An Economic Theory of Wrongs

Loren K. Fryxell

University of Oxford and Global Priorities Institute

loren.fryxell@economics.ox.ac.uk

October 22, 2024

Oxford Theory Workshop

Introduction

Motivation

Suppose some criminal or civil wrong has been committed.

How should the government respond?

Four Types of Responses

There are four types of responses to criminal and civil wrongs:

1. Punishment
   - Could be as retribution or for deterrence or both
2. Incapacitation
   - Disabling an individual from committing further crimes
3. Rehabilitation
   - Decreasing the chance an individual commits further crimes
4. Reparations
   - Compensating the victim(s) for harms caused

Focus on Punishment

The framework I present applies to all four types

But in this talk, I will focus primarily on the punishment interpretation

Interpreting the framework for incapacitation, rehabilitation, and reparations is simple, but nevertheless not trivial

I'd love to discuss it more, but we won't have time today

Pure Punishment

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 assume secondary moral principles (like a moral upper bound on punishment, or punishing only one person for one crime) which I think ought to be derived from primary moral principles (like punishing the innocent is bad)

I want to start with the fundamentals

If we think that certain types of policies are morally inadmissible, that result should come out of the model

Outline

1. Build the model
2. Describe the results (3 Theorems, 2 Corollaries)

Primitives

Individuals

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

One of these individuals is “nature”, which captures the possibility that no crime was in fact committed

E.g., $N = \{\text{Alice, Bob, Charlie, Nature}\}$

Acts

It will be useful to distinguish between "actions" and "acts"

An *action* is a description of something someone could do, or several people could do together

- e.g., murder, fraud, trespassing, running a red light

An *act* is simply an action plus who committed it

- e.g., Alice commits murder, Bob and Charlie commit murder, David and Ester commit fraud, Eric runs a red light

Acts

Let A be the set of all acts that can occur

- $a =$ "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, δ_a is probably close to one
- for fraud, trespassing, and running a red light, δ_a is probably further from one

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 $L_a(e)$

The Punishment Plan

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

input: evidence observed at the crime scene

output: a (possibly random) punishment for **every** individual

- remember we are starting from a blank slate
- in principle, we might want to punish multiple people given the occurrence of a single act
- conditional on a particular evidence realization, we might also want to give people stochastic punishments
- (if either is morally inadmissible, I want it to come out of the model vs simply being assumed)

The Punishment Plan

I will measure 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

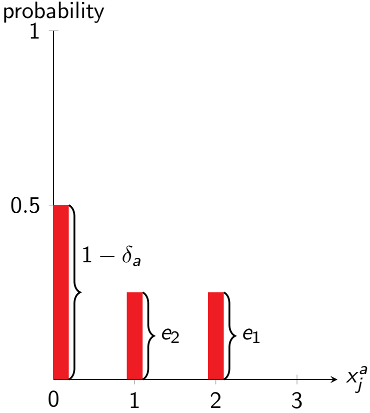
The Punishment Plan

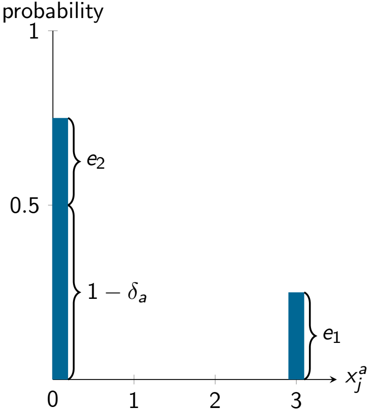
Some Example Punishment Plans

- 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 and whose probability of guilty is between .6 and .9 lightly

Useful Object: Punishment Distribution

Let $x_j^a = (\delta_a, x_j(L_a); (1-\delta_a), 0)$ be the punishment distribution individual j faces upon committing act a





Behavioral Response Function

I don't want to immediately assume that individuals are rational actors who decide to commit crimes based on costs and benefits

