A Left Realist Critique of the Political Value of Adopting Machine Learning Systems in Criminal Justice

Fabio Massimo Zennaro[1 [0000−0003−0195−8301]]

Department of Informatics, University of Oslo, 0316 Oslo, Norway
fabiomz@ifi.uio.no

Abstract. In this paper we discuss the political value of the decision to adopt machine learning in the field of criminal justice. While a lively discussion in the community focuses on the issue of the social fairness of machine learning systems, we suggest that another relevant aspect of this debate concerns the political implications of the decision of using machine learning systems. Relying on the theory of Left realism, we argue that, from several points of view, modern supervised learning systems, broadly defined as functional learned systems for decision making, fit into an approach to crime that is close to the law and order stance. Far from offering a political judgment of value, the aim of the paper is to raise awareness about the potential implicit, and often overlooked, political assumptions and political values that may be undergirding a decision that is apparently purely technical.

Keywords: Machine learning · Supervised learning · Fairness · Left realism · Law and order.

1 Introduction

The success of machine learning systems in recent years (see, for instance, the breakthroughs in image recognition [23] or audio classification [20]) has led to a wide adoption of automated systems in several fields and applications, ranging from biomedical reseach to computer security [16].

The prototypical application of machine learning consists of *supervised learning*, that is learning a functional relationship between inputs and outputs by relying on samples demonstrating this relationship. Normally, such a relationship is inferred by tuning the learned model to be as close to data as possible, that is, practically, by optimizing the accuracy of its predictions. This simple learning process, designed to reflect a human inductive learning process, can be easily applied to a surprisingly large number of problems. For instance, any sort of decision-making may be reduced to the problem of taking a decision (output) as a function of a set of signals (input). In several instances, when deployed, this form of machine learning can achieve a level of accuracy that can equal or surpass human experts.

The enthusiasm generated by the success of these systems combined with their flexibility and applicability has led, most recently, to their deployment in socially-critical applications, including criminal justice [9]. From a formal point of view, criminal justice problems, such as sentencing or parole decisions, can be easily modeled in functional terms: for instance, deciding on a sentence or a parole may be seen as the result of a complex evaluation performed by a judge on a set of information. At first sight, it seems that machine learning systems could efficiently support or automate decision-making in criminal justice, maybe even improving its accuracy and removing human biases.

However, it has been noted that algorithms designed to optimize the accuracy of their model may lead to decisions that would not be considered fair from a social or legal point of view [6]. Decisions in modeling, choices in optimization, and biases implicit in the data inevitably cause learned models to be socially unfair. As a result, fair machine learning quickly developed as an active area of research that aims to tackle the problem of designing algorithms able to learn models that are not only accurate but also fair. Debate in fair machine learning has been actively concerned with the definition of fairness and its implications. For instance, [12] studied the cost of adopting fairness criteria in learning; [36] reviewed several measures of fairness; [10] and [22] derived trade-offs and impossibility results in satisfying multiple definitions of fairness; [34] reviewed the trade-off between accuracy and fairness; [27] evaluated how the enforcement of fairness may impact individuals in the long term; [35] explored the use of causal models in defining fairness.

Even if challenges in fair machine learning constitute an interesting and important strand of research, these topics are not the main focus of this work. In this paper, we would like to shift the level of discussion from *the fairness of adopted machine learning algorithms* to *the fairness of adopting machine learning algorithms*. In other words, instead of focusing on the fairness of the outcome suggested by an algorithm, we want to question the more fundamental decision of relying on machine learning systems in the specific setting of criminal justice.

Holding that any technological application has a political value, we try to investigate what political outlook may drive the adoption of machine learning in the field of criminal justice. While it may be suggested that such an adoption is just the necessary and unavoidable consequence of technological progress (i.e., machine learning systems may improve our decision making, therefore it is right and natural to adopt them), we want to argue, instead, that such a choice has a political valence, and, therefore, it should be discussed (also) in a political arena. Thus, *while elsewhere the adoption of machine learning in socially-sensitive fields may be taken for granted, in this paper we question this assumption and analyze the potential political significance that machine learning systems, as tools available to society, may carry.*

In particular, we will focus on the field of criminal justice because of its relevance to political decision-making and its active involvement with statistical models. Indeed, the use of statistical tools in this field has a long history [3], and debates about fairness are very active [6]. Here, disregarding specific definitions

of fairness, we will instead consider: *what is the political significance of using modern supervised learning systems in criminal justice? What are the hidden assumptions and potential implications of adopting such systems*?

Relying on the analysis and critique of the problem of crime proposed by *Left realism* [25], we consider the use of standard modern machine learning systems in the light of relevant criminological theories. We restrict our attention to supervised learning systems because of their success and wide adoption. We suggest that features of these systems such as causal-agnosticism, opacity, and reactive stance are particularly suited for a *law-and-order* approach to the problem of crime. Moreover, we draw an instructive analogy between the issue of deploying close-camera television (CCTV) recording systems in the 1980s and the contemporary adoption of machine learning systems. In conclusion, we discuss the limitations of our analysis and point out current developments in machine learning research that may be consistent with different political values.

This paper itself has no intention of expressing a political judgment of value with respect to the problem of adopting machine learning in the sphere of criminal justice (and in other fields). Its main aim is instead to raise awareness concerning the potential political weight of the adoption of machine learning systems in order to instigate an informed debate.

