving defense of others. The analysis also reveals that in simple models where the punishment is monetary, there is value to allowing affirmative defenses only in these latter cases, since $\sigma = 0$ implies no benefits in situations where $1 \leq \underline{b}$.

3 Affirmative Defenses with Imperfect Information

I now consider the case where courts make errors both in determining whether the defendant committed the act, and in determining whether the person is entitled to a defense. These errors occur, because courts do not possess perfect information, and must rely on noisy signals to make decisions about defendants. In particular, courts may commit type-1 and type-2 errors. In offense related determinations, type-1 errors refer to incorrectly finding that the defendant committed the offense, whereas in defense related determinations they correspond to incorrectly finding that the defendant is not entitled to a defense. Type-2 errors correspond to the opposite mistakes.

When these judicial errors are possible, the defendant is no longer certain that he will be convicted or acquitted. Thus, when he is required to submit evidence pertinent to his case, he may have to make a strategic decision, and choose to submit the type of evidence that increases his odds of acquittal. In reality, the types of evidence that increase the odds of raising a successful affirmative defense often require the person to admit that he has committed the offense. In fact, this is often embedded in the nature of the affirmative defense: in the quintessential necessity hypothetical studied by lawyers, the defendant asserts that he broke into a cabin to avoid freezing to death in a blizzard. This requires disclosing that he, in fact, broke in to the cabin. To capture this trade-off, I assume that people who commit the offense have the option of disclosing infor-

mation that is conclusive of the fact that they committed the act, but which generates a noisy signal about whether they committed the act under justifiable or excusable circumstance. This corresponds to assuming that the defendant bears what is called *the burden of production*, and that only people who have committed the offense can meet this burden. A procedural cost minimization rationale for this assumption is provided in Hay and Spier (1997). Moreover, the assumption is consistent with the evidentiary processes of many courts, and also simplifies the derivation of probabilities of conviction.

This framework guarantees that people who have not committed the offense focus on proving this fact. The choice for a guilty person, i.e. a person who has committed the offense without a justification or excuse, however, is not as clear. Presumably, the person's choice will depend on his belief about which issue the prosecutor has stronger inculpatory evidence on. To formalize this idea, I assume that guilty defendants are of different types, denoted as a $t \in [0,1)$, which are revealed to them as private knowledge after they commit the offense. The defendant's type is indicative of the relative strength of his affirmative defense case compared to his case with respect to whether he committed the offense.[13] A large t implies a weak case with respect defenses. As explained in Appendix B, this assumption, coupled with standard monotone likelihood ratio properties (MLRP) regarding noisy signals emitted by defendants imply that the probabilities of conviction for a type t defendant, conditional on not raising an affirmative defense, and raising an affirmative defense, respectively, can be expressed as:

$$\beta^o(\alpha^o) \text{ and } \beta^d(t, \alpha^d) \tag{5}$$

Here, $\alpha^o \in [0,1]$ denotes the probability of conviction for a person who has not committed the offense. Similarly, $\alpha^d \in [0,1]$ is the probability with which a person who has a justification or an excuse is convicted after raising an affirmative defense. Therefore, $\alpha^{i\in\{o,d\}}$ represent type-1 errors that the court may commit when assessing whether a person has committed the offense, and when assessing whether the person has a valid defense, respectively. As is noted in the prior literature,[14] there is an equivalence between courts choosing a threshold evidentiary standard, and fixing the type-1 error at a particular level. Thus, to simplify the exposition, here, I take the government's choice variables as α^o and α^d, and relegate the formal derivation of these probabilities from evidence generation processes to Appendix B. This simplified notation allows the expression of probabilities of conviction for guilty individuals as functions of the government's choice of type-1 errors and their types, as in (5).

Some properties of $\beta^{i\in\{o,d\}}$ are worth highlighting, as they play an important role in derivations that follow. In particular,

$$\begin{aligned} \beta^o(0) &= \beta^d(t,0) = 0; \text{ and} \\ \beta^o(1) &= \beta^d(t,1) = 1 \text{ for all } t \in [0,1) \end{aligned}$$

[13] One can, alternatively, consider two-dimensional types to consider the strength of guilty defendants' cases over defenses and offenses, respectively. An analysis of this case reveals that it complicates the notation without providing any insights beyond what is presented here.

[14] *See, e.g.*, Demougin and Fluet (2005) and Fluet and Mungan (2017).

Moreover, MLRP implies that

$$\frac{\partial \beta^i}{\partial \alpha^i} > 0 \text{ for } i \in \{o, d\} \tag{6}$$

That a high t indicates a weaker defense case is reflected by

$$\frac{\partial \beta^d}{\partial t} > 0 \tag{7}$$

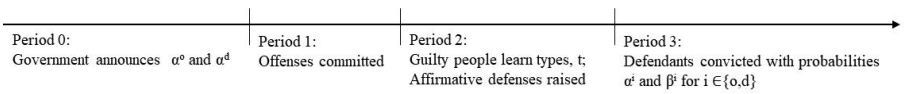
Moreover, to guarantee that there are always some guilty defendants who choose to raise a defense and some who choose to not, I assume that

$$\lim_{t \to 1} \beta^d(t, \alpha^d) = 1 \text{ and } \beta^d(0, \alpha^d) = \alpha^d \tag{8}$$

This last property ensures that no attention is diverted towards trivial and suboptimal cases where one of the options is never exercised by guilty individuals. Assuming that types are continuously distributed, such cases obviously cannot be optimal, because the government could increase welfare by slightly reducing the type-1 error over the dimension which is never chosen by guilty individuals, and, hence increase welfare by reducing the number of wrongful convictions. Thus, the only function of (8) is to simplify the analysis.