This will be a special case of the model

Rather, I want to model behavior also starting from a clean slate

Surprisingly, we will be able to say quite a lot without assuming *anything* about individuals' behavioral responses

Later, we will make some assumptions about behavior and see where it leads

Behavioral Response Function

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

input: the punishment distribution x_j^a each individual $j \in G_a$ 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)

Behavioral Response Function

Note that:

- Each G_a's behavior depends *arbitrarily* on the distribution of punishment each of them face conditional on committing act a
- "Rational" (expected utility maximizing) behavior is a *special case* of this model

Government's Preferences

I want to write down all the fundamentals that a government might plausibly care about based on primary moral principles

The "total 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:

$$\bar{x}_j^{\text{innocent}} = \sum_{a \in A \setminus A_j} \delta_a R_a((x_g^a)_{g \in G_a}) \sum_{e \in E} L_a(e) \mathbb{E} x_j(e)$$

The government has preferences $\succeq$ over all tuples of the form

$$\Big((R_a)_{a \in A},\ \big((x_j^a)_{a \in A_j},\ \bar{x}_j^{\text{innocent}} \big)_{j \in N} \Big)$$

Examples of Government Preferences

Utilitarian

$$\max_x \sum_{j\in N} \Big[-\sum_{a\in A_j} \delta_a R_a((x_g^a)_{g\in G_a}) \sum_{e\in E} L_a(e)\mathbb{E}x_j(e)$$
$$-\sum_{a\in A\setminus A_j} \delta_a R_a((x_g^a)_{g\in G_a}) \sum_{e\in E} L_a(e)\mathbb{E}x_j(e)\Big] - \Big[\sum_{a\in A} c_a R_a((x_g^a)_{g\in G_a})\Big]$$

where c_a is the social damage of act a.

Examples of Government Preferences

Weighted Utilitarian

$$\max_x \sum_{j \in N} \Big[- \sum_{a \in A_j} \delta_a R_a((x_g^a)_{g \in G_a}) \sum_{e \in E} L_a(e) \lambda_j \mathbb{E} x_j(e)$$
$$- \sum_{a \in A \setminus A_j} \delta_a R_a((x_g^a)_{g \in G_a}) \sum_{e \in E} L_a(e) \mathbb{E} x_j(e) \Big] - \Big[\sum_{a \in A} c_a R_a((x_g^a)_{g \in G_a}) \Big]$$

for any $\lambda_j \in [0,1]$, where c_a is the social damage of act a.

Examples of Government Preferences

Retributive

$$\max_x \sum_{j\in N}\Big[\sum_{a\in A_j}\delta_a R_a((x_g^a)_{g\in G_a})\sum_{e\in E}L_a(e)u_j^a(x_j(e))$$
$$-\sum_{a\in A\setminus A_j}\delta_a R_a((x_g^a)_{g\in G_a})\sum_{e\in E}L_a(e)\mathbb{E}x_j(e)\Big]-\Big[\sum_{a\in A}c_a R_a((x_g^a)_{g\in G_a})\Big]$$

for any concave u_j^a which is increasing up to $x_j^{a,\,\text{ideal}}$ and decreasing thereafter with slope no lower than -1.

Primitives Recap

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

- N is a set of individuals
- A is a set of acts
- E is a set of evidence
- $x : E \to \Delta(\mathbb{R}^n_+)$ 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

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

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

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

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

Assumptions

Universal vs Single Action Model

In the universal model,

1. the government has uncertainty about
   - *who* committed the action (Alex, Bob, or Nature) **and**
   - *what* action was committed (e.g., involuntary manslaughter, voluntary manslaughter, or murder)
2. multiple people can commit the same action as a group

These are important features of the criminal justice system, and it is important that our model includes them

That said, it is helpful to look at a basic version of this model to gain some intuition: the special case of a *single action*

