Machine Learning and Deterrence

Daniel L. Chen[*] Henrik Sigstad[†]

July 15, 2022

Abstract

What is the impact of artificial intelligence on the legal system? In a general deterrence model, we show that although machine learning might optimally reduce type I and type II errors, basing legal decisions on machine predictions can undermine incentives to abide by the law. We discuss under which assumptions machine learning methods can be adapted to obtain optimal deterrence. In a planned empirical application on a corpus of 14 million Brazilian labor lawsuits, we assess the amount of statistical discrimination under various machine learning decision rules.

1 Introduction

Artificial intelligence (AI), fueled by machine learning, rapidly transforms society. One sector where AI both shows great promise and spurs controversy is the legal sector. Lawyers and judges have, until recently, largely been insulated from automation. Now, machine learning algorithms can analyze millions of pages of legal texts in minutes and outperform human lawyers in predicting outcomes of certain court cases.[1] Predictive analytics are used by lawyers and litigants to guide filing and settlements decisions (Sadka, Seira, and Woodruff 2018) and are increasingly demanded and adopted by judges. Artificial intelligence in courts promises speedier decisions, considerable cost savings, reduced legal uncertainty, and the elimination of human biases (Ludwig and Mullainathan 2021).

Algorithms can, however, introduce new problems. An extensive literature in computer science and economics discusses concerns that algorithms might reproduce historical patterns of discrimination (Rambachan et al. 2020). This literature has largely ignored how machine learning can impact the broader aims of the judicial system beyond discrimination. Perhaps the chief aim of justice is *deterrence*—preventing violations of the law. While existing machine learning applications in courts center around bail decisions where deterrence is not the main goal (Kleinberg et al. 2018b), artificial intelligence is creeping into other aspects of judicial decision-making where deterrence is central. For instance, Estonia is developing an algorithm to adjudicate small claims disputes, and China

[*] Toulouse School of Economics, Institute for Advanced Study in Toulouse. E-mail: daniel.chen@iast.fr

[†] University of Oslo. E-mail: henrik.sigstad@econ.uio.no

[1] E.g., Rory Cellan-Jones "The robot lawyers are here—and they're winning", BBC, January 26, 2022, https://www.bbc.com/news/technology-41829534.

is experimenting with robot judges to settle e-commerce disputes.[2] Moreover, the use of machine learning by lawyers and litigants to decide settlements of cases could adversely impact deterrence.

In this paper, we use a model to study how machine learning algorithms in courts impact incentives to abide by the law and how they can be designed to improve deterrence. Our baseline model considers the consequences of basing judicial decisions purely on machine predictions. While such a direct usage of machine learning is unrealistic in real settings, this stylized case helps to clarify key mechanisms that generalize to more realistic applications of AI, such as using machine learning to *guide* judicial decisions or settlements. To make our argument transparent, we consider an ideal environment for machine learning. In particular, we assume a machine learning algorithm that perfectly estimates the probability that the defendant is guilty given the observed data. Thus, we essentially assume an algorithm trained on an infinite number of past court cases where guilt was perfectly determined. We also disregard selection issues such as the selection of cases into litigation.

In this ideal setting, we find that basing judicial decisions on machine predictions is optimal if the aim is to avoid judicial errors. In fact, for any trade-off between type I errors (convicting innocents) and type II errors (acquitting guilties), optimal punishment relies on predicted guilt. But this encouraging result does not imply that machine learning is optimal if the aim is to provide ex ante incentives to abide by the law. To study that question, we consider a general deterrence model where agents engage in crime if the gain from crime exceeds the increase in expected punishment. We distinguish between *evidence*—data causally affected by the choice to engage in crime—and *fixed characteristics*—data observed independent of this choice, such as gender or social status. In this model, basing judicial decisions on predicted guilt can severly lower deterrence. For instance, if all agents with a specific fixed characteristic are innocent, the machine learning rule never punishes agents of this type, no matter how strong the evidence is. While this is optimal to reduce judicial errors, it is fatal for incentives: No agent of this type has incentives to stay innocent. Similarly, if all agents with a specific fixed characteristic engage in crime, the machine learning rule always punishes agents of this type, no matter how weak the evidence. Such agents also have no incentives to abstain from crime. Sub-optimal deterrence arises not only in these extreme cases. For instance, when the crime rate within a type is large enough, agents of this type are punished even at evidence more likely to be seen when the agent is innocent than when she engages in crime. In that case, we can improve deterrence *and* reduce punishment costs by lowering the punishment threshold.

Allowing the machine learning punishment rule to be updated with new data as agents respond to the rule only exacerbates these problems. For instance, assume the profit from crime can be deduced from the fixed characteristics. Then all agents of a given fixed characteristics behave in the same way, and the fixed characteristics become perfect predictors of crime. A machine learning punishment rule will always punish criminal types and never punish innocent types, providing nobody with incentives to abide by the law. The optimal response to this punishment rule is for all agents to engage in crime. Thus, in the unique equilibrium, all agents engage in crime and the punishment rule always punishes. There could be other equilibria when the profit from crime is not a direct function of the

[2] Eric Niiler "Can AI Be a Fair Judge in Court? Estonia Thinks So", Wired, January 26, 2022, https://www.wired.com/story/can-ai-be-fair-judge-court-estonia-thinks-so/ and Joshua Park "Your Honor, AI", Harvard International Review, January 26, 2022, https://hir.harvard.edu/your-honor-ai/.

fixed characteristics. These equilibria, however, exhibit a positive crime rate and might be unstable.

Can machine learning models be adapted to improve deterrence? In the ideal scenario, data is labeled as either evidence or fixed characteristics. It is then tempting to predict guilt using only evidence. But it turns out that judicial decisions based on evidence-predicted guilt suffer from many of the same problems as when fixed characteristics are included as predictors. In particular, there are no equilibria with a zero crime rate, and equilibria with low crime rates might be unstable. Instead, we consider a modification of the punishment rule—focusing on the odds of being guilty given the evidence rather than predicted guilt, and normalizing by the population ratio between criminals and innocents. It turns out that this modified punishment rule is equivalent to an optimal non-discriminatory punishment rule. Thus, in our model, a machine learning punishment rule can sustain optimal deterrence when we can distinguish evidence from fixed characteristics.

In reality, however, we are unlikely to obtain such ideal training data. In realistic scenarios, machine learning algorithms are fed large quantities of unprocessed text containing both evidence and descriptive information about the defendant. While a human might do a reasonably good job at categorizing pieces of information as fixed characteristics or evidence, this task is essentially impossible for a machine learning algorithm. Distinguishing fixed characteristics from evidence is a question of causal inference. As piece of information should count as evidence if it would not have been observed had the defendant not committed the crime. Without experimental data or strong assumptions, we can not learn anything about this counterfactual.

Instead of proposing a data-driven way of distinguishing evidence from fixed characteristics, we consider situations where parts of the data are manually labeled as fixed characteristics or evidence. This is a plausible scenario. Human annotators likely have a good intuition about pieces of information that are clearly fixed characteristics, and experienced judges will be able to list clear examples of evidence. We consider three approaches. First, if all *relevant* evidence is labeled, optimal deterrence can be obtained by training the algorithm just on the labeled evidence. We view this situation as unlikely. As an alternative where we lose less data, we consider training the algorithm on all data not labeled as fixed characteristics. This approach is also unlikely to succeed. For instance, if skin color is excluded from the training data, a machine learning algorithm can likely reconstruct this variable from other fixed characteristics, such as neighborhood. Instead, we prefer a third approach—in the spirit of Kleinberg et al. (2018a)—where all data is used for prediction and the punishment rule is "debiased" ex post using the labeled fixed characteristics. This method replicates the optimal non-discriminatory punishment rule if the labeled fixed characteristics sufficiently control for the propensity of a defendant to engage in crime. One advantage of this method is that we can exploit structured data, such as biographical data on the defendant, to debias the algorithm. Machine predictors based on draw text typically rely on vector representations that are difficult to interpret, while identifying fixed characteristics in external structured data is more feasible. None of these solutions, however, are guaranteed to work. We believe that any use of machine predictions to guide legal decisions should be accompanied with interpretable representations of the identified evidence to be manually inspected by human lawyers or judges.