Finally, it is worth highlighting a possibility which I have not yet considered. People who have justifications and excuses may, counterintuitively, find it preferable to not raise a defense, if the standard used in the determination of offenses provide significantly greater protections to defendants than those used in the determination of defenses. A simple observation reveals that type-1 errors that generate this result can never be optimal: since guilty individuals face a higher probability of conviction upon raising an affirmative defense than defendants who truly have affirmative defense, a choice of α^o and α^d that induces people with justifications or excuses to refrain from raising affirmative defenses will also induce guilty individuals to refrain from raising affirmative defenses. Thus, the government can improve upon a system that causes this result by reducing α^d until only people with actual justifications and excuses choose to raise affirmative defenses. This policy change increases welfare by reducing punishment costs, while leaving the incentives of people who are responsive to policies, unchanged. Thus, in the remaining analysis, I focus only on cases where $\alpha^d \leq \beta^o(\alpha^o)$, and formalize, in note 18 below, the suboptimality of cases where $\alpha^d > \beta^o(\alpha^o)$.

Given the properties and assumptions listed above, the timeline of events can be summarized as follows.



3.1 Individuals' Decision Making Processes and Behavior

Individuals' behavior can be analyzed via backward induction. In the second period, a person with a justification chooses to raise an affirmative defense, since[15]

$$\alpha^d \leq \beta^o(\alpha^o) \tag{9}$$

Whenever this condition holds, it follows, via (8), that there exists a critical type, $\bar{t}(\alpha^o, \alpha^d)$ such that

$$\beta^d(\bar{t}, \alpha^d) = \beta^o(\alpha^o) \tag{10}$$

such that guilty defendants choose to raise an affirmative defense if, and only if, $t \leq \bar{t}$.

Thus, in period 1, a person who has no justification or excuse, faces a probability of conviction of[16]

$$\beta(\alpha^o, \alpha^d) \equiv P(t \leq \bar{t})E[\beta^d | t \leq \bar{t}] + P(t > \bar{t})\beta^o \tag{11}$$

where the arguments of $\bar{t}$ and $\beta^{i \in \{o,d\}}$ are omitted to simplify notation, and P and E denote the likelihood of an event and the expectation operator, respectively. Therefore, a person who has no justification or excuse expects a net benefit of $b - \beta$ from committing crime. The same individuals face a conviction probability of α^o if they choose not to commit crime. Therefore, a person who has no justification or excuse commits crime if

$$b^*(\alpha^o, \alpha^d) \equiv \beta(\alpha^o, \alpha^d) - \alpha^o < b \tag{12}$$

On the other hand, a person who has a justification or excuse, commits the act if

$$b^{**}(\alpha^o, \alpha^d) \equiv \alpha^d - \alpha^o < b \tag{13}$$

It is possible for some people with justifications to be deterred from committing crime, if $1 > \underline{b}$, since the strongest standard for offenses (i.e. $\alpha^o = 0$) combined with the weakest standard for defenses (i.e. $\alpha^d = 1$) maximizes deterrence for these individuals and causes them to refrain from committing the act if they have benefits smaller than the sanction. This case represents situations where the justification would not exist but for the presence of positive externalities conferred on to third parties through the actions of the actor, and is analyzed in Appendix C. The analysis of this case requires additional notation, and does not reveal results that significantly differ from the case where $1 < \underline{b}$. Thus, to explain the main results in a compact manner the analysis in section 3.3. focuses on cases where justified acts generate benefits to the actor that exceed the sanction (i.e. $\underline{b} > 1$).

[15] I assume that defendants who are indifferent choose to raise an affirmative defense.

[16] When $\alpha^d > \beta^o(\alpha^o)$, it naturally follows that $\beta(\alpha^o, \alpha^d) = \beta^o(\alpha^o)$.

3.2 Ranking Standards of Proof

Before analyzing the optimality of various standards of proof, a couple of observations can be made to compare the strength of various standards that correspond to the type-1 errors that the government may choose to use. The terms that play a key role in making such comparisons are $\frac{\partial\beta(\alpha^o,\alpha^d)}{\partial\alpha^o}$ and $\frac{\partial\beta(\alpha^o,\alpha^d)}{\partial\alpha^d}$. These terms refer to the impact that the standards of proof over dimensions o and d have on the ex-ante probability of convicting a person who, in reality, has committed the act without a justification or excuse. In particular, a change in the standard that causes a one unit change in the probability of type-1 error over dimension i has the effect of increasing the ex-ante probability of convicting a guilty person by $\frac{\partial\beta}{\partial\alpha^i}$. This can be true, only if the threshold signal used by the court in determining verdicts is one which is $\frac{\partial\beta}{\partial\alpha^i}$ times likely to be produced by a guilty individual rather than one who is not guilty. In more familiar jargon, $\frac{\partial\beta}{\partial\alpha^i}$ corresponds to the likelihood ratio pertaining to the evidence generating process. Thus, a large $\frac{\partial\beta}{\partial\alpha^i}$ corresponds to a strong standard of proof, since it leads to convictions only if the signal is more frequently produced by guilty individuals than individuals who are not guilty due to the presence of the exculpating factor in question (i.e. non-commission of the offense if $i = o$, and a justification or excuse if $i = d$). The special case where the standard of proof is such that $\frac{\partial\beta}{\partial\alpha^i} = 1$ has a very intuitive meaning: it represents that standard which convicts an individual, if, and only if, the signal emitted by that person is more frequently emitted by a guilty individual rather than an individual who possesses the exculpating factor being investigated. Due to this reason, this standard is called *preponderance of the evidence,* as discussed in Demougin and Fluet (2006). This standard is very useful in discussing the relationship between various standards and burdens of proof. However, defining some terms is useful to eliminate potential ambiguities.