Single Action Model

Suppose there is

1. a single action (equivalently, the government can tell which action was committed when inspecting the crime scene, so that the only uncertainty is about who committed the action)
2. a single criminal (the government knows only one individual, or nature, committed the action)

In this special case, the set of *acts* (action + who did it) is just the set of individuals: $A = N$

For ease of exposition, I will present all results within the single action model

All results do generalize appropriately to the universal model, with some important lessons

Single Action Model

Suppose there is just a single action, call it *murder*

The set of acts is then just the set of individuals

$$A = \{a \text{ (Alice did it)}, b \text{ (Bob did it)}, c \text{ (Nature did it)}\}$$

Government's Preferences

We will only make *one assumption* on the primitives

Recall that the government has preferences $\succeq$ over all tuples of the form

$$\Big((R_a)_{a\in A},\, \big((x_j^a)_{a\in A_j},\, \bar{x}_j^{\text{innocent}}\big)_{j\in N}\Big)$$

Assumption 1. Holding all else constant, the government's preference $\succeq$ is strictly decreasing in $\bar{x}_j^{\text{innocent}}$ 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)

Notice that all the example government preferences we went over in the beginning are allowed (and many more)

Results

Theorem 1

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

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

This is surprising

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

1. We made almost no assumptions. All we assumed was the government doesn't like punishing j when j is innocent.
2. It implies that an optimal punishment plan depends *only on* the posterior probability of guilt. In other words, we can focus only on punishment plans which map *posteriors for j* to punishments for j.
3. It also implies that, at least in some cases, we should punish multiple people for the same act (even if we know only one person committed it)

1. Almost 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

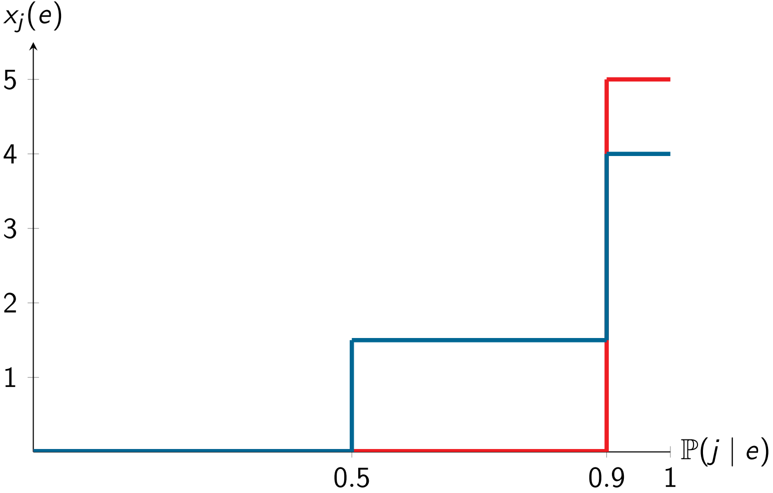
You committing a crime has three costs:

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. Punishment depends only on posteriors



3. Punishing multiple people

Implication. An optimal punishment plan ignores the relative ordering of the suspects.

For example:

Scenario 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

Scenario 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 Scenario 2 than in Scenario 1.*

Discussion

1. You might think punishing multiple people is repugnant, but Alice herself prefers this plan

By switching from a non-monotonic to a monotonic plan for Alice, Alice faces the same punishment distribution upon committing the crime, and the only thing that changes is she gets (wrongly) punished less often when someone *else* commits the crime