To assess the properties of machine learning decision rules in practice, we collaborate with a Brazilian legal intelligence firm to predict outcomes of labor-related lawsuits. The firm accesses 14 million lawsuits, including all legal briefs filed before the final decision. To obtain fixed characteristics of the firm and the plaintiff, we rely on matched employer-employee data for the entire Brazilian formal sector. We plan to construct machine learning based punishment rules based purely on predicted guilt and as well as our proposed solutions. We rely on the ratio between type I and type II errors to measure statistical discrimination. In particular, machine learning predictors tend to allow more type II errors for firms with a high ex ante likelihood of being guilty than for other firms. Such statistical discrimination is optimal for reducing overall errors but detrimental to deterrence. With this method, we can assess the severity of statistical discrimination under a machine learning punishment rule and the extent to which our proposed solutions lower statistical discrimination. The results of this exercise will be ready for the conference.

Our results build on a literature on the use of probabilistic reasoning in courts. In particular, we are not the first to note that reducing judicial errors might lower deterrence. While early work maintains that accuracy improves deterrence (Kaplow and Shavell 1994), several authors (**Demougin2006Preponderance**; Lando 2002) argue that judges should disregard their prior beliefs about a defendant's guilt. Many legal procedures—especially rules of evidence—restrict the ability of judges to engage in Bayesian inference (Daughety and Reinganum 2000). For instance, evidence proving that a defendant has a high prior likelihood of engaging in crime is generally inadmissible.[3] These rules increase the likelihood of committing errors but can be rationalized if the aim is deterrence (Sanchirico 2001). Mirroring our result that a machine learning decision-rule trained only on evidence can ensure optimal deterrence, **Demougin2006Preponderance** show that Bayesian inference under exclusionary rules of evidence can be optimal. That punishments based on Bayesian beliefs are inefficient is also well known in the literature on moral hazard models (Holmstrom 1979). There, in the optimal contract, the principal needs to withdraw the bonus when the agent produces low output even though, in equilibrium, the principal knows the agent exerted high effort.

In Section 2, we present our model of machine learning punishment. We discuss how to adapt machine learning algorithms to improve deterrence in Section 3. In Section 4, we discuss how our results generalize to settings where machine learning does not directly decide cases but is instead used to *guide* judicial decisions or settlements. We explain our planned empirical application in Section 5. Section 6 concludes.

2 A Model of Machine Learning Punishment

In this section, we propose a simple model of machine learning punishment and show that using predicted guilt to decide court cases is optimal for reducing judicial errors but not for optimizing deterrence. Assume a probability space $(\Omega, \mathcal{F}, \Pr)$ with each $\omega \in \Omega$ being an *agent*. The random variable $A \in \{0, 1\}$ indicates whether the agent is guilty of a crime.[4] Let $Z \in \mathcal{Z}$ be a random vector of *evidence*. We assume that evidence might be *causally* affected by the choice of engaging in crime.

[3] According to Federal Rules of Evidence, Rule 404, "evidence of a person's character or character trait is not admissible to prove that on a particular occasion the person acted in accordance with the character or trait."

[4] By assuming only two possible actions, we abstract away chilling effects—that the prospect of punishment might deter socially desired actions—as considered by **Kaplow2011Optimal**.

In particular, we denote *potential evidence* by Z_1 if A is set to 1 and Z_0 if A is set to 0. We only observe Z_1 if $A = 1$ and Z_0 if $A = 0$, thus $Z = AZ_1 + (1 - A) Z_0$. Denote by $F \in \mathcal{F}$ a random vector of fixed characteristics. Unlike evidence, the fixed characteristics of an agent do not depend on whether A is set to 0 or 1. We will refer to F as the agent's *type*. Let $X = \{F, Z\} \in \mathcal{X}$ be the combined vector of evidence and fixed characteristics. A key challenge will be that an algorithm can not distinguish fixed characteristics from evidence when fed a vector of observational data X. A *punishment rule* is a function $\pi : \mathcal{X} \to [0, r]$ where r is the maximal possible punishment. For a given punishment rule, define the *conviction rule* as the mapping $c(X) = \mathbf{1}[\pi(X) > 0]$. We assume that we observe the joint distribution of (A, X). By assuming this, we abstract away issues of imperfect training data and assume that we can form the best possible prediction of A given X. In other words, we assume

Assumption 1. *We can perfectly estimate* $\mathrm{E}[A \mid X]$ *by machine learning.*

Informally, the implicit assumption is that we have access to a large body of previous cases where guilt was perfectly determined. We also abstract away selection issues, such as the selection of cases into litigation. We seek to relax some of these assumptions later. Under these optimal conditions, it is tempting to use machine learning to determine punishment. In particular, one might consider a conviction rule that punishes if predicted guilt is above a threshold.[5]

Definition 1. π is a *machine learning conviction rule* if

$$\mathrm{E}[A \mid X] > k \Rightarrow c(X) = 1$$

$$\mathrm{E}[A \mid X] < k \Rightarrow c(X) = 0$$

for a constant $k \in [0, 1]$.

In fact, in this ideal setting, such a machine learning conviction rule is optimal if the aim is to reduce judicial errors. To see this, define the *type I error rate*—the probability of falsely punishing an innocent—by $\Pr[c(X) = 1 \mid A = 0]$ and the *type II error rate*—the probability of falsely acquitting a guilty—by $\Pr[c(X) = 0 \mid A = 1]$. We are agnostic about the optimal trade-off between type I and type II errors, but consider conviction rules that reduce both rates as better:

Definition 2. A conviction rule *optimally reduces errors* if there is no rule with lower type I *and* type II error rates.

With this definition, a machine learning conviction rule is optimal. Furthermore, it is the only optimal rule:

Proposition 1. *c optimally reduces errors $\Leftrightarrow$ c is a machine learning conviction rule.*

The threshold k might depend on the preferred trade-off between type I and type II errors. For instance, if convicting one innocent is deemed as bad as letting ten guilty go free, one can set $k = 0.91$. Thus, there is an optimal machine learning conviction rule for any trade-off between type I and type II errors.

[5] When $\mathrm{E}[A \mid X] = k$, we allow for randomization.

2.1 Optimal Deterrence

If the aim of justice is deterrence rather than reducing errors, a machine learning conviction rule might not be optimal. To discuss deterrence, we need assumptions about the agents' incentives. Assume agents differ by an unobserved profit from engaging in crime, defined by the random variable $\Pi > 0$. Let the random variable $C \in \{0,1\}$ indicate whether the agent is an *irrational criminal*—engaging in crime no matter what the punishment is. Assume that for each type $f \in \mathcal{F}$ there is a share $\varepsilon > 0$ of irrational criminals:

Assumption 2. *For all $f \in \mathcal{F}$, $\Pr[C = 1 \mid F = f] = \varepsilon > 0$ with $\varepsilon \to 0$.*

We need some irrational agents to get a positive crime rate under the optimal punishment rule—if all agents have $A = 0$, a machine learning predictor can not be trained. We assume $\varepsilon \to 0$ to simplify our results.[6] The remaining rational agents commit crimes if the profit from crime exceeds the increased expected cost of punishment. In particular, we assume

$$A = \begin{cases} 1 & \text{if } \Pi > \mathrm{E}\left[\pi(F, Z_1) - \pi(F, Z_0) \mid F\right] \text{ or } C = 1 \\ 0 & \text{if } \Pi \leq \mathrm{E}\left[\pi(F, Z_1) - \pi(F, Z_0) \mid F\right] \text{ and } C = 0 \end{cases}$$

We implicitly assume that an agent does not know her realization of potential evidence, Z_0 and Z_1. Instead, she knows the distribution of potential evidence among agents of her type F and maximize expected utility given this belief. To obtain a simple closed form solution to optimal punishment, we maintain the following assumption.

