A Theory of Punishment

Loren K. Fryxell*

January 2, 2024

(Latest Version Here)

This is an older version of the project. A significantly updated draft is coming soon. See slides for new material.

Abstract

I propose a general framework with which to analyze the optimal punishment as deterrence in response to crime. Each criminal act, detected with some probability, generates a random piece of evidence and a consequent probability of guilt for each citizen. I consider a utilitarian planner with no artificial moral constraints. In particular, I assume no upper bound on punishment—such a bound can only rise endogenously from the utilitarian objective. Punishment is pure, i.e., costless. If citizens are expected utility maximizers, a repugnant conclusion is reached—it is optimal to punish *only* with the realization of the *most incriminating* evidence. Allowing for more general behavior yields a weaker but more satisfactory result—optimal punishment is always *non-decreasing* in the quality of evidence.

1 Introduction

A crime has been committed. How should the government respond? There are several types of responses to crime. *Punishment as deterrence* uses punishment to deter potential offenders from committing crimes. *Punishment as retribution* seeks to punish those guilty of crimes because it is "intrinsically good" to do so. *Incapacitation* seeks to temporarily prevent offenders from committing further crimes. *Rehabilitation* seeks to reduce the future crime rate of offenders. *Reparations* seek to compensate victims for the harm caused by the offender. In practice, a particular response may fall into several of these categories. For example, a prison sentence might simultaneously serve as punishment as deterrence, punishment as retribution, incapacitation, and rehabilitation.

*Department of Economics and Global Priorities Institute, University of Oxford (e-mail: loren.fryxell@economics.ox.ac.uk). I wish to thank Jeff Ely, Eddie Dekel, Alessandro Pavan, and Asher Wolinsky for their invaluable comments, discussions, and guidance. All errors are my own.

In this paper, I consider the response of *punishment as deterrence* alone. In particular, I consider a government that can freely inflict pure "pain" or "disutility", which potentially deters crime, but has no other social benefits (including incapacitation effects, rehabilitation effects, or direct benefits from the punishment). There are a few ways we can think about this. First, we may consider the use of electric shocks of various intensities and durations. These cause pure disutility with no social benefits. Second, we may consider the use of hard, but fruitless, labor—for example, digging a hole and filling it in again. Third, we may consider this abstractly as the "disutility" component of a richer sentence, like prison, that includes other effects.

To choose an optimal response to crime, the government must decide on its objective. I consider a utilitarian government—one who cares only about maximizing the unweighted sum of its citizens' utility. Would such a government ever find it optimal to inflict pure disutility upon its citizens in order to deter socially detrimental behavior (crime)? If so, what can we say about the structure of optimal punishment? I construct a simple model of crime and punishment in order to explore these questions. A crime is committed. Evidence is left at the crime scene according to an individual-specific distribution. The government detects the crime with some probability and, if detected, observes the evidence. This gives rise to a posterior probability of guilt for each individual. The government's choice variable is a *punishment plan*—a mapping from evidence to a punishment for each individual. Individual crime rates are determined by the distribution of punishment they will face upon committing the crime (a combination of their evidence distribution and the punishment plan).

Since the seminal work of Becker (1968), economists and legal scholars have analyzed several models of optimal punishment. In this literature, utilitarian governments are (surprisingly) rare. For example, it is common to exclude from welfare the benefits to the offender of committing the crime (e.g., Stigler (1970)). It is also common to consider the disutility of the innocent as worse than the disutility of the guilty (e.g., Siegel and Strulovici (2019, 2021)). Finally, it is almost ubiquitous to assume an upper bound on punishment (e.g., Becker (1968); Stigler (1970); Siegel and Strulovici (2019, 2021)).

One might view an upper bound on punishment as a physical or technological constraint (it is physically impossible to inflict more disutility than some amount). However, it appears that such bounds are more often intended to reflect *moral* constraints. Indeed, maximal punishment is often considered to be life-in-prison or execution, but clearly there exist feasible punishments that an offender would prefer less than these. Moreover, many of these models imply that maximal punishment is, in fact, optimal, which would be especially striking if interpreted as the maximum punishment humanly (not ethically) possible.