Definition 1 *(i) The 'likelihood ratio of a signal' refers to the frequency with which the signal is produced by a guilty individual divided by the frequency with which the signal is produced by a person who possesses the exculpating factor in question. (ii) A standard of proof is a threshold value $s \in [0, \infty)$, such that the court convicts the defendant whenever the likelihood ratio of the signal emitted by the defendant exceeds this threshold value. (iii) Standard s' is stronger than standard s'', if $s' > s''$. (iv) $s = 1$ is the preponderance of the evidence standard. (v) A standard allocates the burden of proof on the defendant if $s < 1$, and it allocates the burden of proof on the prosecution if $s > 1$.*

Definition 1 makes references to signals and likelihood ratios, which are formally considered only in Appendix B. Lemma 1, below, establishes the relationships between these concepts and $\frac{\partial\beta}{\partial\alpha^i}$, whose intuitive meaning is discussed above. This structure attempts to expedite the derivation of main results and keep the analysis focused. The reader who is interested in signal generating processes may consult Appendix B before moving forward.

Lemma 1 *(i) The standard used for issue $i \in \{o, d\}$ is the preponderance of the*

evidence standard if and only if $\frac{\partial \beta}{\partial \alpha^i} = 1$. *(ii) On issue* $i \in \{o, d\}$, *the defendant has the burden of proof if* $\frac{\partial \beta}{\partial \alpha^i} < 1$, *and the prosecution has the burden of proof if* $\frac{\partial \beta}{\partial \alpha^i} > 1$. *(iii) A stronger standard is employed in the determination of issue* i *than issue* j *if* $\frac{\partial \beta}{\partial \alpha^i} > \frac{\partial \beta}{\partial \alpha^j}$.

Proof. See Appendix B

Lemma 1 reveals how one can rank and compare standards of proof only by focusing on how standards chosen by the government impact type-1 errors relative to type-2 errors. Before, proceeding, it may be useful to acknowledge that, definition 1 and lemma 1, above, refer to the *burden of proof* and the *strength* of standards in a different way than some of the usages encountered in the existing literature.

The function of the burden of proof, here, is to identify the party whom the noisy signal must favor for it to win. The signal favors the defendant, if it is more frequently produced by a person who possesses the exculpatory factor the defendant is claiming to possess than a guilty person. The signal favors the prosecution, if it is more frequently produced by a guilty person than a person who possesses the exculpatory factor the defendant is claiming to possess. This use of the phrase coincides, to a large extent, with what is often called the *burden of persuasion* in the law,[17] because in most cases where the defendant bears the burden of persuasion, he has to produce evidence that is more consistent with his claim than the alternative.

The usage of the word *strength* is not completely aligned with the way one thinks of the strength associated with a standard in real legal disputes. The misalignment arises (only) in cases where the burden of proof is on the defendant. This can be illustrated by noting that requiring the defendant to prove his case with clear and convincing evidence is thought to correspond to a 'stronger' standard of proof than preponderance of the evidence, since more is demanded from the defendant. However, definition 1 would rank the former standard as being weaker than the latter, because it ranks standards based on how much protection they afford to the defendant, and calls standards that afford greater protections, 'stronger'. This definition has the advantage of having a monotonic relationship with likelihood ratios, and thus, allows more compact descriptions of results.

Next, optimal burdens of proof are characterized by using the rankings identified in this section.

3.3 Optimal Burdens of Proof

In identifying the optimal burdens and standards of proof, the analysis focuses on the case where people with justifications are never deterred (i.e. $\underline{b} > 1$). (The alternative case, whose analysis reveals similar results, is analyzed in Appendix C.) This condition is met, for instance, when the justifications in question are

[17] *See, e.g.,* Hay and Spier (1997), where *burden of proof* is used to describe something that is quite simlar to the *burden of production* in legal disputes.

based on acts which confer benefits to the actor which are greater than the harms from the offense (i.e. $\underline{b} > h$). In these cases, to simplify notation, we may denote the measure of individuals with justifications or excuses, as φ and express the total net-benefit caused by the acts of individuals with the symbol μ. Thus, welfare, which equals the expected sum of net-benefits, can be expressed as:

$$W(\alpha^o, \alpha^d) = \int_{b^*}^{\overline{b}} (b - h - \beta\sigma) m(b) db - \sigma[\varphi\alpha^d + M(b^*)\alpha^o] + \mu \qquad (14)$$

where the arguments of b^* and β are omitted, and derivations below contain similar omissions to abbreviate expressions.

The impacts of changes in α^o and α^d on welfare, whenever $\alpha^d \leq \beta^o(\alpha^o)$, are:[18]

$$\begin{aligned} \frac{\partial W}{\partial \alpha^o} &= \left(\frac{\partial \beta}{\partial \alpha^o} - 1 \right) (h + (\sigma - 1)(\beta - \alpha^o)) m(b^*) \qquad (15) \\ &\quad -\sigma \left[[1 - M(b^*)] \frac{\partial \beta}{\partial \alpha^o} + M(b^*) \right]; \text{ and} \\ \frac{\partial W}{\partial \alpha^d} &= \frac{\partial \beta}{\partial \alpha^d} (h + (\sigma - 1)(\beta - \alpha^o)) m(b^*) \qquad (16) \\ &\quad -\sigma \left[[1 - M(b^*)] \frac{\partial \beta}{\partial \alpha^d} + \varphi \right] \end{aligned}$$

It follows that the optimal α^o and α^d are interior, assuming that the optimal solution involves some deterrence.[19] Thus, the optimal standards of proof are characterized by the following first order conditions:

$$\frac{\partial \beta}{\partial \alpha^o} = \frac{\sigma M(b^*) + (h + (\sigma - 1)(\beta - \alpha^o)) m(b^*)}{(h + (\sigma - 1)(\beta - \alpha^o)) m(b^*) - \sigma[1 - M(b^*)]}; \text{ and} \qquad (17)$$

$$\frac{\partial \beta}{\partial \alpha^d} = \frac{\sigma\varphi}{(h + (\sigma - 1)(\beta - \alpha^o)) m(b^*) - \sigma[1 - M(b^*)]} \qquad (18)$$

An inspection of (17) and (18) reveals the following.