Assumption 3. *Assume $Z_1, Z_0 \perp F$.*

This is a strong assumption—we essentially assume all types produce the same potential evidence. In reality, certain types might be more inclined to produce certain types of evidence, even conditional on guilt. We will later discuss the consequences of relaxing this assumption. Under these assumptions, we seek to find the punishment rule that deters crime among rational agents at minimal social cost.[7]

Definition 3 (Optimal punishment.)**.** A punishment rule π is *optimal* if it deters all rational agents at minimal punishment $\mathrm{E}[\pi(X)]$.

When all rational agents are deterred from crime, minimizing $\mathrm{E}[\pi(X)]$ amounts to minimizing the punishment of innocents. Define the *strength of evidence* $z \in \mathcal{Z}$ by the likelihood ration $s(z) \equiv \frac{\Pr[Z_1 = z]}{\Pr[Z_0 = z]}$, the probability of producing evidence z if the agent is guilty divided by the probability of producing z if the agent is innocent. As in **Becker1968Crime**, the optimal punishment rule is to punish at a high level when the strength of evidence crosses a certain threshold:

[6] When $\varepsilon \to 0$ we do not have to worry about the utility of irrationals when designing the optimal punishment rule.

[7] For simplicity, we assume society wants to deter *all* crime. In reality, it might be optimal to allow some crime in equilibrium, to reduce the social cost of punishment.

Proposition 2. *The* unique *optimal punishment rule can be written as*

$$\pi\left(f,z\right)=\begin{cases}L\left(f\right) & \textit{if } s\left(z\right)\geq B\left(f\right)\\ 0 & \textit{if } s\left(z\right)<B\left(f\right)\end{cases}$$

for functions B and L.

The intuition for this result—similar to Becker's argument—is that reserving punishment for the strongest levels of evidence can reduce wrongful convictions while keeping incentives unchanged. The amount of punishment $L\left(f\right)$ can be lower than the maximal punishment r only when punishment occurs exclusively at the strongest possible evidence ($\arg\max_{z} s\left(z\right)$). The thresholds $B\left(f\right)$ might differ according to types f. If a type f includes agents with large profits from crime Π, punishment must happen at weaker evidence to deter all agents of this type. The optimal punishment rule thus engages in discrimination—some agents are more likely to be punished just because of their type. Such discrimination might not be desirable due to fairness concerns, and we might want to constrain the punishment rule to not take into account fixed characteristics:

Definition 4. A punishment rule is *non-discriminatory* if $\pi\left(f,z\right)=\pi\left(f',z\right)$ for all $f,f'\in\mathcal{F}$ and $z\in\mathcal{Z}$

The optimal non-discriminatory rule has the same structure as the optimal rule, except that the threshold for punishment and the amount of punishment does not depend on f.

Proposition 3. *The* unique *optimal non-discriminatory punishment rule is*

$$\pi\left(f,z\right)=\begin{cases}l & \textit{if } s\left(z\right)\geq k\\ 0 & \textit{if } s\left(z\right)<k\end{cases}$$

for constants $k\in\left[0,1\right]$ and l.

We now consider how a punishment rule based on machine learning compares to this optimal punishment rule.

2.2 Machine Learning and Deterrence

How does machine learning impact incentives to follow the law? To analyze this, we consider the following punishment rule.

Definition 5. Define a *machine learning punishment rule* as

$$\pi\left(X\right)=\begin{cases}l & \text{if } \mathrm{E}\left[A\mid X\right]>k\\ 0 & \text{if } \mathrm{E}\left[A\mid X\right]\leq k\end{cases}$$

for constants $k\in\left[0,1\right]$ and l.

The structure of punishing a constant amount l when predicted guilt crosses a certain threshold is chosen to mimick the optimal punishment rule. It turns out that such a machine learning punishment rule shares several properties with the optimal rule. In particular, an agent is punished by a machine learning punishment rule if the strength of evidence crosses a certain threshold:

Proposition 4. *An agent with $Z = z$ and $F = f$ is punished by a machine learning punishment rule iff*

$$s(z) > \frac{1 - \mathrm{E}[A \mid F = f]}{\mathrm{E}[A \mid F = f]} \frac{k}{1-k}$$

Also, as in the optimal rule, the threshold depends on the type f. Unfortunately, however, the machine learning punishment rule is not optimal. For now, we keep A fixed to look at the "short term" effects of a machine learning punishment rule on incentives. In Section 2.3, we consider the "long term" effects—what happens when A is allowed to endogenously respond to the punishment rule. To see why the machine learning punishment rule creates suboptimal incentives, it is useful to look at some examples. First, consider an "innocent type" $f \in \mathcal{F}$ consisting only of innocent agents and irrational criminals: $\mathrm{E}[A \mid F = f] = \varepsilon$. For this type, the machine learning rule punishes if $s(z) > \lim_{\varepsilon \to 0} \frac{1-\varepsilon}{\varepsilon} \frac{k}{1-k} = \infty$. Thus such innocent types are never punished, no matter how strong the evidence is. This is optimal if the aim is to reduce errors, but has fatal consequences for incentives: No agent of this type has incentives to stay innocent. Conversely, members of "criminal types" consisting only of agents engaging in crime ($\mathrm{E}[A \mid F = f] = 1$) are always punished, no matter how weak the evidence is. These agents also have no incentives to abstain from crime. More generally, if $\mathrm{E}[A \mid F = f] > k$, agents of type f are punished even when the evidence favors innocence ($s(z) < 1$). This can never be optimal: By increasing the threshold for punishing above 1 one can both improve incentives *and* reduce punishment. The reason that the machine learning punishment rules fails to be optimal is that it engages in statistical discrimination against "criminal types" and in favor of "innocent types". Note that machine learning engages in discrimination in this model even though there is no discrimination in the training data, contrary to the common wisdom that algorithmic bias is caused by biased data.[8]

2.3 Endogenous Data

What happens if agents optimally respond to the machine learning punishment rule and the training data is updated in real time? To study this, fix a type f. Denote the machine learning punishment rule for type f when $\mathrm{E}[A \mid F = f] = \alpha$ by:

$$\pi_\alpha(z) \equiv \begin{cases} l & s(z) > \frac{1-\alpha}{\alpha} \frac{k}{1-k} \\ 0 & s(z) \leq \frac{1-\alpha}{\alpha} \frac{k}{1-k} \end{cases}$$

For any given punishment rule π_α, denote by $g(\alpha)$ the share not deterred among type f:

$$g(\alpha) = \Pr[\Pi > \mathrm{E}[\pi_\alpha(Z_1)] - \mathrm{E}[\pi_\alpha(Z_0)] \mid F = f]$$

[8] See **Rambachan2019Bias** for a related point.

In equilibrium, we need the share not deterred by π_α to be exactly α. Thus

Definition 6. An *equilibrium* is a fixed point: $g(\alpha) = \alpha$

We are unable to solve analytically for this equilibrium. Instead, we will discuss certain pathological features of equilibrium. The most striking example of how machine learning punishment rule can go wrong is when all agents of the same type have the same profit from crime. In that case, all agents of a given type behave in the same way. Thus, the machine learning algorithm can use the agent's type as a perfect predictor of crime. Agents of "criminal types" are always punished and agents of "innocent types" are never punished. Nobody has incentives to abstain from crime. We thus get

Proposition 5 (No deterrence when Π is observable.)**.** *Assume $\Pi = h(F)$ and a machine learning punishment rule. Then all agents engage in crime in equilibrium.*

We now consider what happens when the profit from crime can not be perfectly proxied by observable fixed characteristics. To analyze the properties of equilibrium, it is useful to note the following:

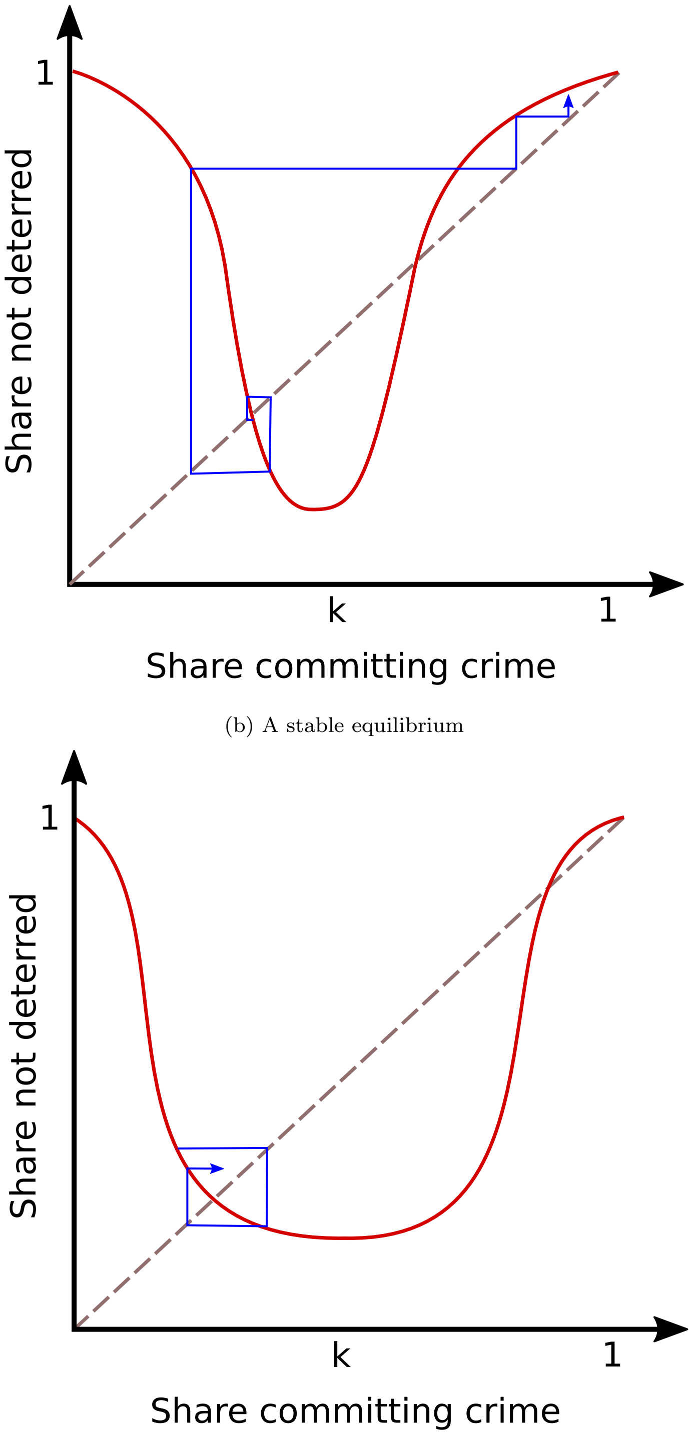
Proposition 6. *g is u-shaped with lowest point at $\alpha = k$ and $g(0) = g(1) = 1$.*

Thus incentives to abstain from crime is zero when all agents are innocent, increasing in the share of criminals until $\mathrm{E}[A \mid F = f] = k$, and then decreasing again. In Figure 1, we show two examples of g functions compatible with Proposition 6. Proposition 6 implies that there are most three equilibria. All agents committing crime is always an equilibrium. In addition, there might be two other equilibria with positive crime rates. There is never an equilibrium with zero crime. Furthermore, the equilibria with the lower crime rates might be unstable. For instance, assume g is as in Figure 1 a) and assume we start in the lowest crime equilibrium. If the crime rate increases slightly, the machine learning rule punishment punishes at weaker evidence leading to a lower crime rate than in equilibrium. This lower crime rate leads to a more lenient punishment rule which causes a crime rate that is even larger than the initial deviation from equilibrium. The dynamic responses lead towards the equilibrium where all agents commit crimes, the only stable equilibrium. A low crime equilibrium can be stable, however, as shown in Figure 1 b). Stable equilibria can occur if the density of indifferent agents in equilibrium is low. In other words, when agents close to indifference differ sufficiently in their profit from crime.

3 Adapting Machine Learning Punishment

Can the machine learning punishment rule be improved? In this section, we first show that a machine learning punishment rule trained only on evidence can mimick the optimal punishment rule. Then, we discuss three different ways to approximate the optimal rule when evidence and fixed characteristics is not perfectly distinguished in the data.

Figure 1: Example equilibria


(a) An unstable equilibrium

(b) A stable equilibrium

Note: Example of possible g functions. There are three equilibria characterized by $g(\alpha) = \alpha$. The arrows indicate the dynamics after the crime rate deviates slightly from the low crime equilibrium value.

3.1 Evidence-based Machine Learning Punishment

Since statistical discrimination against certain types lead to , it is tempting to use a machine learning punishment rule that is based only on evidence. Setting issues of distinguishing evidence from fixed characteristics aside, would such a rule be optimal? In particular, we might want to use a punishment rule of the form

$$\pi\left(f,z\right)=\begin{cases} l & \text{if } \mathrm{E}\left[A\mid Z=z\right]>k\\ 0 & \text{if } \mathrm{E}\left[A\mid Z=z\right]\leq k\end{cases}$$

for constants k and l. This rule is, unfortunately, also not optimal and also suffer from the issues discussed in Sections 2.2 and 2.3. In particular, there are up to three equilibria—all with a positive crime rate—and the equilibria with low crime rates might be unstable. While the rule does not discriminate based on f, it statistically discriminates based on the overall level of crime in society, $\mathrm{E}\left[A\right]$. It is impossible to deter all rational agents from crime since when $\mathrm{E}\left[A\right]=\varepsilon$ and $\varepsilon\rightarrow0$ no agents are ever punished. It turns out, however, that a modification of this rule is able to mimick the optimal non-discriminatory punishment rule.

Definition 7. Define an *evidence-based machine learning punishment rule by*

$$\pi\left(f,z\right)=\begin{cases} l & \text{if } \frac{\Pr[A=1|Z=z]}{\Pr[A=0|Z=z]}\Big/\frac{\Pr[A=1]}{\Pr[A=0]}>k\\ 0 & \text{if } \frac{\Pr[A=1|Z=z]}{\Pr[A=0|Z=z]}\Big/\frac{\Pr[A=1]}{\Pr[A=0]}\leq k\end{cases}$$

for constants k and l.

Instead of focusing on predicted guilt, this formula depends on the *odds* of being guilty given the evidence, $\Pr\left[A=1\mid Z=z\right]/\Pr\left[A=0\mid Z=z\right]$. We normalize the odds by the ratio of criminals to innocents in the whole population to avoid a punishment rule that depends on the population crime rate. This rule punishes with a positive probability even when all rational agents are innocent. In fact, one can show that $\frac{\Pr[A=1|Z=z]}{\Pr[A=0|Z=z]}\Big/\frac{\Pr[A=1]}{\Pr[A=0]}=s\left(z\right)$. Thus

Proposition 7. *The optimal non-discriminatory punishment rule can be implemented by an evidence-based machine learning punishment rule.*

All parts of the evidence-based machine learning punishment rule can be estimated with machine learning. Thus, under the maintained assumptions, if we know which pieces of information should count as evidence, machine learning can be used to sustain optimal punishment. In reality, however, it might be difficult to distinguish evidence from fixed characteristics.