Since our objective function is already utilitarian, any additional moral constraint is

a departure from utilitarianism. In other words, the justification that a utilitarian would never punish above some bound is precisely that such punishment is never an *unconstrained* maximum of a utilitarian objective function.

This paper presents two main findings.

The first is that optimal punishment must *always* be non-decreasing in the posterior probability of guilt. This sounds intuitive, but is surprising for two reasons. First, this result holds *no matter* how individuals respond to punishment (indeed, they may even commit more crimes with more punishment). Second, this implies that punishment must ignore the relative ranking of suspects. In particular, for a given posterior probability of guilt, the optimal punishment cannot depend on whether an individual is the "top" suspect.

The second is that, for a large class of individual behavior including all rational (expected utility maximizing) actors, optimal punishment must *only* punish upon the realization of the *most-incriminating* evidence, no matter how rare. This is a repugnant conclusion, and leads me to believe that these seemingly weak assumptions placed on behavior (which include all standard economic models of behavior) does not reflect reality.

The rest of the paper is organized as follows. Section 2 introduces the model. Sections 3 and 4 discuss the main results. Section 5 concludes.

2 Model

Fix some action. Implicitly, this can be thought of as a criminal action, but since this term is somewhat loaded, I would like to think of it simply as any action that results in a net social loss. That being said, for simplicity I will refer to this action as a "crime". To fix ideas, one can think of this crime as murder.

Fix an observable state of affairs. For each state of affairs (the individuals' criminal records, the time of year, etc.), there may be a different set of primitives. Indeed, if an individual was convicted of assault in the past, we may think that he is more likely to commit murder in the future. Moreover, crime rates are generally higher in the summer than in the winter.

For a given crime and a given state of affairs, let I be a finite set of n individuals. Let $c_i > 0$ be the net cost to society of individual i committing the crime. Note that the crime may be beneficial to some individuals (e.g., individual i), but on net the crime is bad for society. Let $\delta_i \in (0, 1]$ be the probability that a crime committed by i is detected. For an action like murder, δ_i is probably close to one, but for other actions it may not be (e.g., running a red light, trespassing, etc.).

Let Φ be the set of all possible things that can be observed at a crime scene. We

will call this "evidence" and will take it to be finite. For example, we may observe "one of Joe's shoes and a hair whose DNA matches to Joe" at a crime scene, so this is an element of Φ. Let $\lambda_i \in \Delta\Phi$ be the distribution over evidence that i leaves behind upon committing a crime.

Let $\kappa_i : \Delta\mathbb{R}_+ \to \mathbb{R}_+$ be a *behavioral response function* for individual i. This function takes as input the distribution of punishments that i will face upon committing a crime and produces as output i's resulting crime rate (actions per unit time, e.g., murders per year). Note that individuals are not modeled as rational actors who choose the number of crimes to commit based on expected costs and benefits. That would be a special case of this model. Rather, I allow for i's behavior to depend *arbitrarily* on the distribution of punishments he faces upon committing the crime.

Consider the following fact.

Fact 1. If there exists any punishment plan, no matter how extreme, that fully deters crime, then this plan is optimal.

Since there is no crime, there is also no punishment—only the threat of punishment. Do we believe that there exists such a punishment plan? No (if we did we would implement it and enjoy a crime-free, punishment-free state). This can be seen as a critique of rational agent models in which individuals commit crimes if and only if their expected benefits outweigh their expected costs. In such a model, there always exists a punishment large enough (when punishment is unbounded or has a sufficiently high bound) such that all individuals will be fully deterred from committing crimes.

Instead, we might believe there is always a chance of "crimes of passion"—crimes that occur in the heat of the moment that aren't subject to rational deliberation (and hence aren't deterred by large punishments). In particular, suppose individuals act rationally most of the time, but sometimes commit crimes of passion. Formally, for some $p \in [0, 1]$, individuals have crime rate $p\kappa_i(\pi_i(\lambda_i)) + (1 - p)\bar{\kappa}_i$, where p is the fraction of the time that they are rational and $\bar{\kappa}_i$ is their crime rate when acting irrationally. Let's call such an individual p-rational.