Proposition 2 *(i) It is optimal to place the burden on the prosecution to prove that the defendant committed the offense. (ii) It is optimal to employ a stronger*

[18] To see that cases where $\alpha^d \geq \beta^o(\alpha^o)$ are suboptimal, note that in such cases welfare is unaffected by α^d and equals:

$$\int_{b^*}^{\overline{b}} (b - h - \beta^o\sigma) m(b) db - \sigma[\varphi\beta^o + M(b^*)\alpha^o] + \mu.$$

Thus, whenever $\alpha^d > \beta^o(\alpha^o)$, welfare can be enhanced by reducing α^d below $\beta^o(\alpha^o)$, since

$$\frac{\partial W(\alpha^o, \beta^o(\alpha^o))}{\partial \alpha^d} = -\sigma\varphi < 0$$

which follows from (27).

[19] That some degree of deterrence is desirable implies that the optimal α^d and α^o are positive. Note 18, above implies that the optimal α^d is smaller than 1, and (15) implies that $\frac{\partial W(1, \alpha^d)}{\partial \alpha^o} < 0$. Thus, the optimal solutions are interior.

standard for proving that the defendant committed the offense than the standard used for determining whether the defendant has a valid defense, if the number of people who do not commit the act is greater than the number of people who have valid defenses or if the marginal deterrence benefits are sufficiently large (i.e. if $\sigma\varphi < \sigma M(b^*) + (h + (\sigma - 1)(\beta - \alpha^o))m(b^*)$*). (iii) It is optimal to place the burden on the defendant to prove that he has a valid defense if, and only if,* $\sigma[\varphi + [1 - M(b^*)]] < (h + (\sigma - 1)(\beta - \alpha^o))m(b^*)$.

Proof. (i) The condition in (17) implies that $\frac{\partial \beta}{\partial \alpha^o} > 1$ as long as $\sigma > 0$. This implies, via lemma 1, that the burden is on the prosecution. (ii) (17) and (18) imply that $\frac{\partial \beta}{\partial \alpha^o} > \frac{\partial \beta}{\partial \alpha^d}$ if $\sigma\varphi > \sigma M(b^*) + (h + (\sigma - 1)(\beta - \alpha^o))m(b^*)$, in which case, it follows via lemma 1 that there is a stronger standard for proving that the defendant committed the offense than the standard used for determining whether the defendant has a valid defense. (iii) It follows via (18) that $\frac{\partial \beta}{\partial \alpha^d} < 1$ if $\sigma[\varphi + [1 - M(b^*)]] < (h + (\sigma - 1)(\beta - \alpha^o))m(b^*)$, in which case, it follows, via lemma 1, that the burden of proof is on the defendant. ■

Proposition 1 summarizes the main result; the asymmetry in the way burdens of proof are allocated across issues pertaining to defenses versus offenses can be rationalized by focusing on the incentive effects and the statistical rareness of circumstances giving rise to valid affirmative defenses. The difference in incentive effects can be observed by noting that the marginal deterrence gains associated with an increase in α^o are proportional to $\left(\frac{\partial \beta}{\partial \alpha^o} - 1\right)$ whereas an increase in α^d causes similar gains that are proportional to $\frac{\partial \beta}{\partial \alpha^d}$. This is due to the asymmetry in deterrence effects previously explained. Similarly, the marginal increase in the imprisonment cost caused by an increase in α^o is greater than the analogous increase caused by a similar increase in α^d, as long as the number of people who elect not to commit the offense is greater than the number of people who have justifications and excuses. The final result, highlighted in proposition 1 suggests that whether the defense or prosecution ought to bear the burden in proving the presence or absence of a defense depends on the number of people who are on the margin, i.e. indifferent between committing crime and not committing crime, the harm from crime, the cost of imprisonment, and the number of people who are, in actuality, not guilty, either because they have a justification or have not committed the crime.

Overall, the results are, in a certain way, consistent with what we observe across jurisdictions in the United States: there is uniformity across jurisdictions in allocating the burden of proof to the prosecution in proving that the defendant committed the act. The model suggests that this is an optimal practice. There is, however, variation across jurisdictions, in terms of who carries the burden of proof when an affirmative defense is raised. The model suggests that either allocation of the burden of proof can be optimal, depending on the relationship between deterrence benefits, and punishment costs avoidable, on the margin.

4 Informativeness and the Social Value of Affirmative Defenses

The preceding analysis demonstrates that affirmative defenses enhance welfare, even when courts have imperfect information. Moreover, this result is derived in a fairly general framework, and, thus, it may lead one to believe that allowing all types of defenses, even one which simply relies on the defendant's benefit from crime being large, may enhance welfare. In this section, I demonstrate that the value from allowing the defendant to raise an affirmative defense is limited by the informativeness of the type of evidence which gives rise to the defense. In particular, as the evidence generating process associated with the defense gets less informative, the potential welfare gains from allowing the defense are diminished. This suggests that it is inefficient to allow the defense, when there are costs associated with reviewing the defendant's claims.

To formalize the informativeness of the evidence generating process, note that the difference between $\beta^d(t, \alpha^d)$ and α^d can be thought of as a measure of the extent to which a court can distinguish between a type t offender and a person who has a valid defense by using an evidentiary standard that generates these probabilities of conviction. For similar reasons,

$$E[\beta^d(t, \alpha^d)] - \alpha^d \tag{19}$$

can be thought of as a measure of the potential of a standard of proof to discriminate between guilty defendants and defendants who have valid defenses. Thus, if we denote by $k(t, I)$ the probability density function over types, where $I \in (0, 1)$ measures the informativeness of the evidence generating process, we can express (19) as:

$$\int_0^1 \beta^d(t, \alpha^d)k(t, I)dt - \alpha^d \tag{20}$$

Intuitively, as k assigns greater densities to types whose guilt can more easily be ascertained, it causes the court to better distinguish between people with valid defenses, and defendants who raise false defenses. This becomes obvious when one imagines the extreme case (which is ruled out by the assumption that $t \in [0, 1)$) where all guilty people are type $t = 1$, such that they are convicted with certainty when they raise an affirmative defense. The opposite extreme (which, again, is ruled out by the assumption that $k(t, I) > 0$ for all $t \in [0, 1)$) corresponds to a case where the evidence generating process is completely uninformative; all guilty people are type $t = 0$, such that they all have the same probability of being convicted as people with valid justifications. This corresponds to the case where the court is unable to distinguish between people with valid defenses and guilty people, to any degree. Cases in between are more or less informative according to the weight that k places on high types and low types. These observations can be operationalized by assuming that I is a shift parameter, such that $\frac{d\int_0^a k(t,I)dt}{dI} < 0$ and $\lim_{I\to 0} \int_0^a k(t, I)dt = 1$ for any $a > 0$.