3.2 Imperfectly Labeled Data

In reality, fixed characteristics and evidence can not easily be distinguished. While a trained human judge might be able to disentangle evidence from fixed characteristics, a machine learning algorithm can not. Knowing which pieces of information should count as evidence is essentially a causal question. Information that is causally affected by an agent's choice of engaging in crime should be labeled evidence and information that is not affected should be labeled fixed characteristics. Without experimental data, an algorithm—let alone a predictive algorithm—can not succeed in this task. Instead, in this section, we consider the use of machine learning algorithms to determine punishment when parts of the data are labeled as fixed characteristics or evidence. Human annotators can easily label some pieces of information as fixed characteristics—e.g., race, neighborhood of residence, and education level of the defendant. Similarly, some pieces information might be manually labeled as evidence by experienced judges. Thus, assume the data is given by the vector

$$X = \{X_1, \ldots, X_n\}$$

and that each entry in X is either labeled as evidence, labeled as a fixed characteristic, or unlabeled. With this data, we consider three approaches. First, we consider using only data known to be evidence when training the machine learning model. Second, we consider training the algorithm on all data except what is known to be fixed characteristics. Finally, we consider using all data for prediction and debias ex post using the known fixed characteristics.

Training algorithm on data known to be evidence

Let W be the vector of data known to be evidence. Since the entries in W is a subset of all pieces of evidence in Z, there is a function h such that $W = h(Z)$. Using W instead of Z leads to optimal punishment if W contains all the relevant pieces of evidence:

Proposition 8. *The optimal punishment rule with no discrimination can be written as*

$$\pi(w) = \begin{cases} l & \textit{if } \frac{\Pr[A=1|W=w]}{\Pr[A=0|W=w]} \Big/ \frac{\Pr[A=1]}{\Pr[A=0]} > k \\ 0 & \textit{if } \frac{\Pr[A=1|W=w]}{\Pr[A=0|W=w]} \Big/ \frac{\Pr[A=1]}{\Pr[A=0]} \leq k \end{cases}$$

if $h(z) = h(z') \Rightarrow s(z) = s(z')$.

In reality, however, it is highly unlikely that all the relevant evidence is labeled as evidence in the training data. In this case, deterrence may be suboptimal. An alternative route that does not throw out as much data is to just exclude data known to be fixed characteristics.

Training algorithm on all data except known fixed characteristics

Instead of training a machine learning algorithm only on data known to be evidence one can train it on all data known to be not fixed characteristics. For instance, one might leave out obvious fixed characteristics such as skin color from the training data. This is unlikely to work well, however, since machine learning algorithms are generally able to reconstruct such excluded characteristics. For instance, neighborhood might be a good proxy for skin color. A better approach—similar to Kleinberg et al. (2018a)—is to use the fixed characteristics to *debias* the machine learning predictions ex post.

Training algorithm on all data and debias ex post

Let G be the vector of data known to be fixed characteristics. If, after controlling for G, other fixed characteristics are unrelated to the profit of crime, optimal non-discriminatory punishment can be acheived by the following rule:

Proposition 9. *If $\Pi \perp F \mid G$, the optimal punishment rule with no discrimination can be written as*

$$\pi(x,g)=\begin{cases} l & \text{if } \frac{\Pr[A=1|X=x]}{\Pr[A=0|X=x]} \Big/ \frac{\Pr[A=1|G=g]}{\Pr[A=0|G=g]} > k \\ 0 & \text{if } \frac{\Pr[A=1|X=x]}{\Pr[A=0|X=x]} \Big/ \frac{\Pr[A=1|G=g]}{\Pr[A=0|G=g]} \leq k \end{cases}$$

for constants $0<k<1$ and l.

Here, we first calculate the odds of being guilty given all the data, including fixed characteristics. Then, we debias by dividing by the share of criminals to innocents among individuals of the same fixed characteristics g. With this approach, it is not necessary to control for all fixed characteristics. We do not even have to control for all predictors of Π—once you control for some predictors you implicitly control for all other perfectly correlated predictors. Thus, we do not have to worry about other fixed characteristics being able to reconstruct skin color once we have controlled for skin color. Since Π is unobserved, however, the assumption $\Pi \perp F \mid G$ can not be directly assessed. This method must thus be applied with caution.

4 Machine learning guiding judicial decisions and settlements

So far, we have considered deciding court cases based purely on machine learning. Such a direct usage of machine learning is unrealistic. This section considers how our conclusions generalize to more realistic use cases. In particular, we discuss judges using machine learning algorithms to *guide* decisions and lawyers using machine predictions to inform settlement decisions.

4.1 Machine learning guiding judicial decisions

How would our conclusions be modified if machine learning predictors only *guide* judicial decisions? In some court cases, the judge might costlessly determine the correct outcome without the assistance of an algorithm. In such cases, the presence of a machine prediction will not influence judicial decisions. However, it is likely that in many cases, the judge will at least partly rely on machine predictions to save time and effort and perhaps reduce errors. The issues discussed in Section 2 likely arise if machine predictions influence judicial decisions. Exactly how these issues manifest themselves depends on how the judge combines the machine predictions with external information. For instance, if the judge knows the crime rate among similar defendants, she might informally engage in the debiasing exercise proposed in Section 3.2. The effect of machine predictions on deterrence depends on the judge's objective function. If the judge aims to provide incentives to abide by the law, she might decide to let machine predictions influence her only when she believes it improves deterrence.[9] If, instead, the judge aims to avoid reversals of her decisions, the presence of a machine learning algorithm can do more harm. In ongoing work, we formalize how judges might rationally incorporate machine predictions under various assumptions and consider its consequences for incentives to abide by the law.

4.2 Machine learning guiding settlement decisions

Here, we consider a civil case with a plaintiff (he) and a defendant (she). Assume a judge perfectly determines guilt if the case goes to trial.[10] Moreover, assume the parties have access to a machine learning algorithm that predicts the judge's decision perfectly given the data, X.[11] In other words, they can learn $\mathrm{E}\left[A \mid X\right]$, the predicted probability that the judge decides in favor of the plaintiff given X. Assume the parties base their settlements decision on this prediction only.[12] Denote the fixed cost of going to trial by $c > 0$, payed by the loser of the trial. We assume the defendant pays the plaintiff a fine k if found guilty. The parties can decide to settle instead of going to trial. Denote the settlement payment by the random variable V. The expected value of going to trial is $-\left(c+k\right)\mathrm{E}\left[A \mid X\right]$ for the defendant and $k\,\mathrm{E}\left[A \mid X\right] - c\left(1 - \mathrm{E}\left[A \mid X\right]\right)$ for the plaintiff. Since we have assumed symmetric information, the parties can rationally agree on a range of settlements values. Assume the parties engage in Nash bargaining. In particular, the surplus from settling the case, c, is split evenly. The settlement value is then

$$V = \left(c+k\right)\mathrm{E}\left[A \mid X\right] - \frac{c}{2}$$

Now, consider the ex ante incentives for the defendant to abide by the law. Assume the plaintiff can credibly threaten to go to court no matter the defendant's guilt. Let V_1 be the settlement value if A is set to 1 and V_0 the settlement value if A is set to 0. Then, the defendant will abide by the law if and only if

$$\Pi < \mathrm{E}\left[V_1 - V_0\right] = \left(c+k\right)\left(\mathrm{E}\left[A \mid F, Z_1\right] - \mathrm{E}\left[A \mid F, Z_0\right]\right)$$

Similar to when predicted guilt is used to decide court cases, defendants of types $f \in \mathcal{F}$ where $\mathrm{E}\left[A \mid F = f\right]$ is either large or small have reduced incentives to abide by the law. For instance, if $\mathrm{E}\left[A \mid F = f\right] = 1$ defendants of type f settle cases at the same value independent of guilt and thus have no incentives to abide by the law. Also, as in Section 2.3, once we take into account that the machine learning algorithm can be updated with new data, there are no equilibria with full compliance with the law and the equilibria where some abide by the law might be unstable. Thus, basing settlements decisions on machine learning predictions can lead to the same problems as when machine learning is used to guide judicial decisions.

[9] However, if basing decisions on machine predictions saves effort, she might still accept lower deterrence.

[10] In other words, we assume that the judge will have access to additional evidence upon trial that allows perfect determination of guilt.

[11] For simplicity, we assume both parties have access to the same data.

[12] This strategy is not necessarily optimal if the parties have access to additional information. For instance, the defendant can exploit that she knows whether she is guilty. In ongoing work, we consider this case.