For $p < 1$, arbitrarily large punishments are almost certainly not optimal, since, even if rational behavior is fully deterred, irrational behavior will persist, requiring these large punishments to be carried out. By modeling individual behavior using behavioral response functions, p-rationality is immediately included as a special case.[1]

Next, I introduce the government's choice variable. The government chooses a

[1]Though, as we will see in Section 4, even p-rationality leads to repugnant conclusions and hence seems behaviorally implausible.

punishment plan $\pi : \Phi \to \Delta(\mathbb{R}^n_+)$ mapping observed evidence to potentially random punishments for each individual. The following are examples of punishment plans:

1. if any crime is detected, punish everyone a fixed amount
2. if any crime is detected, punish each individual a fixed amount independently with probability .5
3. punish anyone a fixed amount who eye-witnesses can identify at the crime scene (for each i, punish for an i-specific subset of Φ)
4. punish anyone a fixed amount who eye-witnesses can identify at the crime scene, and punish anyone for which DNA evidence links them to the crime scene twice as much

Punishment, both for the input to κ_i and the output of π, will be measured in *disutility*. We could equivalently consider a particular type of punishment (e.g., electric shocks or hard labor) and work with expected utilities over these punishments, but since the method of punishment is not relevant to the analysis, it is more transparent to simply abstract away from it. In other words, I am not interested in what gives rise to the disutility, I am just interested in how much disutility in fact arises.[2] Notice that feasible punishment (disutility) is unbounded.[3]

For a given punishment plan π, we can define the government's beliefs as follows. The probability of observing ϕ conditional on i committing the crime (and it being detected) is defined by

$$\mathbb{P}_\pi(\phi \mid i) \equiv \lambda_i(\phi).$$

The total probability that i committed the crime (conditional on detection) is defined by the total number of crimes i commits that are detected (per unit time) divided by the total number of crimes that are detected overall (per unit time),

$$\mathbb{P}_\pi(i) \equiv \frac{\delta_i \kappa_i(\delta_i \pi_i(\lambda_i))}{\sum_{j \in I} \delta_j \kappa_j(\delta_j \pi_j(\lambda_j))}.$$

The joint distribution over $I \times \Phi$ is then given by

$$\mathbb{P}_\pi(\phi \text{ and } i) = \mathbb{P}_\pi(\phi \mid i)\mathbb{P}_\pi(i) = \lambda_i(\phi)\frac{\delta_i \kappa_i(\delta_i \pi_i(\lambda_i))}{\sum_{j=1}^n \delta_j \kappa_j(\delta_j \pi_j(\lambda_j))}.$$

[2]Notice that it may take different amounts of a particular punishment to induce the same amount of disutility across individuals.

[3]This can be interpreted in a few ways. First, we might believe punishment is in fact physically unbounded. Second, we might believe that there is a physical bound on punishment (i.e.. a bound on the disutility an individual can experience), but that we would like to consider a hypothetical world in which punishment is unbounded, and ask if in such a world we would ever use arbitrarily large punishments. Third, we might think that there is a physical bound on punishment, but that it is sufficiently high to never bind in our analysis (that is, if optimal punishment calls for arbitrarily large punishments, punishment at the upper bound will also suffice).

The government chooses a punishment plan to minimize the expected utilitarian loss, defined by

$$L(\pi) \equiv \sum_{i \in I} \kappa_i(\delta_i \pi_i(\lambda_i)) \Bigg(c_i + \delta_i \sum_{\phi \in \Phi} \lambda_i(\phi) \left(\mathbb{E}[\pi_1(\phi)] + \ldots + \mathbb{E}[\pi_n(\phi)] \right) \Bigg).$$

For each crime committed by each i, society experiences a loss of c_i. With probability δ_i, the crime is detected and punishments are carried out and $\sum_{\phi \in \Phi} \lambda_i(\phi)(\mathbb{E}[\pi_1(\phi)] + \ldots + \mathbb{E}[\pi_n(\phi)])$ is the expected total punishment. It will prove useful to alternatively express L by