Intuitively, as the evidence generating process gets less informative, the only way the court can better distinguish between a large proportion of defendants who dishonestly raise an affirmative defense and who truthfully raise an affirmative defense is to use a lower standard of proof, i.e. increase the type-1 error α^d. This is because, given a fixed level of α^o, the effect of a reduction in the informativeness of the evidence process is to increase the measure of defendants who dishonestly raise affirmative defenses. Thus, the court can counteract this effect by reducing the protections available to defendants, and thus, deter people without valid justifications or excuses from raising affirmative defenses. In the limiting case where the evidence generation process approaches being completely uninformative, i.e. $I \to 0$, the only way to separate people with valid affirmative defenses from those who do not, is to allow the type-1 error to converge to $\beta^o(\alpha^o)$, since only people with valid defenses are willing to raise them when $\alpha^d = \beta^o(\alpha^o)$. The implications of this observation vis-à-vis the social value of affirmative defenses becomes apparent by noting that, then, social welfare converges to

$$W(\alpha^o, \beta^o(\alpha^o)) = \int_{(\beta^o - \alpha^o)}^{\overline{b}} (b - h - \beta^o \sigma) m(b) db - \sigma[\varphi \beta^o + M(b^*)\alpha^o] + \mu$$

which is the same level of welfare obtained without the use of affirmative defenses. Thus, we have the following result.

Proposition 3 *As the evidence generating process associated with the affirmative defense approaches being completely uninformative, the social value from allowing the affirmative defense approaches zero.*

Proof. See Appendix B.

The formal proof for this proposition is relegated to the appendix, as it involves the use of additional notation, and follows steps that are very similar to those outlined in the text preceding the proposition. An important corollary of this observation, which follows from a simple continuity argument, is the following.

Corollary 1 *If allowing defendants to introduce affirmative defenses generates administrative costs, then there exists sufficiently uninformative evidence generating processes which make it inefficient to allow affirmative defenses.*

Corollary 1 illustrates the idea that affirmative defenses which rest on highly speculative claims and must rely on uninformative evidence generating processes are likely to bring about social benefits that are outweighed by the administrative costs of allowing such defenses. This observation provides a rationale as to why many defenses contain elements that pertain to events that can be reliably and objectively verified and lead to a certain degree of discrimination between honest and deceptive defendants.

5 Conclusion

There is uniformity in the burden of proof that is applicable in a criminal trial when the issue pertains to whether the defendant committed the offense. This is contrasted by great variation across jurisdictions as to who carries the burden of proof, and what the applicable standard is, when the issue pertains to defenses. The contrast between the uniformity in one dimension and the variation in the other dimension may appear troubling or surprising. However, an analysis of the differences between the two contexts reveals quite simple rationales for this contrast. Errors in the determination of whether a person committed an offense, and errors in the determination of whether the person has a valid justification or excuse have different consequences. In particular, incorrectly denying defenses reduce welfare less than similar errors committed in the determination of offenses: they lead to smaller reductions in deterrence, and they increase punishment costs to a lesser degree. Although the dynamics that lead to these observations imply that the prosecution ought to bear the burden of proof in offense related issues, they do not necessarily imply that it is optimal to assign the burden of proving the existence of valid defenses to the defendant. The optimality of such allocations depend on conditions that pertain to the gains from deterrence versus costs of punishment, on the margin.

These insights can be used to study related issues in future research. One may, for instance, attempt to measure the ratios between marginal gains from deterrence and marginal punishment costs across different jurisdictions, and question whether variations in these ratios may explain variations in the way jurisdictions assign burdens of proof for affirmative defenses. Moreover, insights provided in the analysis can be used to study the optimality of partial defenses which were not analyzed in this article. For instance, a defendant in a homicide case who successfully proves the existence of adequate provocation and/or extreme emotional distress is generally punished for a lesser offense than a defendant who has no similar defense. One may investigate whether there exists a deterrence or punishment cost based rationale for the existence of such defenses, and, if so, what the optimal standard of proof should be in demonstrating the validity or invalidity of such partial defenses.

6 Appendix A: Endogenous Sanctions and Detection Probabilities

When the government can choose the punishment which may not exceed a maximum sentence of $\overline{z}$, and when it can set the audit probability p by incurring a cost of $c(p)$ (with $c' > 0 > c''$) it follows that welfare is given by

$$\widetilde{W}(p,z,\Delta) = \int_{pz}^{\overline{b}} (b-h-\sigma pz)m(b)db + \int_{b_J}^{\infty} (b-h-\psi^J \sigma pz)n(b)db + \varphi_E(\widehat{b}-h-\psi^E \sigma pz) - c(p) \tag{21}$$

where Δ refers to the defense policy adopted by the government described in section 2, and

$$b_J \equiv (1-\psi_J)\underline{b} + \psi_J \max\{pz, \underline{b}\} \tag{22}$$

It follows from (21) that whenever $z = z_l < \overline{z}$, and $p = p_h$, welfare can be increased by increasing z_l to some $z_h \in (z_l, \overline{z}]$, while simultaneously decreasing p_h to $p_l = \frac{z_l p_h}{z_h}$. This is because such changes lead to no change in deterrence or punishment costs, while reducing the cost of audit, $c(p)$. Thus, given that defenses and the punishment are chosen optimally, it follows that the impact of a change in the probability of audit on welfare is given by $\frac{\partial \widetilde{W}(p,\overline{z},\Delta)}{\partial p} = -c'(\frac{\overline{b}}{\overline{z}})$ for all $p \geq \frac{\overline{b}}{\overline{z}}$, since $m(\overline{b}) = 0$. Thus, it is optimal to have an audit probability smaller than $\frac{\overline{b}}{\overline{z}}$, which leads to under-deterrence.