5 Planned Empirical Application

We collaborate with a Brazilian legal intelligence firm to predict outcomes of labor-related lawsuits to assess the properties of machine-learning-based decision rules. The firm accesses 14 million lawsuits, including all legal briefs filed before the final decision. This setting is ideal for AI-guided legal decisions. In particular, a large data set with relatively similar cases allows for precise predictions. Also, anecdotally, Brazilian firms demand AI-powered predictions to guide settlement decisions. We will merge the legal data with rich data on fixed characteristics of the firm and the plaintiff using matched employer-employee data for the entire Brazilian formal sector (Relação Anual de Informações Sociais). Important fixed characteristics include race, gender, the plaintiff's employment history, and the firm's size and sector. We plan to train machine learning predictors of case outcomes using (i) all data, (ii) only data identified by lawyers as relevant evidence, (iii) all data excluding known fixed characteristics, and (iv) only fixed characteristics. The latter will be used to debias predictions based on all data. To assess decision rules, we rely on the ratio between type I and type II errors, proxied by prediction errors in a hold-out sample. This ratio should not depend on fixed characteristics under the optimal non-discriminatory punishment rule. A decision rule based on predicted guilt, however, allows more type II errors for firms with a high ex ante likelihood of being guilty than for other firms. Such statistical discrimination is optimal for reducing the overall error rate but is detrimental to deterrence. The relationship between the type I to type II error ratio and guilt predicted by fixed characteristics is thus informative about deterrence. Using this approach, we can assess the severity of statistical discrimination under a punishment rule based on predicted guilt and how our proposed solutions lower such discrimination.

Proposition 10. *Under a machine learning punishment rule, the type I (type II) error rate among agents of type f is increasing (decreasing) in* $\mathrm{E}[A \mid F = f]$.

5.1 Application

We selected the following sample:

- A random sample of 44,000 cases + drop firms with less than 10 cases in total
- We keep only cases that are decided by the judge (removing settled cases, pending cases, and cases where the plaintiff abandons the case)
- We use as fixed characteristics the sector of the firm, the number of past cases involving the firm (including pending and settled cases), the share of past cases that are settled/pending or decided, the share of decided cases where the judge has decided with the plaintiff,
- We consider the last three years and the last year
  - For firms that are censored, last three years is all the way back

6 Conclusions

Basing legal decisions on predicted guilt can be optimal for reducing errors but can lower incentives to abide by the law. We have disccussed several ways to modify machine learning models to improve deterrence. None of these solutions are perfect, however, and whether they work depends on strong untestable assumptions. We thus believe that any usage of machine learning to guide legal decisions should be accompanied by interpretable representations of the evidence identified by the system. These representations might be used to draw the attention of the lawyer or judge to potentially important aspects of the case and to manually verify whether the system indeed identifies relevant pieces of evidence. Methods for interpretable machine learning (Molnar 2020) can be used to achieve this goal. In this paper, we have considered an ideal setting for machine learning: an infinite training sample with no selection issues and perfectly observed guilt. In future work, it will be important to assess additional issues that might arise from relaxing these conditions. Also, we have assumed that data can be perfectly classified as either evidence or fixed characteristics. In reality, certain data might have both properties—i.e., some individuals might have a higher propensity to produce a specific piece of evidence even conditional on guilt. How to deal with this case is an important question for future work. X

References

Daughety, Andrew F and Jennifer F Reinganum (2000). "Appealing judgments". In: *The Rand Journal of Economics*, pp. 502–525.

Holmstrom, Bengt (1979). "Moral Hazard and Observability". In: *The Bell Journal of Economics* 10.1, pp. 74–91. ISSN: 0361915X. DOI: 10.2307/3003320.

Kaplow, Louis and Steven Shavell (1994). "Accuracy in the Determination of Liability". In: *The Journal of Law and Economics* 37.1, pp. 1–15.

Kleinberg, Jon et al. (2018a). "Algorithmic Fairness". In: *Aea papers and proceedings*. Vol. 108, pp. 22–27.

Kleinberg, Jon et al. (2018b). "Human Decisions and Machine Predictions". In: *The Quarterly Journal of Economics* 133.1, pp. 237–293.

Lando, Henrik (2002). "When is the preponderance of the evidence standard optimal?" In: *The Geneva Papers on Risk and Insurance-Issues and Practice* 27.4, pp. 602–608.

Ludwig, Jens and Sendhil Mullainathan (Sept. 2021). *Fragile Algorithms and Fallible Decision-Makers: Lessons from the Justice System*. Working Paper 29267. National Bureau of Economic Research. DOI: 10.3386/w29267. URL: http://www.nber.org/papers/w29267.

Molnar, Christoph (2020). *Interpretable machine learning*. Lulu. com.

Rambachan, Ashesh et al. (2020). "An economic perspective on algorithmic fairness". In: *AEA Papers and Proceedings*. Vol. 110, pp. 91–95.

Sadka, Joyce, Enrique Seira, and Christopher Woodruff (2018). *Information and Bargaining through Agents: Experimental Evidence from Mexico's Labor Courts*. Tech. rep. National Bureau of Economic Research.

Sanchirico, Chris William (2001). “Character evidence and the object of trial”. In: *Colum. L. Rev.* 101, p. 1227.

A Proofs

Proof. (Proposition 1). Define $c(X) \equiv \mathbf{1}[\pi(X) > 0]$. Denote type I and type II errors of punishment rule π by

$$e_1(\pi) = \Pr[\pi(X) > 0 \mid A = 0]$$

$$e_2(\pi) = \Pr[\pi(X) = 0 \mid A = 1]$$

Assume π satisfies

$$\pi(X) > 0 \Leftrightarrow \mathrm{E}(A \mid X) > k$$

and π' is a punishment rule that has lower type I and type II error. Denote by B the event that π incarcerates but not π'.

$$B \equiv (\pi'(X) = 0, \pi(X) > 0)$$

Similarly, define

$$C = (\pi(X) = 0, \pi'(X) > 0)$$

Then must have

$$e_1(\pi') < e_1(\pi) \Leftrightarrow \Pr[A = 0, C] < \Pr[A = 0, B]$$

$$e_2(\pi') < e_2(\pi) \Leftrightarrow \Pr[A = 1, B] < \Pr[A = 1, C]$$

Since $\pi(X) > 0 \Leftrightarrow \mathrm{E}(A \mid X) > k$, we have

$$\Pr[A = 1 \mid B] > \Pr[A = 1 \mid C]$$

$$\Leftrightarrow \frac{\Pr[A = 1, B]}{\Pr[B]} > \frac{\Pr[A = 1, C]}{\Pr[C]}$$

Together with $\Pr[A = 1, B] < \Pr[A = 1, C]$, this implies $\Pr[C] > \Pr[B]$. But $\Pr[A = 0 \mid B] < \Pr[A = 0 \mid C]$ and $\Pr[A = 0, C] < \Pr[A = 0, B]$ imply $\Pr[C] < \Pr[B]$, a contradiction. To show that no other punishment rule optimally reduces errors, assume a punishment rule π does not satisfy $\pi(X) > 0 \Leftrightarrow \mathrm{E}(A \mid X) > k$. Then there exist values $x_1 \in \mathcal{X}$ and $x_2 \in \mathcal{X}$ such that $\mathrm{E}[A \mid X = x_1] > \mathrm{E}[A \mid X = x_2]$, $\pi(x_1) = 0$, and $\pi(x_2) > 0$. Assume $\Pr[X = x_2] < \Pr[X = x_1]$.[13] Consider a modified punishment rule

$$\tilde{\pi}(x) = \begin{cases} \pi(x) & x \notin \{x_1, x_2\} \\ \pi(x_1) & x = x_2 R + x_1(1 - R) \\ \pi(x_2) & x = x_1 \end{cases}$$

[13] The case $\Pr[X = x_1] < \Pr[X = x_2]$ is analog.