$$\begin{aligned} L(\pi) &= \sum_{i \in I} \kappa_i(\delta_i \pi_i(\lambda_i)) c_i \\ &\quad + \sum_{j \in I} \delta_j \kappa_j(\delta_j \pi_j(\lambda_j)) \sum_{i \in I} \frac{\mathbb{P}_\pi(i)}{\delta_i} \delta_i \sum_{\phi \in \Phi} \mathbb{P}_\pi(\phi \mid i) \left(\mathbb{E}[\pi_1(\phi)] + \ldots + \mathbb{E}[\pi_n(\phi)] \right) \\ &= \sum_{i \in I} \kappa_i(\delta_i \pi_i(\lambda_i)) c_i \\ &\quad + \sum_{j \in I} \delta_j \kappa_j(\delta_j \pi_j(\lambda_j)) \sum_{\phi \in \Phi} \mathbb{P}_\pi(\phi) \left(\mathbb{E}[\pi_1(\phi)] + \ldots + \mathbb{E}[\pi_n(\phi)] \right). \end{aligned}$$

The left term is the net loss from all crimes committed (detected or not). The right term is the expected total punishment from all crimes that are committed and detected.

3 Theorem 1

It turns out that we can say a fair amount about an optimal punishment plan without placing *any* restrictions on individual behavior. In particular, an optimal punishment plan must be *non-decreasing* in the posterior probability of guilt.

Theorem 1. *An optimal punishment plan π is non-decreasing in the posterior probability of guilt. That is, for all $i \in I$ and $\phi, \phi' \in \Phi$,*

$$\mathbb{P}_\pi(i \mid \phi') > \mathbb{P}_\pi(i \mid \phi) \implies \mathbb{E}[\pi_i(\phi')] \geq \mathbb{E}[\pi_i(\phi)].$$

Proof. Suppose by contradiction that for some $\phi', \phi \in \Phi$, $\mathbb{P}_\pi(i \mid \phi') > \mathbb{P}_\pi(i \mid \phi)$ and $\mathbb{E}[\pi_i(\phi')] < \mathbb{E}[\pi_i(\phi)]$ in an optimal punishment plan π. Consider another punishment plan π' such that $\pi'_i(\phi') = \pi'_i(\phi) \equiv \overline{\pi}'_i$, where $\overline{\pi}'_i$ gives punishment $\pi_i(\phi')$ with probability $\frac{\lambda_i(\phi')}{\lambda_i(\phi') + \lambda_i(\phi)}$ and $\pi_i(\phi)$ with probability $\frac{\lambda_i(\phi)}{\lambda_i(\phi') + \lambda_i(\phi)}$. If i commits a crime, she faces the same punishment

distribution under both plans,[4] so her crime rate is unchanged. The punishment plans are identical for all $j \neq i$, so everyone else's crime rate is unchanged. Hence, $\mathbb{P}_\pi = \mathbb{P}_{\pi'} \equiv \mathbb{P}$.

Note that

$$\begin{aligned}\mathbb{E}[\overline{\pi}_i'] &= \frac{\lambda_i(\phi')}{\lambda_i(\phi') + \lambda_i(\phi)}\mathbb{E}[\pi_i(\phi')] + \frac{\lambda_i(\phi)}{\lambda_i(\phi') + \lambda_i(\phi)}\mathbb{E}[\pi_i(\phi)] \\ &= \frac{\mathbb{P}(\phi')\mathbb{P}(i \mid \phi')}{\mathbb{P}(\phi')\mathbb{P}(i \mid \phi') + \mathbb{P}(\phi)\mathbb{P}(i \mid \phi)}\mathbb{E}[\pi_i(\phi')] + \frac{\mathbb{P}(\phi)\mathbb{P}(i \mid \phi)}{\mathbb{P}(\phi')\mathbb{P}(i \mid \phi') + \mathbb{P}(\phi)\mathbb{P}(i \mid \phi)}\mathbb{E}[\pi_i(\phi)].\end{aligned}$$