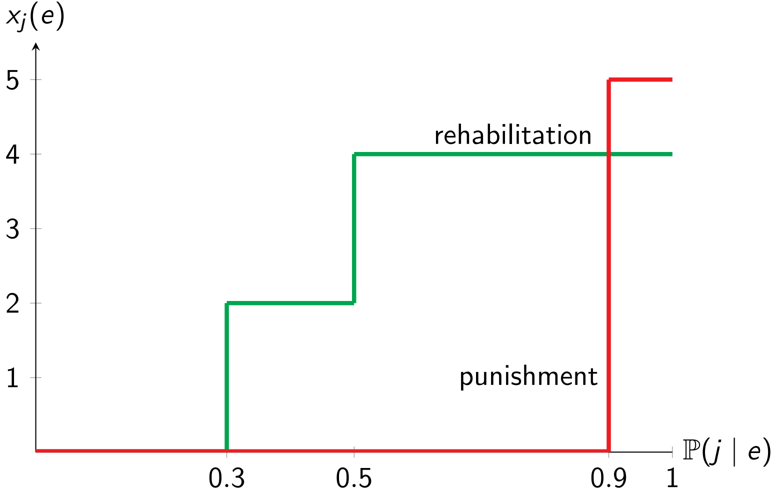
Discussion

2. You might think that the optimal punishment plan won't punish people with posterior probability of guilt $< .5$, even lightly, so we will never end up punishing multiple people for the same crime

I think this is a very reasonable view. That said, for the other treatments (namely, rehabilitation and reparations), I think it's very plausible that we would want to treat an individual who's posterior probability of guilt is, say .4

And, interestingly, this implies that we might send multiple people to rehabilitation for the same crime

Rehabilitation and Punishment for Sexual Assault



Upcoming

In the upcoming two theorems (and two corollaries), we will dig in and see if we can say more

It turns out it will be useful to tackle retributive and non-retributive theories separately

Retribution

A Retributive Theory of Justice

A primary view in criminal justice is that of *retribution* or *desert* (the condition of deserving something)

Under this view, an individual who is guilty of a crime *deserves* to be punished, irrespective of possible side-effects (e.g., deterrence, incapacitation, rehabilitation, reparations to the victim)

A Retributive Theory of Justice

This is not my view (and one of my main reasons for writing this paper is to explore / make the case for a non-retributive approach to crime)

But I've done the analysis under this view to understand it better and for completeness

It turns out to be extremely simple (and, in my view, alluringly elegant)

A Retributive Theory of Justice

We already know from Theorem 1 that

1. the optimal punishment x_j^* depends only on the posterior probability that j committed the action $\mathbb{P}(j \mid e)$
2. x_j^* is non-decreasing in $\mathbb{P}(j \mid e)$

Hence, we can restrict attention to punishment plans which are just a function of $\mathbb{P}(j \mid e)$ for each j

Since a retributivist does not care about the effect of punishment on behavior (through deterrence, incapacitation, or rehabilitation), we do not even have to model behavioral responses

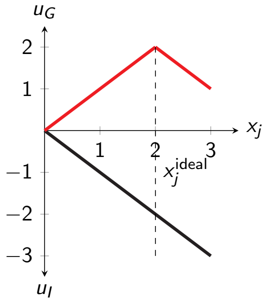
We simply care intrinsically about *punishing the guilty* and *not punishing the innocent*

The Government's Problem

For each j, the government seeks to

$$\max_{x_j \geq 0} \mathbb{P}(j \mid e) u_G(x_j) - (1 - \mathbb{P}(j \mid e)) x_j$$

