An Economic Theory of Criminal and Civil Law

Loren K. Fryxell

University of Oxford and Global Priorities Institute

loren.fryxell@economics.ox.ac.uk

September 2, 2024

Transatlantic Theory Workshop 2024

Introduction

Today

Today, I will talk about how to optimally respond to crime with punishment

I believe the results can generalize to civil, in addition to criminal, law

Indeed, I will sometimes use acts which are not crimes as examples

I also believe the results can generalize to other responses to crime (there are four other responses: retribution, incapacitation, rehabilitation, reparations)

But today we will just focus on the interpretation of crime and punishment

Motivation

Suppose an individual commits a crime

How should the government respond?

Pure Punishment

Today, we consider the tool of *pure punishment*

That is, the government can freely inflict pure "harm" or "disutility" on its citizens

How should a benevolent government use this tool, if it all, on its citizens?

Goal

The goal is to build a model from first principles with which to understand this question

Importantly, I do not want to bake any moral assumptions into the model—such as an upper bound on punishment

I want to start with a blank slate

If we think that certain policies (like extreme punishments) are morally inadmissible, that result should come out of the model

Primitives

Individuals

Let N be a set of n *individuals* in society

We should think of one of these individuals as "nature" (which will capture the possibility that no crime was in fact committed)

Acts

Let A be the set of all crimes or *acts* that may have occurred

- $a =$ "Alice committed murder"
- $b =$ "Bob and Charlie committed fraud"
- $c =$ "Someone died of natural causes"

Let $G_a \subseteq N$ be the set of individuals who are guilty of committing act a

- $G_a =$ "Alice", $G_b =$ "Bob and Charlie", and $G_c =$ "Nature"

Let $\delta_a \in (0, 1]$ be the probability the government detects that some act was committed conditional on the occurrence of act a

- for murder/an unusual death, this is likely close to one (if a murder/unusual death occurs, we almost always observe that)
- but for other actions this may not be the case (e.g. running a red light, trespassing, fraud)

Evidence

Let E be the set of all possible *evidence* that can be observed upon the government's detection that some act has occurred

- the set of all possible things that can be observed at a crime scene
- for example, we may find "one of Alice's shoes and a hair whose DNA matches to Alice"

Let $L_a \in \Delta(E)$ be the distribution of evidence (likelihood distribution) conditional on the occurrence of act a and the detection that some act occurred

- when Alice commits murder a, conditional on the detection that some act occurred, the government observes evidence e with probability $L_a(e)$

Government's Choice Variable: The Punishment Plan

Let $\pi : E \to \Delta(\mathbb{R}^n_+)$ be a *punishment plan* mapping what the government sees to what the government does

input: evidence observed at the crime scene

output: a distribution over punishments for **every** individual

- remember we are starting with a blank slate
- in principle, we can and might want to punish multiple people given the occurrence of a single act
- if this is morally inadmissible, I want it to come out of the model vs simply being assumed
- this could be a deterministic punishment for each individual, but need not be

Government's Choice Variable: The Punishment Plan

Remember that I am measuring punishment in units of *disutility*

I remain agnostic about what gives rise to this disutility (physical labor, electric shocks, time in prison, grading problem sets, etc.)

A punishment plan just specifies how much *disutility* an individual should receive

Government's Choice Variable: The Punishment Plan

Some Example Punishment Plans

- If we detect any crime, punish everyone a little bit
- Punish everyone whose probability of guilt is greater than .4
- Punish only the top suspect if their probability of guilt is greater than .4
- Punish anyone who eye-witnesses can identify at the crime scene
- Punish anyone whose probability of guilt is greater than .9 severely, whose probability of guilty is between .6 and .9 moderately, and whose probability of guilty is between .3 and .6 lightly

Individual's Response: Behavioral Response Function

Let $R_a : \Delta(\mathbb{R}_+)^{G_a} \to \mathbb{R}_+$ be a *behavioral response function*

input: the distribution over punishments that each individual $g \in G_a$ faces upon committing the act a

output: the resulting crime rate for act a (how many times G_a commit a per year on average)

Individual's Response: Behavioral Response Function

A few things to note:

- I am not modeling individuals as rational actors who decide to commit crimes based on costs and benefits
- That would be a special case of this model
- I am allowing each individual's behavior to depend *arbitrarily* on the distribution of punishment they face conditional on committing the crime

Government's Objective

The "total punishment" j receives per year on average is:

$$\mathbb{E}(\pi_j) = \sum_{a \in A} \delta_a R_a(\delta_a \pi_{G_a}(L_a)) \sum_{e \in E} L_a(e) \mathbb{E}\pi_j(e)$$

Government's Objective

Let A_j be the set of acts for which j is a perpetrator

The "guilty punishment" j receives per year on average (i.e., the total punishment j receives per year on average for acts *they did commit*) is:

$$\mathbb{E}(\pi_j^{\text{guilty}}) = \sum_{a \in A_j} \delta_a R_a(\delta_a \pi_{G_a}(L_a)) \sum_{e \in E} L_a(e) \mathbb{E}\pi_j(e)$$

Government's Objective

The "innocent punishment" j receives per year on average (i.e., the total punishment j receives per year on average for acts *they did not commit*) is:

$$\mathbb{E}(\pi_j^{\text{innocent}}) = \sum_{a \in A \setminus A_j} \delta_a R_a(\delta_a \pi_{G_a}(L_a)) \sum_{e \in E} L_a(e) \mathbb{E}\pi_j(e)$$

$$\text{Note } \mathbb{E}(\pi_j) = \mathbb{E}(\pi_j^{\text{guilty}}) + \mathbb{E}(\pi_j^{\text{innocent}})$$

Government's Objective

The government has preferences $\succeq$ over

$$\left((R_a)_{a \in A}, \left(\mathbb{E}(\pi_j^{\text{guilty}}), \mathbb{E}(\pi_j^{\text{innocent}}) \right)_{j \in N} \right) \in \mathbb{R}_+^{|A|+2|N|} \equiv X$$

- So far we have made no assumptions on anything: likelihood functions, behavioral response functions, nor government preferences
- At this point, we can even have individuals who commit more crimes with more punishment and/or government's who prefer more crimes to less
- As it turns out, I will need to make surprisingly few assumptions to get the main results (just one assumption for today), and both of the above points will be allowed

Primitives Recap

The primitives are $(N, A, E, \pi, (G_a, \delta_a, L_a, R_a)_{a \in A}, \succeq)$, where

- N is a set of individuals
- A is a set of acts
- E is a set of evidence
- $\pi : E \to \Delta(\mathbb{R}^n_+)$ is a punishment plan
- $G_a \subseteq N$ is the set of individuals guilty of committing act a
- $\delta_a \in (0, 1]$ is the probability of detecting act a conditional on its occurrence
- $L_a \in \Delta(E)$ is the likelihood distribution over evidence conditional on the occurrence and detection of act a
- $R_a : \Delta(\mathbb{R}_+)^{G_a} \to \mathbb{R}_+$ is a behavioral response function for a
- $\succeq$ is the government's preference relation over X

Government Beliefs

For any specification of the primitives, we may define a probability space $(A \times E, \mathcal{F}, \mathbb{P})$ representing the beliefs of the government

First,

$$\mathbb{P}(e \mid a) = L_a(e) \quad \text{and} \quad \mathbb{P}(a) = \frac{\delta_a R_a(\delta_a \pi_{G_a}(L_a))}{\sum_{b \in A} \delta_b R_b(\delta_b \pi_{G_b}(L_b))}$$

The joint distribution over $A \times E$ is then given by

$$\mathbb{P}(e \text{ and } a) = \mathbb{P}(e \mid a)\mathbb{P}(a) = L_a(e) \frac{\delta_a R_a(\delta_a \pi_{G_a}(L_a))}{\sum_{b \in A} \delta_b R_b(\delta_b \pi_{G_b}(L_b))}$$

Basic Model

Basic Model vs Universal Model

In the (upcoming) paper, I have a basic model and a universal model

The universal model considers the fully general case, as defined by the model primitives we discussed

The basic model applies the analysis to a single criminal act, like murder

The basic model helps to elucidate the intuition of the results

The results do generalize to the universal model, with added nuance

Basic Model

Suppose the set of acts consists simply of a single "action" (e.g, murder, theft, or fraud) taken by each individual