The expected utilitarian loss under π' is lower than under π, since

$$\begin{aligned} & L(\pi') < L(\pi) \\ \iff & \sum_{\hat{\phi} \in \Phi} \mathbb{P}(\hat{\phi}) \left(\mathbb{E}[\pi_1'(\hat{\phi})] + \ldots + \mathbb{E}[\pi_n'(\hat{\phi})] \right) < \sum_{\hat{\phi} \in \Phi} \mathbb{P}(\hat{\phi}) \left(\mathbb{E}[\pi_1(\hat{\phi})] + \ldots + \mathbb{E}[\pi_n(\hat{\phi})] \right) \\ \iff & \mathbb{P}(\phi')\mathbb{E}[\pi_i'(\phi')] + \mathbb{P}(\phi)\mathbb{E}[\pi_i'(\phi)] < \mathbb{P}(\phi')\mathbb{E}[\pi_i(\phi')] + \mathbb{P}(\phi)\mathbb{E}[\pi_i(\phi)] \\ \iff & \mathbb{E}[\overline{\pi}_i'] < \frac{\mathbb{P}(\phi')}{\mathbb{P}(\phi') + \mathbb{P}(\phi)}\mathbb{E}[\pi_i(\phi')] + \frac{\mathbb{P}(\phi)}{\mathbb{P}(\phi') + \mathbb{P}(\phi)}\mathbb{E}[\pi_i(\phi)] \end{aligned}$$

Plugging in for $\mathbb{E}[\overline{\pi}_i']$ and doing some algebra, we have

$$\iff (\mathbb{P}(i \mid \phi') - \mathbb{P}(i \mid \phi))\mathbb{E}[\pi_i(\phi')] < (\mathbb{P}(i \mid \phi') - \mathbb{P}(i \mid \phi))\mathbb{E}[\pi_i(\phi)]$$

which is true by assumption, contradicting that π is an optimal punishment plan. ■

Intuitively, the proof states the following. For any punishment plan π which is *not* weakly increasing in the posterior probability of guilt for some individual i, we may construct another punishment plan π' that *is* weakly increasing in i's posterior probability of guilt such that i faces precisely the same distribution over punishments upon committing a crime and which leaves everyone else's punishments unchanged. Thus, each individual's incentives and behavior remain the same. What is different? The expected punishment for j conditional on someone *else* committing a crime is smaller under π' than under π. By switching to π', we punish j less when j is innocent, keeping everything else the same.

This result shows that there is a fundamental relationship between utilitarianism and punishment. No matter how individuals respond to incentives,[5] a utilitarian policy must never prescribe less punishment when more incriminating evidence is observed.

An immediate, but surprisingly powerful, corollary is the following.

[4] Under π, i faces punishment $\pi_i(\phi')$ with probability $\lambda_i(\phi')$ and punishment $\pi_i(\phi)$ with probability $\lambda_i(\phi)$. Under π', i faces punishment $\pi_i(\phi')$ with probability

$$\lambda_i(\phi') \cdot \frac{\lambda_i(\phi')}{\lambda_i(\phi') + \lambda_i(\phi)} + \lambda_i(\phi) \cdot \frac{\lambda_i(\phi')}{\lambda_i(\phi') + \lambda_i(\phi)} = \lambda_i(\phi')$$

and similarly for $\pi_i(\phi)$.

[5] It is even possible that they commit more crimes with more punishment.

Corollary 1. *An optimal punishment plan ignores the order of suspects. That is, it does not depend on the relative probability of guilt across suspects.*

For example, consider some evidence ϕ for which $\mathbb{P}(\text{Joe} \mid \phi) = 1/3$ and the remaining $2/3$ probability is dispersed evenly over the remaining billion individuals. Intuitively, Joe is our top suspect, and in fact the only individual we have substantial evidence against. We are somewhat sure that he is guilty. Suppose we decide to give him a modest punishment in this case.

Now consider another evidence ϕ' for which $\mathbb{P}(\text{Joe} \mid \phi') = 1/3 + \varepsilon$ and $\mathbb{P}(\text{Bob} \mid \phi') = 2/3 - \varepsilon$. Intuitively, Bob is our top suspect. We are quite sure that Bob is guilty, but there is also a reasonable chance that Joe is guilty.

It is never optimal to punish Joe less in the second scenario than the first. In other words, it doesn't matter that Joe is or is not our top suspect. If we decide to punish him when his probability of guilty is $1/3$, then we must punish him no less when his probability of guilt is larger than $1/3$, regardless of our beliefs about the guilt of other individuals like Bob.

4 Theorem 2

Next, we consider a weak assumption on individual behavior which includes as a special case all *rational* (expected utility maximizing) behavior.