where $u_G(x_j)$ is increasing up until some x_j^{ideal} and decreasing after

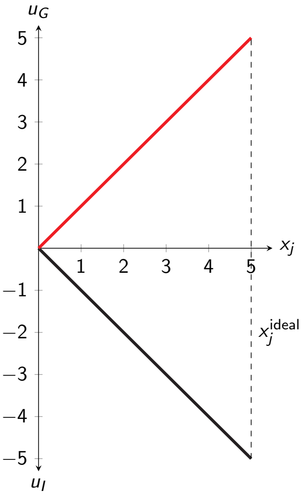


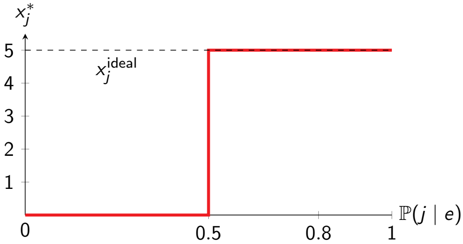
The Solution

Theorem 2. If the government has retributivist preferences, then for each j,

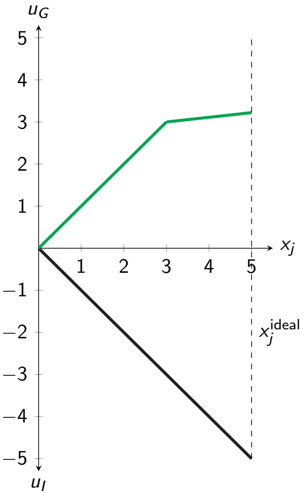
$$x_j^*(e) \begin{cases} = 0 & \text{if } \mathbb{P}(j \mid e) < \frac{1}{1+u_G'(0^+)} \\ \in [0, x_j^{\text{ideal}}] & \text{if } \mathbb{P}(j \mid e) \in [\frac{1}{1+u_G'(0^+)}, \frac{1}{1+u_G'(x_j^{\text{ideal}-})}] \\ = x_j^{\text{ideal}} & \text{if } \mathbb{P}(j \mid e) > \frac{1}{1+u_G'(x_j^{\text{ideal}-})} \end{cases} .$$

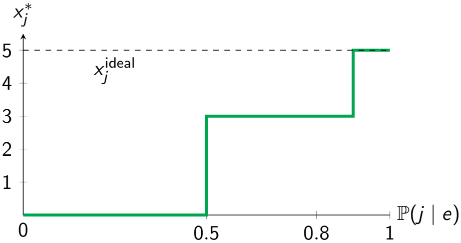
The Solution in Pictures



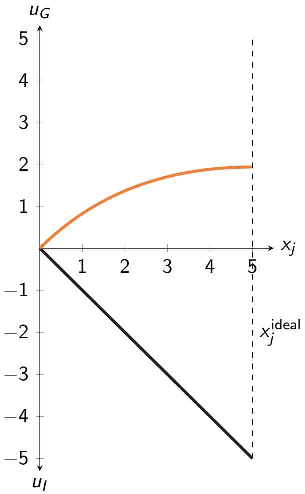


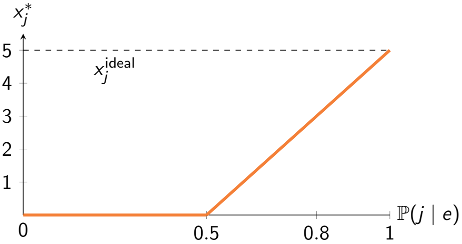
The Solution in Pictures



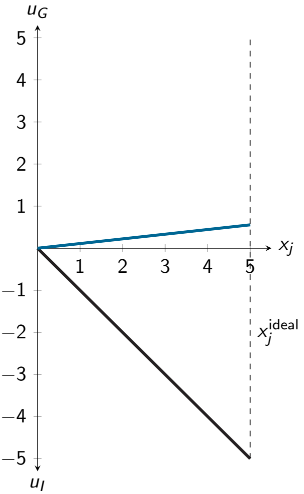


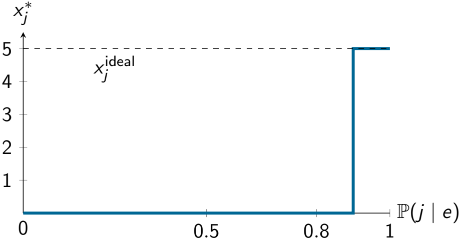
The Solution in Pictures





The Solution in Pictures





Non-Retribution

Non-Retribution

Now consider a government with non-retributive preferences

Hence, we can no longer ignore behavioral response functions as we did in the retributive case

Behavioral Assumption

We will now make some assumptions about individual behavior R_j and see where it leads