Hence, the set of acts can simply be indexed by the individual who committed it

$$A = N \quad \text{and} \quad G_j = \{j\} \text{ for each } j$$

For example, suppose the action we are studying is "murder". Then the set of possible acts is

$$A = \{\text{"Alice did it"}, \text{"Bob did it"}, \text{"Charlie did it"}\}$$

Government's Preferences

As promised, we will only make *one* assumption about the primitives

Recall that the government has preferences $\succeq$ over

$$\left((R_a, \delta_a)_{a \in A}, \left(\mathbb{E}(\pi_j^{\text{guilty}}), \mathbb{E}(\pi_j^{\text{innocent}}) \right)_{j \in N} \right) \in \mathbb{R}_+^{2|A|+2|N|} \equiv X$$

Assumption 1. The government's preference $\succeq$ is strictly decreasing in $\mathbb{E}(\pi_j^{\text{innocent}})$ for each $j \in N$.

Effectively No Assumptions

All we assume is that *the government doesn't like to punish j when j is innocent*

We haven't said anything about

- how the government feels about punishing the guilty (could like it or dislike it, all else equal)
- how the government feels about crime rates (could even prefer more crime)
- how individuals respond to punishment (could even respond to FOSD shifts in punishment with more crime)

Theorem 1

Theorem 1

Theorem 1. An optimal punishment plan π is non-decreasing in the posterior probability of guilt. That is, for any individual $j \in N$ and any evidence $e_1, e_2 \in E$,

$$\mathbb{P}(j \mid e_1) > \mathbb{P}(j \mid e_2) \implies \mathbb{E}\pi_j(e_1) \geq \mathbb{E}\pi_j(e_2).$$

This is surprising

At first, this seems natural. But it is surprising for two reasons

1. No assumptions. All we assumed was the government doesn't like punishing j when j is innocent
2. It implies, at least in some cases, we should punish multiple people for the same act (even though we know only one person committed it)

1. No assumptions

How can this be?

There are many punishment plans that give rise to the *exact same* distribution of punishment for every individual

Among these, those that are monotonic for j **minimize the total punishment to j when j did not commit the crime**

One way to think about this

You committing a crime has three costs:

1. the direct cost to society of the crime
2. the cost to you in the form of the expected punishment you could receive
3. the cost to everyone else in the form of the expected punishment they could receive for being punished for a crime they did not commit

Call the last effect the *collateral damage* of committing a crime

Making j's punishment monotonic *minimizes the collateral damage everyone else imposes on* j, holding the punishment distribution each individual faces constant

2. Punishing multiple people

Corollary 1. An optimal punishment plan ignores the relative ordering of the suspects.

For example:

Scene 1. $\mathbb{P}(\text{Alice} \mid e_1) = 1/3$ and the remaining $2/3$ probability is dispersed evenly over the other billion individuals

Suppose we decide to give Alice a modest punishment in this case: say a week in jail and a remedial course

Scene 2. $\mathbb{P}(\text{Alice} \mid e_2) = 1/3 + \varepsilon$ and $\mathbb{P}(\text{Bob} \mid e_2) = 2/3 - \varepsilon$

> *It is never optimal to punish Alice less in Scene 2 than Scene 1.*

Takeaways

Without making any assumptions other than that the government dislikes punishing innocent people, the optimal policy is **monotonic** in the posterior probability of guilt

This implies that we should punish j relative to her *probability of guilt* rather than punishing only the individual who is *most likely to be guilty*, and hence that we may sometimes punish multiple people for the same crime

Moreover, j herself prefers this policy of receiving punishment even when there are other more likely suspects—the reason this is optimal is precisely that using a monotonic policy for j punishes j less when she is innocent

Theorem 2

Behavioral Assumption

We will now make a very standard assumption about individual behavior R_j and see where it leads

I think some foreshadowing might be useful

In my opinion, it leads to a "reductio"

The fact that the optimal plan is so clearly not something we want in practice (at least in many contexts), implies that some assumption—in my view, this behavioral assumption—must be importantly incorrect in these settings