Definition 1. A behavioral response function $\kappa_i : \Delta\mathbb{R}_+ \to \mathbb{R}_+$ *respects the mean* if for any $X, Y \in \Delta\mathbb{R}_+$,

$$\mathbb{E}[X] = \mathbb{E}[Y] \implies \kappa_i(X) = \kappa_i(Y).$$

This says that i's behavior depends only on his expected utility (recall punishment is measured in disutility). It doesn't matter how i treats different expected punishments (he could commit more crimes with higher expected punishment), only that all punishments with the same expectation induce the same behavior. All rational and p-rational agents satisfy this condition.

Theorem 2. *If κ_i respects the mean, then an optimal punishment plan π only punishes i when the* most *incriminating evidence is observed. That is, for any $\phi, \phi' \in \Phi$,*

$$\mathbb{P}_\pi(i \mid \phi') > \mathbb{P}_\pi(i \mid \phi) \implies \pi_i(\phi) = 0.$$

Proof. Suppose by contradiction that for some $\phi', \phi \in \Phi$, $\mathbb{P}_\pi(i \mid \phi') > \mathbb{P}_\pi(i \mid \phi)$ and $\mathbb{E}[\pi_i(\phi)] > 0$ in an optimal punishment plan π.[6] Consider another punishment plan π' such

[6]Since we are only considering non-negative punishment, $\mathbb{E}[\pi_i(\phi)] = 0$ implies $\pi_i(\phi) = 0$ with probability 1.

that $\mathbb{E}[\pi_i'(\lambda_i)] = \mathbb{E}[\pi_i(\lambda_i)]$ with $\mathbb{E}[\pi_i'(\phi')] > \mathbb{E}[\pi_i(\phi')] \geq 0$, $\pi_i'(\phi) = 0$, and all other punishments unchanged. If i commits a crime, she faces the same expected punishment under both plans, and since κ_i respects the mean, her crime rate is unchanged. The punishment plans are identical for all $j \neq i$, so everyone else's crime rate is unchanged. Hence, $\mathbb{P}_\pi = \mathbb{P}_{\pi'} \equiv \mathbb{P}$.

Note that

$$\begin{aligned}
&\mathbb{E}[\pi_i'(\lambda_i)] = \mathbb{E}[\pi_i(\lambda_i)] \\
\iff &\sum_{\hat{\phi} \in \Phi} \lambda_i(\hat{\phi}) \mathbb{E}[\pi_i'(\hat{\phi})] = \sum_{\hat{\phi} \in \Phi} \lambda_i(\hat{\phi}) \mathbb{E}[\pi_i(\hat{\phi})] \\
\iff &\mathbb{P}(\phi' \mid i) \mathbb{E}[\pi_i'(\phi')] = \mathbb{P}(\phi' \mid i) \mathbb{E}[\pi_i(\phi')] + \mathbb{P}(\phi \mid i) \mathbb{E}[\pi_i(\phi)] \\
\iff &\frac{\mathbb{P}(i \mid \phi') \mathbb{P}(\phi')}{\mathbb{P}(i)} \mathbb{E}[\pi_i'(\phi')] = \frac{\mathbb{P}(i \mid \phi') \mathbb{P}(\phi')}{\mathbb{P}(i)} \mathbb{E}[\pi_i(\phi')] + \frac{\mathbb{P}(i \mid \phi) \mathbb{P}(\phi)}{\mathbb{P}(i)} \mathbb{E}[\pi_i(\phi)] \\
\iff &\mathbb{P}(\phi') \mathbb{E}[\pi_i'(\phi')] = \mathbb{P}(\phi') \mathbb{E}[\pi_i(\phi')] + \frac{\mathbb{P}(i \mid \phi) \mathbb{P}(\phi)}{\mathbb{P}(i \mid \phi')} \mathbb{E}[\pi_i(\phi)]
\end{aligned}$$