7 Appendix B: Proofs and Derivation of Conviction Probabilities

Probabilities of Conviction from Signal Generating Processes

People potentially emit two signals, $x \in [0,1]$ and $y \in [0,1]$. The first signal is used as evidence in determining whether the person committed the act, is produced by all individuals, and is always observed by the court. The second signal is produced only by people who committed the act, and is observed by the court only if the defendant raises an affirmative defense. Thus, the observation of any y by the court is conclusive evidence that the person committed the act.

A person who has not committed the offense emits signal x with probability $f(x|n)$ where n stands for *not committed*, whereas a person who has committed the act emits the same signal with probability $f(x|a)$ where a stands for *act committed*. Here, the functions $f(x|i \in \{a,n\})$ are probability density functions with support $[0,1]$. Similarly, a person who has a valid justification or excuse, produces a signal y with probability $g(y|j)$ where j stands for *justification or excuse* and a person who does not have a valid justification or excuse produces signal y with probability $g(y|t)$ where $t \in [0,1)$ is the guilty defendant's type revealed to him after he commits the act. The signal y is observed by the court only if the defendant raises an affirmative defense. $g(x|i \in j \cup [0,1))$ are also probability density functions with support $[0,1]$. Types are distributed with the probability density function $k(t)$. The monotone likelihood ratio (MLRP) holds, such that $\frac{\partial\left(\frac{f(x|a)}{f(x|n)}\right)}{\partial x} < 0$ and $\frac{\partial\left(\frac{g(y|t)}{g(y|j)}\right)}{\partial y} < 0$ for all $t \in [0,1)]$. This means that small values of x are more consistent with the person having committed the act than large values of x. Similarly, small values of y are more consistent with the person having a justification or excuse than large values of y. Standards of proof correspond to threshold signal values, denoted $\overline{x}$ and $\overline{y}$, such that the court produces a verdict that favors the defendant when $x < \overline{x}$ and when $y < \overline{y}$.

Denoting the cumulative distribution functions associated with f and g as F and G, it follows that type-1 errors generated by the standards chosen by the

government are

$$\begin{aligned} \alpha^o(\overline{x}) &= F(\overline{x}|n); \text{ and} \\ \alpha^d(\overline{y}) &= G(\overline{y}|j) \end{aligned} \tag{23}$$

since both functions are increasing in the standard, it follows that $\overline{x}$ and $\overline{y}$ can be expressed as functions of targeted levels of type-1 error, as follows:

$$\begin{aligned} \overline{x} &= F^{-1}(\alpha^o|n); \text{ and} \\ \overline{y} &= G^{-1}(\alpha^d|j) \end{aligned} \tag{24}$$

where $F^{-1}(\alpha^o|n)$ and $G^{-1}(\alpha^d|j)$ denote the inverses of the functions $F(\overline{x}|n)$ and $G(\overline{y}|j)$.

Using this notation, one can express the probability of correct verdicts as follows:

$$\begin{aligned} \beta^o(\alpha^o) &= F(F^{-1}(\alpha^o|n)|a); \text{ and} \\ \beta^d(t,\alpha^d) &= G(G^{-1}(\alpha^d|j)|t) \end{aligned} \tag{25}$$

Assuming that $\overline{x}$ and $\overline{y}$ are chosen such that $\beta^o(\alpha^o) \geq \alpha^d$, it follows that the ex-ante probability of being convicted, conditional on committing an offense without a justification, is:

$$\beta = \int_0^{\overline{t}} \beta^d(t,\alpha^d)k(t)dt + \int_{\overline{t}}^1 \beta^o(\alpha^o)k(t)dt \tag{26}$$

where $\overline{t}$ is defined in (10).

Proof of Lemma 1 Differentiating β with respect to α^d and α^o reveals that

$$\begin{aligned} \frac{\partial\beta}{\partial\alpha^o} &= \frac{\int_{\overline{t}}^1 f(\overline{x}|a)k(t)dt}{f(\overline{x}|n)}; \text{ and} \\ \frac{\partial\beta}{\partial\alpha^d} &= \frac{\int_0^{\overline{t}} g(\overline{y}|t)k(t)dt}{g(\overline{y}|j)} \end{aligned} \tag{27}$$

Thus, $\frac{\partial\beta}{\partial\alpha^o}$ and $\frac{\partial\beta}{\partial\alpha^d}$ correspond to the standards defined in definition 1, since, due to MLRP,

$$\begin{aligned} \frac{\int_{\overline{t}}^1 f(x|a)k(t)dt}{f(x|n)} &\geq \frac{\partial\beta}{\partial\alpha^o} \text{ iff } y \leq \overline{y}; \text{ and} \\ \frac{\int_0^{\overline{t}} g(y|t)k(t)dt}{g(y|t)} &\geq \frac{\partial\beta}{\partial\alpha^d} \text{ iff } x \leq \overline{x} \end{aligned} \tag{28}$$

This implies that $\frac{\partial\beta}{\partial\alpha^o}$ and $\frac{\partial\beta}{\partial\alpha^d}$ can be replaced with s in definition 1 to obtain the results stated in the lemma.■