where R is a random variable with $\Pr[R=1] = \frac{\Pr[X=x_2]}{\Pr[X=x_1]}$. We have

$$e_1(\pi) - e_1(\tilde{\pi}) = \Pr[X = x_2 \mid A = 0] - \Pr[X = x_1 \mid A = 0]$$

$$= \frac{\Pr[A=0 \mid X = x_2]}{\Pr[A=0]} \Pr[X = x_2] \Pr[R=1] - \frac{\Pr[A=0 \mid X=x_1]}{\Pr[A=0]} \Pr[X = x_1]$$

$$= (\Pr[A=0 \mid X=x_2] - \Pr[A=0 \mid X=x_1]) \frac{\Pr[X=x_1]}{\Pr[A=0]}$$

$$= (\mathrm{E}[A \mid X=x_1] - \mathrm{E}[A \mid X=x_2]) \frac{\Pr[X=x_1]}{\Pr[A=0]} > 0$$

Similarly

$$e_2(\pi) - e_2(\tilde{\pi}) > 0$$

□

Proof. (Proposition 2). Fix an $f \in \mathcal{F}$ and consider the optimal punishment rule for this type. We need to show that the optimal punishment rule is of the form

$$\pi^*(z) = \begin{cases} l & \text{if } s(z) \geq b \\ 0 & \text{if } s(z) < b \end{cases}$$

for constants b and l. Assume π is an optimal punishment rule and is not of the above form. Then there exists $z_1 \in \mathcal{Z}$ and $z_2 \in \mathcal{Z}$ such that $\pi(z_1) < \bar{K}$, $\pi(z_2) > 0$ and $s(z_1) > s(z_2)$. Consider the punishment rule

$$\tilde{\pi}(z) = \begin{cases} \pi(z) & z \notin \{z_1, z_2\} \\ \pi(z) + \delta & z = z_1 \\ \pi(z) - \delta \frac{\Pr[Z_1=z_1] - \Pr[Z_0=z_1]}{\Pr[Z_1=z_2] - \Pr[Z_0=z_2]} & z = z_2 \end{cases}$$

where $0 < \delta < \bar{K} - \pi(x)$. It is straight forward to verify that

$$\mathrm{E}[\tilde{\pi}(Z_1) - \tilde{\pi}(Z_0)] = \mathrm{E}[\pi(Z_1) - \pi(Z_0)]$$

Thus $\tilde{\pi}$ also deters all rational agents. Under Assumption 2, the cost of punishment equals

$$\lim_{\varepsilon \to 0} \mathrm{E}[\pi(Z)] = \mathrm{E}[\pi(Z_0)]$$

$$\lim_{\varepsilon \to 0} \mathrm{E}[\tilde{\pi}(Z)] = \mathrm{E}[\tilde{\pi}(Z_0)]$$

Since $\mathrm{E}[\tilde{\pi}(Z_0)] < \mathrm{E}[\pi(Z_0)]$, $\tilde{\pi}$ deters at a lower cost. □

Proof. (Proposition 3). Under a non-discriminatory punishment rule, we must ignore F. Thus the optimal non-discriminatory punishment rule is the optimal punishment rule when $\mathcal{F} = \{1\}$—when all agents are of the same type. The result then follows from Proposition 2. □

Proof. (Proposition 4). Consider an agent with $Z = z$ and $F = f$. The agent is punished if and only if

$$\mathrm{E}\left[A \mid F = f, Z = z\right] > k$$

We have

$$\Pr\left[A = 1, F = f, Z = z\right]$$

$$= \Pr\left[Z = z \mid A = 1, F = f\right] \Pr\left[A = 1, F = f\right]$$

$$= \Pr\left[Z_1 = z\right] \mathrm{E}\left[A \mid F = f\right] \Pr\left[F = f\right]$$

Similarly

$$\Pr\left[A = 0, F = f, Z = z\right] = \Pr\left[Z_0 = z\right] \left(1 - \mathrm{E}\left[A \mid F = f\right]\right) \Pr\left[F = f\right]$$

This gives

$$\mathrm{E}\left[A \mid F = f, Z = z\right] = \frac{\Pr\left[A = 1, F = f, Z = z\right]}{\Pr\left[F = f, Z = z\right]}$$

$$= \frac{\Pr\left[Z_1 = z\right] \mathrm{E}\left[A \mid F = f\right]}{\Pr\left[Z_1 = z\right] \mathrm{E}\left[A \mid F = f\right] + \Pr\left[Z_0 = z\right] \left(1 - \mathrm{E}\left[A \mid F = f\right]\right)}$$

$$= \frac{s\left(z\right) \mathrm{E}\left[A \mid F = f\right]}{s\left(z\right) \mathrm{E}\left[A \mid F = f\right] + \left(1 - \mathrm{E}\left[A \mid F = f\right]\right)}$$

Thus, this agent is punished if and only if

$$\frac{s\left(z\right) \mathrm{E}\left[A \mid F = f\right]}{s\left(z\right) \mathrm{E}\left[A \mid F = f\right] + \left(1 - \mathrm{E}\left[A \mid F = f\right]\right)} > k$$

$$\Leftrightarrow s\left(z\right) > \frac{1 - \mathrm{E}\left(A \mid F = f\right)}{\mathrm{E}\left(A \mid F = f\right)} \frac{k}{1 - k}$$

□

Proof. (Proposition 5). Fix a type $f \in \mathcal{F}$. When $\Pi = h\left(F\right)$, all agents of type f face the same incentives. Thus either all choose $A = 1$ or all choose $A = 0$. If all choose $A = 1$ then, by Proposition 4, a machine learning punishment rule punishes for all $z \in \mathcal{Z}$. If the punishment rule punishes for all $z \in \mathcal{Z}$, it is optimal for agents to choose $A = 1$. Thus, this is a possible equilibrium. If all agents of type f choose $A = 0$, the machine learning punishment rule does not punish for any $z \in \mathcal{Z}$. Then it is optimal for all agents to choose $A = 1$. Thus, the only possible equilibrium is for all agents to engage in crime (choose $A = 1$). □

Proof. (Proposition 6). The derivative of g has the opposite sign of the derivative of

$$h\left(\alpha\right) \equiv \mathrm{E}\left[\pi_\alpha\left(Z_1\right)\right] - \mathrm{E}\left[\pi_\alpha\left(Z_0\right)\right] = l\left(\Pr\left[s\left(Z_1\right) > \frac{1-\alpha}{\alpha}\frac{k}{1-k}\right] - \Pr\left[s\left(Z_0\right) > \frac{1-\alpha}{\alpha}\frac{k}{1-k}\right]\right)$$

with respect to α. Since $\frac{1-\alpha}{\alpha}\frac{k}{1-k}$ is decreasing in α and

$$a > 1 \Rightarrow \Pr\left[s\left(Z_1\right) = a\right] \geq \Pr\left[s\left(Z_0\right) = a\right]$$

$$a < 1 \Rightarrow \Pr\left[s\left(Z_1\right) = a\right] \leq \Pr\left[s\left(Z_0\right) = a\right]$$

we get (using $\frac{1-\alpha}{\alpha}\frac{k}{1-k} > 1 \Leftrightarrow \alpha < k$)

$$\alpha < k \Rightarrow h'\left(\alpha\right) > 0$$

$$\alpha > k \Rightarrow h'\left(\alpha\right) < 0$$

Thus g is u-shaped with lowest point at $\alpha = k$. Finally, since $h\left(0\right) = h\left(1\right) = 0$ and $\Pi > 0$, we get $g\left(0\right) = g\left(1\right) = 1$. □

Proof. (Proposition 7). Using Bayes' rule

$$\frac{\Pr\left(A = 1 \mid Z = z\right)}{\Pr\left(A = 0 \mid Z = z\right)} \frac{\Pr\left(A = 0\right)}{\Pr\left(A = 1\right)}$$

$$= \frac{\Pr\left(Z = z \mid A = 1\right)}{\Pr\left(Z = z \mid A = 0\right)} = \frac{\Pr\left(Z_1 = z\right)}{\Pr\left(Z_0 = z\right)} = s\left(z\right)$$