1. We will assume individual behavior is fully rational (EU) and notice that the result is clearly undesirable
   - In my view, this helps elucidate how/where individual behavior is not fully rational (EU) in the context of crime
2. We will then assume an upper bound on punishment and see how this changes the result, but does not solve the undesirability
3. We will then assume a lighter form of rationality, which seems potentially reasonable to me, and see that it leads to strong but reasonable conclusions

Behavioral Assumption

Definition. A behavioral response function $R_j : \Delta(\mathbb{R}_+) \to \mathbb{R}_+$ *respects the mean* if for any $X_1, X_2 \in \Delta(\mathbb{R}_+)$,

$$\mathbb{E}(X_1) = \mathbb{E}(X_2) \implies R_j(X_1) = R_j(X_2).$$

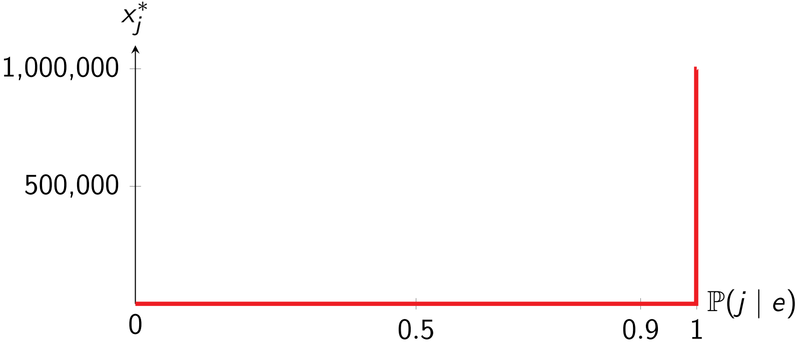
- Remember that I am measuring punishment in **utils**, not in hours of labor or duration in prison
- This is **not** a risk neutrality assumption
- All expected utility agents satisfy this condition
- And many more, e.g., an individual who maximizes expected utility, but also commits "crimes of passion"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

Theorem 3

Theorem 3. If R_j respects the mean, then an optimal punishment plan x 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 $e_1, e_2 \in E$,

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

Theorem 3: Illustration



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, we might be hitting the boundary of what is *physically* possible (not just morally possible)

And of course analyzing this case tells us what happens with a morally-imposed upper bound as well

Corollary 1

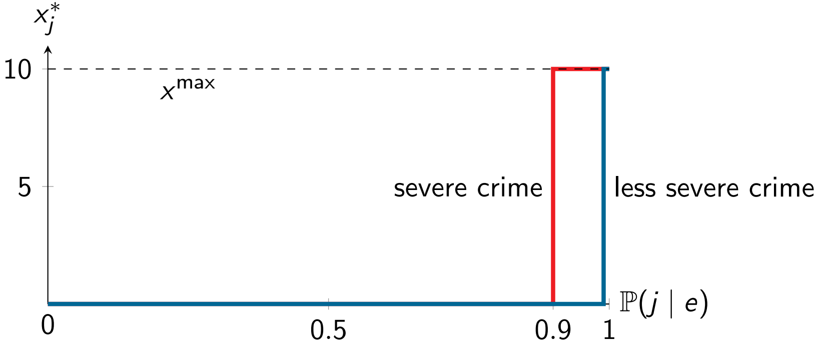
Suppose there is an upper bound on punishment $x^{\max}$

Corollary 1. For any $x^{\max} > 0$, any individual $j \in N$ who respects the mean, and any evidence $e_1, e_2 \in E$,

$$\mathbb{P}(j \mid e_1) > \mathbb{P}(j \mid e_2) \quad \text{and} \quad \mathbb{E}x_j(e_1) < x^{\max} \implies x_j(e_2) = 0.$$

That is, all punishment must be "concentrated at the top" and (if the upper bound is binding) all punishment is maximal. To increase/decrease the severity of punishment, we simply increase/decrease its likelihood.