Proof of Proposition 3 First, note that for all α^d and α^o such that $\alpha^d < \beta^o(\alpha^o)$, it follows that $\lim_{I\to 0}\beta = \lim_{I\to 0}\int_0^{\bar{t}}\beta^d(\alpha^d,t)k(t,I)dt + \lim_{I\to 0}\beta^o(\alpha^o)\int_0^a k(t,I)dt = \lim_{I\to 0}\int_0^{\bar{t}}\beta^d(\alpha^d,t)k(t,I)dt = \beta^d(\alpha^d,0) = \alpha^d$. Thus, for any α^d and α^o such that $\alpha^d \in (\alpha^o, \beta^o(\alpha^o))$, it follows that $\lim_{I\to 0}W(\alpha^o,\alpha^d) = \int_{(\alpha^d-\alpha^o)}^{\bar{b}}(b-h-\alpha^d\sigma)m(b)db - \sigma[\varphi\alpha^d + M(b^*)\alpha^o] + \mu$. Next, consider the pair $\widehat{\alpha}^d, \widehat{\alpha}^o$ such that $\widehat{\alpha}^d > \alpha^d$ and $\widehat{\alpha}^o = [\beta^o]^{-1}(\alpha^d)$ where $[\beta^o]^{-1}$ denotes the inverse of β^o. It follows that $\beta^d(t,\alpha^d) \geq \widehat{\alpha}^d > \alpha^d = \beta^o(\alpha^d)$ for all $t \in [0,1)$. Thus, no one raises an affirmative defense, and, therefore, the pair $\widehat{\alpha}^d$,$\widehat{\alpha}^o$ generates more deterrence, and at a lower imprisonment cost than α^d, α^o (because $\widehat{\alpha}^o < \alpha^o$) as I converges to 0, which implies greater welfare. In symbols

$$W(\widehat{\alpha}^o,\widehat{\alpha}^d) = \int_{(\alpha^d-\widehat{\alpha}^o)}^{\bar{b}}(b-h-\alpha^d\sigma)m(b)db - \sigma[\varphi\alpha^d + M(\alpha^d-\widehat{\alpha}^o)\widehat{\alpha}^o] + \mu$$

$$> \int_{(\alpha^d-\alpha^o)}^{\bar{b}}(b-h-\alpha^d\sigma)m(b)db - \sigma[\varphi\alpha^d + M(\alpha^d-\alpha^o)\alpha^o] + \mu = \lim_{I\to 0}W(\alpha^o,\alpha^d) \text{ iff}$$

$$\int_{(\alpha^d-\alpha^o)}^{(\alpha^d-\widehat{\alpha}^o)}(h-b)m(b)db > \sigma[M(\alpha^d-\alpha^o)[\alpha^d-\alpha^o] - M(\alpha^d-\widehat{\alpha}^o)[\alpha^d-\widehat{\alpha}^o]]$$

which holds, since, $\widehat{\alpha}^o < \alpha^o$. Thus, as $I \to 0$, there are no gains from allowing an affirmative defense which induces a type-1 error of $\alpha^d \in (\alpha^o, \beta^o(\alpha^o))$. Moreover, by assumption, some degree of deterrence is desirable, and, thus, $\alpha^d \leq \alpha^o$ cannot generate any welfare beyond the optimal regime which involves some deterrence. Finally, welfare generated in a regime where $\alpha^d \geq \beta^o(\alpha^o)$ equals welfare generated in a regime where no affirmative defenses are allowed, and, thus, the maximum welfare obtainable in a regime with affirmative defenses converges to the welfare obtainable in a regime without defenses.

8 Appendix C: Justifications Based on Positive Externalities

The previous section supposes that people who have justifications always commit the act, regardless of the specific policies that pertain to the existence of defenses. However, if the act is justified because the person commits an act that is meant to protect others from harm, or to provide others certain benefits, then the fact the act is justified does not necessarily imply that the person will commit the act. In particular, it could be the case that $\underline{b} < 1 < h < \underline{b} + v$; which implies that the person's private benefit may be smaller than the maximum expected sanction (which equals 1), and yet, the act is justified because its commission provides a large benefit of v to third parties. In these circumstances, as previously noted in (13), people will be deterred from committing the act if $b^{**} \equiv \alpha^d - \alpha^o < b$. Note that it is possible for b^{**} to be negative in

the unrealistic case where the probability of being wrongful punished for an act one has not committed (α^o) and the probability of successfully raising an affirmative defense ($1-\alpha^d$) are both very high. In these cases, people who actually perceive costs associated with committing the act may end up committing it. Situations giving rise to these possibilities may not be optimal. Nevertheless, these possibilities cannot immediately be ruled out, and, therefore, to eliminate potential discontinuities in the welfare function, I now assume that $\underline{b} < -1$.

With these modifications, social welfare can be expressed as

$$W(\alpha^o, \alpha^d) = \int_{b^*}^{\bar{b}} (b-h-\beta\sigma)m(b)db + \int_{b^{**}}^{\infty} (b+v-h-\alpha^d\sigma)n(b)db - \sigma[[M(b^*)+N(b^{**})]\alpha^o] \tag{29}$$

which reflects the fact that errors in the determination of valid justifications may have an undesirable deterrence effect.