Behavioral Assumption

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

$$\mathbb{E}(X) = \mathbb{E}(Y) \implies R_j(X) = R_j(Y).$$

- Remember that I am measuring punishment in **utils**, not in hours of labor or duration in prison
- This is **not** a risk neutrality assumption
- All expected utility agents satisfy this condition
- And many more, e.g., an individual who maximizes expected utility, but also commits "crimes of passion" with some probability p

Theorem 2

Theorem 2. If R_j respects the mean, then an optimal punishment plan π only punishes j when the *most* incriminating evidence is observed. That is, for any individual $j \in N$ who respects the mean and any evidence $e_1, e_2 \in E$,

$$\mathbb{P}(j \mid e_1) > \mathbb{P}(j \mid e_2) \implies \pi_j(e_2) = 0.$$

Theorem 2: An Example

We observe e, DNA evidence pointing to Joe, and $\mathbb{P}(\text{Joe} \mid e) = .99$

We decide to give Joe some punishment in this case

But it is conceivable to have observed e', DNA evidence pointing to Joe **and** the testimony of 50 eye-witnesses, which gives $\mathbb{P}(\text{Joe} \mid e') = .999$

So it cannot be optimal to punish Joe upon merely observing e

How can this be?

Same logic as before

There are many punishment plans that give rise to the *exact same* ~~distribution of punishment~~ expected punishment for every individual

Among these, those that ~~are monotonic for j~~ place all punishment on the most incriminating evidence for j **minimize the total punishment to j when j did not commit the crime**

So again, j herself prefers us to use such a policy because all it is doing is reducing her punishment when she is innocent

Upper bound on punishment

This policy requires giving arbitrarily large punishments for arbitrarily small probability events

I made a point of not assuming an upper bound on punishment (because if such a bound is optimal, I wanted this to arise from the model)

But at this point you might be wondering if punishments that high are even *physically* possible

(or you might also just be wondering what happens when you have a morally-imposed upper bound)

Corollary 2

Suppose there is an upper bound on punishment $\pi^{\max}$

Corollary 2. For any $\pi^{\max} > 0$ and individual $g \in N$ who respects the mean and evidence $e_1, e_2 \in E$,

$$\mathbb{P}_\pi(g \mid e_1) > \mathbb{P}_\pi(g \mid e_2) \quad \text{and} \quad \mathbb{E}\pi_g(e_1) < \pi^{\max} \implies \pi_g(e_2) = 0.$$

That is, all punishment must be "concentrated at the top", and this punishment must be maximal (except for the least incriminating evidence with positive punishment)

Corollary 2: Intuition

Intuitively, we have two tools to give someone some amount of expected punishment:

1. the severity of punishment
2. the probability of punishment

Corollary 2 says that, for any individual who respects the mean, we should only use the latter to adjust the punishment

For example, murder and stealing bubblegum should have the same severity of punishment—the maximum punishment possible

If we want to punish less for stealing bubblegum, then we should simply lower the probability of this punishment by requiring a lower burden of proof

Takeaways

In the context of crime,

assuming that individuals are expected utility maximizers (or, more generally, respect the mean)

implies that the optimal policy always concentrates **maximal punishment on the most incriminating evidence(s)**

Moreover, j herself prefers it this way—the reason this is optimal is precisely that using this policy for j punishes j less when she is innocent, holding everything else constant

Takeaways

I view this as a "reductio" that it is a good behavioral assumption that, in the context of crime, individuals are expected utility maximizers

In particular, this result casts a doubtful eye on the fact that expected utility maximizers are fully responsive to very low probabilities of very high punishments

(Indeed, we know this to be true from experimental work)

Conclusion

Conclusion

I presented a general model of crime

We assumed nothing about the government's preference other than that it prefers to punish the innocent less, all else constant

We assume nothing about individual behavior and found that optimal punishment is **monotonic** in the posterior probability of guilt

This implies that we should punish relative to the *probability of guilt* rather than punishing only the *most likely to be guilty*, and hence that we may sometimes punish multiple people for the same crime

Conclusion

We then made the standard assumption that individuals maximize their expected utility and found that optimal punishment concentrates **maximal** punishment on the most incriminating evidence(s)

I find this to be a reductio for the assumption that individuals behave in accordance with **expected utility maximization** in the context of crime

Thank You!

Questions, Comments, or Concerns?