□

Proof. (Proposition 8). We have that

$$\frac{\Pr\left(A = 1 \mid W = w\right)}{\Pr\left(A = 0 \mid W = w\right)} \frac{\Pr\left(A = 0\right)}{\Pr\left(A = 1\right)}$$

$$= \frac{\Pr\left(W = w \mid A = 1\right)}{\Pr\left(W = w \mid A = 0\right)} = \frac{\Pr\left[h\left(Z_1\right) = w\right]}{\Pr\left[h\left(Z_0\right) = w\right]}$$

$$= \frac{\sum_{h(z)=w} \Pr\left[Z_1 = z\right]}{\sum_{h(z)=w} \Pr\left[Z_0 = z\right]}$$

$$= \frac{\sum_{h(z)=w} s\left(z\right) \Pr\left[Z_0 = z\right]}{\sum_{h(z)=w} \Pr\left[Z_0 = z\right]}$$

Assume $h\left(z\right) = h\left(z'\right) \Rightarrow s\left(z\right) = s\left(z'\right)$ and define s_w as the common strength of evidence $s\left(z\right)$ for all z with $h\left(z\right) = w$. Then

$$\frac{\Pr\left(A = 1 \mid W = w\right)}{\Pr\left(A = 0 \mid W = w\right)} \frac{\Pr\left(A = 0\right)}{\Pr\left(A = 1\right)} = s_w$$

Thus, the proposed punishment rule can be written as

$$\pi\left(w\right) = \begin{cases} l & \text{if } s_w \geq k \\ 0 & \text{if } s_w < k \end{cases}$$

This is clearly equivalent to the optimal non-discriminatory punishment rule. □

Proof. (Proposition 9). The optimal non-discriminatory punishment rule can be written as

$$\pi(f,z)=\begin{cases} l & \text{if } s(z)\geq k \\ 0 & \text{if } s(z)<k \end{cases}$$

Assume $\Pi \perp F \mid G$. For $x=(f,z)$ we then have

$$s(z)=\frac{\Pr(A=1\mid X=x)}{\Pr(A=0\mid X=x)}\frac{\Pr(A=0\mid F=f)}{\Pr(A=1\mid F=f)}$$

□

Proof. (Proposition 10). Under a machine learning punishment rule the agent is punished iff

$$s(z)>\frac{1-\mathrm{E}[A\mid F=f]}{\mathrm{E}[A\mid F=f]}\frac{k}{1-k}\equiv\alpha_f$$

The type I error rate thus equals $\Pr[s(Z)>\alpha_f\mid A=0]=\Pr[s(Z_0)>\alpha_f]$. The result follows since α_f is decreasing in $\mathrm{E}[A\mid F=f]$. Similarly, the type II error rate $\Pr[s(Z)\leq\alpha_f\mid A=1]=\Pr[s(Z_1)\leq\alpha_f]$ is decreasing in $\mathrm{E}[A\mid F=f]$. □

A.1 Endogenous Filing of Cases

In our main analysis, we have implictly assumed that the algorithm has access to evidence and whether the agent was guilty for a large number of randomly drawn agents (Assumption 1). In real applications of ML in courts, algorithms can be trained only on cases that are brought to court, a very selected sample of all potential "cases" (**Priest1984Selection**). In this section, we discuss how our conclusions might change once we take into account that not all potential cases are brought to trial. How our conclusions change depends on how we model the selection of cases into litigation.

In the most general case, let D indicate whether a court case is filed against the agent and in place of Assumption 1, assume

Assumption 4. *We can perfectly estimate* $\mathrm{E}[A\mid X,D=1]$ *by machine learning.*

Consider the following modified machine learning conviction rule

Definition 8. π is a *selected sample machine learning conviction rule* if

$$\mathrm{E}[A\mid X,D=1]>k\Rightarrow c(X)=1$$

$$\mathrm{E}[A\mid X,D=1]<k\Rightarrow c(X)=0$$

for a constant $k\in[0,1]$.

We still have that a machine learning conviction rule is optimal to reduce errors (REDEFINE ERRORS ONLY WITHIN SAMPLE):

Proposition 11. *c optimally reduces errors $\Leftrightarrow$ c is a selected sample machine learning conviction rule.*

For deterrence, assume

$$A = \begin{cases} 1 & \text{if } \Pi > \mathrm{E}\left[D\left[\pi\left(F, Z_1\right) - \pi\left(F, Z_0\right)\right] \mid F\right] \text{ or } C = 1 \\ 0 & \text{if } \Pi \leq \mathrm{E}\left[D\left[\pi\left(F, Z_1\right) - \pi\left(F, Z_0\right)\right] \mid F\right] \text{ and } C = 0 \end{cases}$$

Definition 9 (Optimal punishment.)**.** A punishment rule π is *optimal* if it deters all rational agents at minimal punishment $\mathrm{E}\left[D\pi\left(X\right)\right]$.

(Discussing alternative of having fixed cost per court case + linearity in footnote). When all rational agents are deterred from crime, minimizing $\mathrm{E}\left[D\pi\left(X\right)\right]$ amounts to minimizing the punishment of innocents. Define the *strength of evidence* $z \in \mathcal{Z}$ by the likelihood ration $s\left(z\right) \equiv \frac{\Pr[Z_1=z]}{\Pr[Z_0=z]}$, the probability of producing evidence z if the agent is guilty divided by the probability of producing z if the agent is innocent. As in **Becker1968Crime**, the optimal punishment rule is to punish at a high level when the strength of evidence crosses a certain threshold:

Proposition 12. *The* unique *optimal punishment rule can be written as*

$$\pi\left(f, z\right) = \begin{cases} L\left(f\right) & \textit{if } s\left(z\right) \geq B\left(f\right) \\ 0 & \textit{if } s\left(z\right) < B\left(f\right) \end{cases}$$

for functions B and L.

The intuition for this result—similar to Becker's argument—is that reserving punishment for the strongest levels of evidence can reduce wrongful convictions while keeping incentives unchanged.

oes machine learning impact incentives to follow the law? To analyze this, we consider the following punishment rule.

Definition 10. Define a *machine learning punishment rule* as

$$\pi\left(X\right) = \begin{cases} l & \text{if } \mathrm{E}\left[A \mid X\right] > k \\ 0 & \text{if } \mathrm{E}\left[A \mid X\right] \leq k \end{cases}$$

for constants $k \in [0, 1]$ and l.

The structure of punishing a constant amount l when predicted guilt crosses a certain threshold is chosen to mimick the optimal punishment rule. It turns out that such a machine learning punishment rule shares several properties with the optimal rule. In particular, an agent is punished by a machine learning punishment rule if the strength of evidence crosses a certain threshold:

Proposition 13. *An agent with $Z = z$ and $F = f$ is punished by a machine learning punishment rule iff*

$$s\left(z\right) > \frac{1 - \mathrm{E}\left[A \mid F = f\right]}{\mathrm{E}\left[A \mid F = f\right]} \frac{k}{1-k}$$

Also, as in the optimal rule, the threshold depends on the type f. Unfortunately, however, the machine learning punishment rule is not optimal. For now, we keep A fixed to look at the "short term" effects of a machine learning punishment rule on incentives.

A court case is brought against an agent $A = a$ and $X = x$ with probability $\gamma(a, x)$ and assume the algorithm . One important special case is *selection on observables*—when γ only depends on the observable information X and not on whether the agent in fact is guilty: $\gamma(x) \equiv \gamma(1, x) = \gamma(0, x)$. Informally, selection on observables assumes that potential plaintiffs who decide on whether to file a lawsuit do not have access to more evidence than what the algorithm receives. This is a natural benchmark, but could be violated if for some reason the evidence the plaintiff sits on can not be reliably conveyed to the court.[14] We then have the following results

Proposition 14. *XXX*

- Proposition X applies
- XXX Proposition X-X still holds under selection on observables and non-zero litigation probability.

As long as the

To build intuition, assume that there is a probability first consider the case

In this Section, we discuss complications that arise when (past) cases

[14] For instance, the evidence might be hearsay.