Differentiating W with respect to α^o and α^d reveals that

$$\frac{\partial W}{\partial \alpha^o} = \left(\frac{\partial \beta}{\partial \alpha^o} - 1\right)\delta_o + \delta_d - \sigma\left[[1-M(b^*)]\frac{\partial \beta}{\partial \alpha^o} + M(b^*) + N(b^{**})\right] \tag{30}$$

$$\frac{\partial W}{\partial \alpha^d} = \frac{\partial \beta}{\partial \alpha^d}\delta_o + \delta_d - \sigma\left[[1-M(b^*)]\frac{\partial \beta}{\partial \alpha^d} + [\varphi - N(b^{**})]\right] \tag{31}$$

where δ_o and δ_d are the marginal deterrence effects defined as follows:

$$\begin{aligned} \delta_o &\equiv (h + (\sigma-1)(\beta-\alpha^o))m(b^*); \text{ and} \\ \delta_d &\equiv (v - h - (\sigma-1)(\alpha^d-\alpha^o))n(b^{**}) \end{aligned} \tag{32}$$

Thus, the optimal solutions can be characterized as follows:

$$\frac{\partial \beta}{\partial \alpha^o} = \frac{\sigma\left[M(b^*)+N(b^{**})\right] + \delta_o - \delta_d}{\delta_o - \sigma[1-M(b^*)]}; \text{ and} \tag{33}$$

$$\frac{\partial \beta}{\partial \alpha^d} = \frac{\sigma\left[\varphi - N(b^{**})\right] + \delta_d}{\delta_o - \sigma[1-M(b^*)]} \tag{34}$$

Inspecting (33) and (34) reveals that variants of the results reported in proposition 1 emerge when policies deter the commission of some justifiable acts. The only difference between (17)-(18) and (33)-(34) is caused by the emergence of the term $\sigma N(b^{**}) - \delta_d$ in the numerator of both conditions, and with opposite signs. Thus, all results in proposition 1 can be restated by adding and subtracting the term $[\sigma N(b^{**}) - \delta_d]$ where necessary.

This term corresponds to the punishment costs saved as a result of the deterrence of some individuals with justifications ($\sigma N(b^{**})$) net of marginal deterrence effects [i.e. δ_d] caused by policy changes. An increase in α^d reduces imprisonment costs, by lowering the frequency with which undeterred people with justifications are punished, whose measure is $\varphi - N(b^{**})$, which is inversely related to $N(b^{**})$. On the other hand, an increase in α^o has the opposite effect: it leads to more frequent punishment of deterred individuals who would

have had justifications, and this group has a measure of $N(b^{**})$. Similarly, the two policy variables have opposite deterrence effects on these individuals: since $b^{**} = \alpha^d - \alpha^o$, an increase in α^d enhances whereas an increase in α^o reduces the deterrence of these individuals. This is what leads to the opposite effects represented in (33) and (34). The over-all impact of these effects on the optimal standards of naturally depends on whether there are benefits or costs associated with the deterrence of these individuals, and, if there are costs associated with deterrence, whether they are greater than the imprisonment costs associated with the deterrence of these individuals. Therefore, it is impossible to make unambiguous statements about how the optimal standard for defenses changes when it is possible for some justifiable acts to be deterred as a result of changes in policies. However, it is possible to identify a few intuitive conditions under which, $\delta_d > \sigma N(b^{**})$, and thus, the deterrability of justifiable acts causes the optimal standard of proof in establishing defenses to be stronger. In particular, this condition can be re-written as:

$$\frac{n(b^{**})}{N(b^{**})}(v - h - (\sigma - 1)(\alpha^d - \alpha^o)) > \sigma \tag{35}$$

Thus, when the gap between the benefit to others and the harm from crime (i.e. $v-h$) is large, and when marginal people with justifications are highly responsive to expected sanctions (i.e. $\frac{n(b^{**})}{N(b^{**})}$ is large), the deterrability of justified acts causes the optimal standard to be stronger.

References

[1] Demougin, D. and C. Fluet "Deterrence versus Judicial Error: A Comparative View of Standards of Proof." Journal of Institutional and Theoretical Economics 161 (2005): 193-206.

[2] Demougin, D. and C. Fluet "Preponderance of Evidence." European Economic Review 50 (2006): 963-976.

[3] Fluet, C. and M. Mungan "The Signal-Tuning Function of Liability Regimes" George Mason Law & Economics Research Paper No. 17-37. (2017).

[4] Garoupa, N. "Explaining the Standard of Proof in Criminal Law: A New Insight" Supreme Court Economic Review (2018, Forthcoming).

[5] Gray, Anthony Davidson. "The presumption of innocence under attack." New Criminal Law Review 20 (2017): 569-615.

[6] Hay, B., and K. Spier. "Burdens of proof in civil litigation: An economic perspective." The Journal of Legal Studies 26 (1997): 413-431.

[7] Hitchler, W.. "Duress as a Defense in Criminal Cases." Virginia Law Review (1917): 519-545.

[8] Kaplow, L. "On the Optimal Burden of Proof" Journal of Political Economy 119 (2011): 1104-1140.

[9] Lando, H. "Prevention of Crime and the Optimal Standard of Proof in Criminal Law" Review of Law and Economics 5 (2009): 33-52.

[10] Lando, H. and M. Mungan "The Effect of Type-1 Error on Deterrence" International Review of Law and Economics 53 (2018): 1-8.

[11] Langlais, E. "Detection Avoidance and Deterrence: Some Paradoxical Arithmetic" Journal of Public Economic Theory 10 (2008): 371-382.

[12] Malik, A. S. "Avoidance, Screening and Optimum Enforcement" RAND Journal of Economics, 21 (1990): 341-353.

[13] Mungan, M. "A Utilitarian Justification for Heightened Standards of Proof in Criminal Trials" Journal of Institutional and Theoretical Economics 167 (2011): 352-370.

[14] Polinsky, M. and S. Shavell "Public Enforcement of Law" in Handbook of Law and Economics 403-454 (A. M. Polinsky & S. Shavell, eds. 2007)

[15] Rizzolli, M. and M. Saraceno. "Better That Ten Guilty Persons Escape: Punishment Costs Explain the Standard of Proof." Public Choice, 155 (2013): 395-411.

[16] Sanchirico, C.W. "Detection Avoidance and Enforcement Theory: Survey and Assessment" *Institute for Law and Economics Research Paper No. 10-29*, University of Pennsylvania Law School, (2010).

[17] Tabbach, A. D. "The Social Desirability of Punishment Avoidance" *Journal of Law, Economics, and Organization* 26 (2010): 265-289.