The expected utilitarian loss under π' is lower than under π, since

$$\begin{aligned}
&L(\pi) > L(\pi') \\
\iff &\sum_{\hat{\phi} \in \Phi} \mathbb{P}(\hat{\phi}) \left(\mathbb{E}[\pi_1(\hat{\phi})] + \ldots + \mathbb{E}[\pi_n(\hat{\phi})] \right) > \sum_{\hat{\phi} \in \Phi} \mathbb{P}(\hat{\phi}) \left(\mathbb{E}[\pi_1'(\hat{\phi})] + \ldots + \mathbb{E}[\pi_n'(\hat{\phi})] \right) \\
\iff &\mathbb{P}(\phi') \mathbb{E}[\pi_i(\phi')] + \mathbb{P}(\phi) \mathbb{E}[\pi_i(\phi)] > \mathbb{P}(\phi') \mathbb{E}[\pi_i'(\phi')] \\
\iff &\mathbb{P}(\phi') \mathbb{E}[\pi_i(\phi')] + \mathbb{P}(\phi) \mathbb{E}[\pi_i(\phi)] > \mathbb{P}(\phi') \mathbb{E}[\pi_i(\phi')] + \frac{\mathbb{P}(\phi) \mathbb{P}(i \mid \phi)}{\mathbb{P}(i \mid \phi')} \mathbb{E}[\pi_i(\phi)] \\
\iff &\mathbb{P}(i \mid \phi') > \mathbb{P}(i \mid \phi)
\end{aligned}$$

which is true by assumption, contradicting that π is an optimal punishment plan. ■

Intuitively, the proof proceeds analogously to Theorem 1. For any punishment plan π which inflicts positive punishment on individual i upon observing something less than the most incriminating evidence, we may construct another punishment plan π' that inflicts no punishment upon observing this evidence and positive punishment upon observing more incriminating evidence, such that i faces precisely the same expected punishment upon committing a crime and which leaves everyone else's punishments unchanged. Thus, each individual's incentives and behavior remain the same. What is different? The expected punishment for j conditional on someone *else* committing a crime is smaller under π' than under π. By switching to π', we punish j less when j is innocent, keeping everything else the same.

This result shows that there is a fundamental relationship between utilitarianism, punishment, and behavior which depends only on one's expected utility. If an individual respects the mean, a utilitarian policy must never prescribe them positive punishment when more incriminating evidence exists.

For example, suppose we observe DNA evidence pointing to Joe, ϕ, for which $\mathbb{P}(\text{Joe} \mid \phi) = .99$. Suppose we decide to punishment him in this case. But it is conceivable

that we observe DNA evidence pointing to Joe *and* the testimony of 50 eye-witnesses, ϕ', for which $\mathbb{P}(\text{Joe} \mid \phi') = .999$. So it cannot be optimal to punish Joe in the first case.

I consider this a *repugnant conclusion*. The observation that rational agents can generally be fully deterred, implying arbitrarily large punishments are optimal, helped us to see that agents likely aren't rational in these settings. From here, I proposed p-rationality, which seems to capture the intuition that some crimes (of passion) may be undeterrable, implying full deterrence is not possible and arbitrarily large punishments are likely not optimal.

Here, we observe that individuals that respect the mean can, given any punishment plan, be equivalently deterred by placing some amount of punishment on a single piece of evidence (possibly with arbitrarily small probability of realizing). This implies that optimal punishment must place *all* punishment on the *most* incriminating evidence (since this minimizes the probability of punishing innocent bystanders without changing the incentives of potential offenders). Naturally, if these assumptions were true, then we should want to adopt this policy. Our (presumed) hesitation reveals the absurdity of the assumption of rationality, and indeed, p-rationality, in these contexts. I conclude from this that the ubiquitous and seemingly innocuous assumption of respecting the mean likely does not capture individual behavior in the context of crime.

5 Conclusion

I present a simple model of crime and punishment and analyze the optimal response of a utilitarian government. I find that no matter how individuals respond to punishment, optimal punishment is *non-decreasing* in the posterior probability of guilt. Moreover, if individuals respond to punishment in commonly assumed ways, optimal punishment only punishes upon the realization of the *most-incriminating* evidence, no matter how rare. This leads me to question the standard assumptions placed upon individual behavior in economics within the context of crime.

References

Becker, Gary S. 1968. "Crime and Punishment: An Economic Approach." *Journal of Political Economy*, 76(2): 169–217.

Siegel, Ron, and Bruno Strulovici. 2019. "The Economic Case for Probablity-Based Sentencing."

Siegel, Ron, and Bruno Strulovici. 2021. "Judicial Mechanism Design."

Stigler, George J. 1970. "The Optimum Enforcement of Laws." *Journal of Political Economy*, 78(3): 526–536.