Corollary 1: Illustration



Respecting the Mean Up To $p^{\min}$

Definition. A behavioral response function $R_j : \Delta(\mathbb{R}_+) \to \mathbb{R}_+$ *respects the mean* if for any $X_1, X_2 \in \Delta(\mathbb{R}_+)$,

$$\mathbb{E}(X_1) = \mathbb{E}(X_2) \implies R_j(X_1) = R_j(X_2).$$

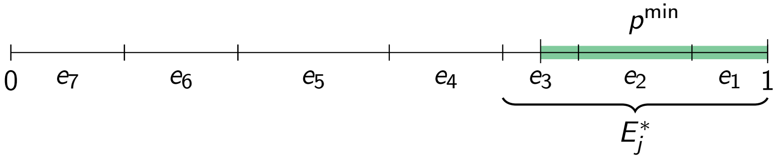
Definition. A behavioral response function $R_j : \Delta(\mathbb{R}_+) \to \mathbb{R}_+$ *respects the mean up to p^{min}* if for any $X_1, X_2 \in \Delta(\mathbb{R}_+)$,

$$\mathbb{E}(X_1) = \mathbb{E}(X_2) \text{ and } X_1 \mid X_2(p^{\min}) \geq_{\text{FOSD}} X_2 \mid X_2(p^{\min}) \\ \implies R_j(X_1) \leq R_j(X_2),$$

where

$$X_2(p^{\min}) = \{x : \mathbb{P}(X_2 > x) < p^{\min}\}.$$

The Burden of Proof



$b(p^{\min})$ is the highest burden of proof such that the evidence which incriminates j at or above that burden of proof occurs with at least probability $p^{\min}$. In this case, $b(p^{\min}) = \mathbb{P}(j \mid e_3)$

E_j^* is the set of evidence which incriminates j at or above the burden of proof $b(p^{\min})$

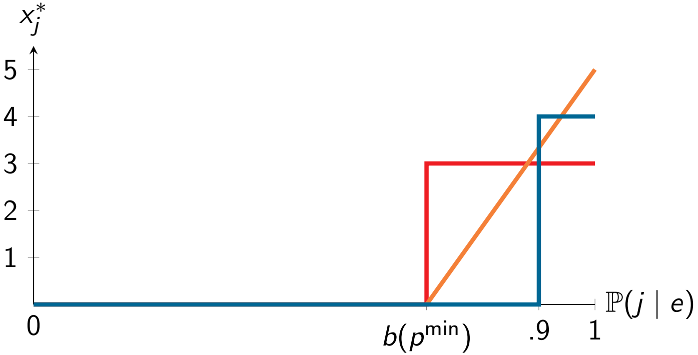
Corollary 2

Corollary 2. If R_j respects the mean up to $p^{\min} > 0$, then an optimal punishment plan x punishes j only upon observing the most incriminating set of evidence E_j^* which occurs with probability at least $p^{\min}$,

$$e \notin E_j^* \implies x_j(e) = 0.$$

That is, we should set the *burden of proof* for punishment at the highest level at which the probability that the punishment materializes is at least $p^{\min}$ when the crime is committed.

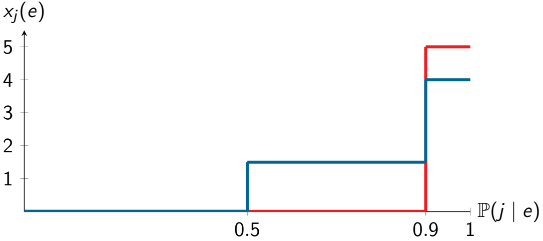
Corollary 2: Illustration



Conclusion

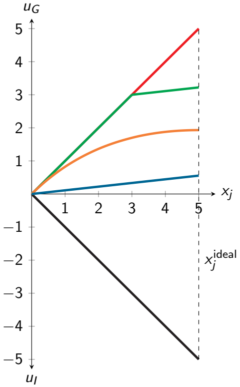
Recap

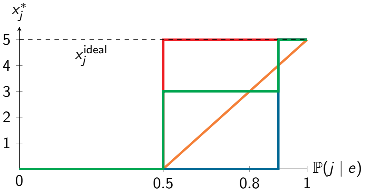
Theorem 1. Suppose the government prefers to treat the innocent less, all else equal (true of both retributive and non-retributive governments). Then the optimal treatment plan for individual j depends only on $\mathbb{P}(j \mid e)$ and, moreover, is non-decreasing in $\mathbb{P}(j \mid e)$.



Recap

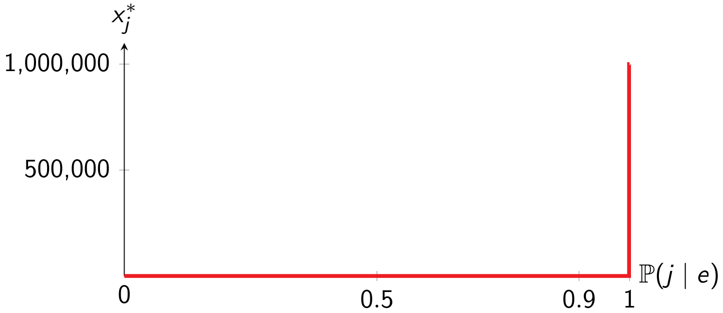
Theorem 2. Suppose the government is retributive and prefers to treat the innocent less. Then the optimal treatment plan for individual j is increasing from zero to x_j^{ideal} and the shape of x^* follows straightforwardly from the utility the government ascribes to punishing the guilty.





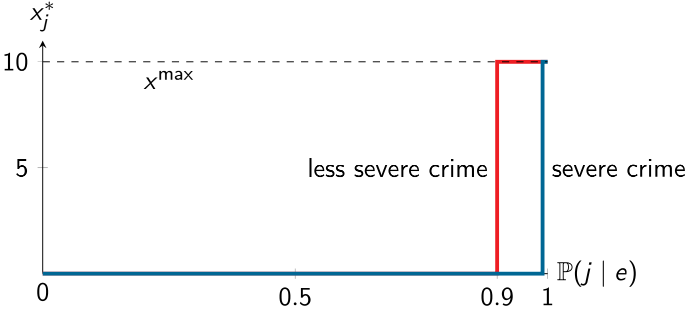
Recap

Theorem 3. Suppose the government is non-retributive and prefers to treat the innocent less. If individuals are fully rational (EU) with respect to crime, then an optimal treatment plan places all treatment on the single most incriminating evidence.



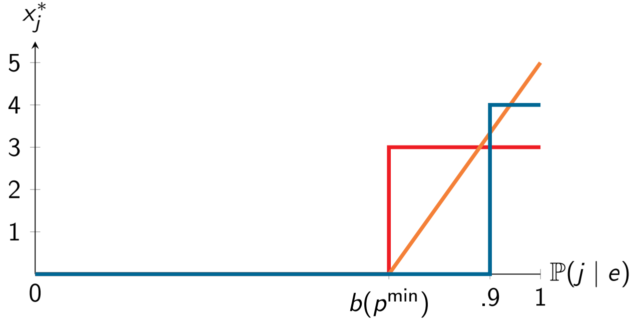
Recap

Corollary 1. Suppose the government is non-retributive and prefers to treat the innocent less. If individuals are fully rational (EU) with respect to crime and there is an upper bound on treatment, then an optimal treatment plan places maximal treatment "at the top".



Recap

Corollary 2. Suppose the government is non-retributive and prefers to treat the innocent less. If individuals are fully rational (EU) only with respect to treatments which are sufficiently likely (occur with probability $\geq p^{\min}$), then an optimal treatment plan places all treatment the most incriminating set of evidence E_j^* which occurs with probability at least $p^{\min}$.



Thank You!

Questions, Comments, or Concerns?