The paper is organized as follows. Section 2 introduces the main concepts in machine learning that are relevant to our discussion. Section 3 offers a presentation of the main criminological theory of interest, that is Left realism. Section 4 provides an analysis of machine learning systems through the conceptual categories of Left realism. Section 5 discusses our observations and their implications. Finally, Section 6 summarizes our contributions.

2 Machine Learning Systems

In this section, we provide a concise conceptual definition of the type of machine learning systems which are being discussed. Machine learning provides many models which can be used for learning in different contexts and which may significantly vary in their expressive and representational power (for an overview of different machine learning models, refer to standard machine learning textbooks, such as [28] or [7]). Given this variety, making sweeping general statements about machine learning systems would be impossible. We therefore focus on supervised learning systems as defined below.

Supervised learning systems. In this paper, we will focus primarily on supervised learning systems because of their easy deployability and their broad success and adoption. Supervised learning systems are statistical models that learn from data a complex functional relationship between an input, encoded as a set of quantitative features, and an output, representing a result dependent on the input. As *statistical* models, supervised learning systems are defined through a

set of assumptions and modeling choices that determine their domain of applicability and their level of abstraction [18]. By *complex* functional relationship we mean that these models learn a non-trivial function which, while commonly used, cannot be easily defined by a human designer (think, for instance, of facial recognition which can be easily performed by humans, thus implying the existence of a functional relationship between facial features and identity, but which can hardly be expressed in an algorithm by a programmer). Formally, a supervised learning system learns a model

$$y = f(\mathbf{x}),$$

where $\mathbf{x}$ is a vector of input features, y is the output, and f is the learned function. The function f is inferred from a data set containing a large number of samples of inputs and outputs. A machine learning algorithm processes each pair of inputs and outputs, $(\mathbf{x}_i, y_i)$, and tunes the learned function f accordingly. By default, most machine learning algorithms tune f by optimizing their accuracy, that is, by minimizing the discrepancy between the results produced by the learned model and results observed in the real-world:

$$\min \mathcal{L}\left(f(\mathbf{x}_i) - y_i\right),$$

over all samples i, and where $\mathcal{L}(\cdot)$ is a chosen loss function evaluating the difference between the predictions of the model and reality.

We take this essential and paradigmatic presentation of a learning system that infers a functional relationship from data via a loss optimization technique as our working definition of a supervised machine learning system. We thus ignore practical details about learning (e.g., how the parametric family over which we optimize is defined or how the evaluation of the degree of generalization is computed), implementation differences between the algorithms (e.g., whether we instantiate a simple linear regression or a neural network) and assumptions underlying concrete instantiations (e.g., independence of the samples, linear separability of the data). Elsewhere, this broad category of machine learning algorithms has been referred to with other terms, such as *regression systems* [3] or *function-based systems* [14].

Supervised learning systems in criminal justice. The adoption of statistical tools in criminal justice has a history that dates back to the 1920s (see [3] for a review of different generations of statistical models used by the criminal justice community).

The use of modern machine learning techniques is a current topic of debate in criminal justice [5, 1], especially in relation to fairness concerns [6]. Indeed, the deployment of supervised learning systems for tasks such as risk assessment and decisional support to judges has been accompanied by discussions on how to evaluate their social fairness. A representative case, that brought the topic to the attention of a wider public, was the Northpointe-ProPublica case. Northpointe was the developer of COMPAS, a predictive tool based on linear regression that

can be used to attribute a numerical risk of recidivism to a defendant [8]. COMPAS was deployed in Broward County, Florida, and the results it produced were analyzed by the nonprofit organization ProPublica, which showed the presence of a racial bias in terms of unequal false negative rates [2] (see [13] for a brief synthesis of the argument). The case of COMPAS is representative both of the types of machine learning systems we are considering, and of the kind of discussion we want to extend, shifting the attention from the question of the fairness of systems like COMPAS to the question of the political value of adopting such systems.

The actual effects and efficacy of machine learning in criminal justice has also been a topic of research [33]. Recently, a careful statistical study of the effect of using machine learning systems for supporting the decision making of judges has been carried out in [21]. These results are extremely valuable in the discussion about the adoption of supervised learning systems from a practical point of view, once politically-agreed quantitative measures of efficacy have been established; in contrast, our work will focus on the question of adopting machine learning systems from a theoretical perspective, that is, considering, before its concrete effects, what political value may be attached to the choice of relying on machine learning systems in criminal justice.

3 Approaches to Crime

In this section, we offer a brief review of the main criminological theory relevant to our work. The theory of Left realism was developed in the 1980s as a consistent alternative to the contemporary approaches to crime.

Theory of Left realism The foundations of Left realism were laid down by Lea and Young in their study and critique of the *law and order* approach to crime adopted in the UK in the 1980s [25]. Left realism was proposed as a new criminological stance in between *left idealism* (a stance biased towards seeing the criminal as the structural product of oppression in an unfair society) and *law and order* (a stance biased towards seeing the criminal as a deviant individual that must be confronted) [25]. Differently form left idealism, Left realism was founded on the central tenet that crime is a serious and real issue affecting everyday life. However, in distinction to law and order approaches, it strongly emphasized the complexity of crime, both in its causes and its prevention. It promoted a cautious and careful approach to crime data, aimed at avoiding simplistic and ungrounded readings and preventing mass-media distortion and moral panic. It suggested that crime and its causes should be examined in terms of discontent, marginalization, and sub-cultural group dynamics. More importantly, it advocated that crime should be fought in terms of deterrence and consensus policing, as a joint effort between communities and law enforcement, not as a battle carried on by the state through military policing [25]. Over time, attention to Left realism theories has lost traction, partly because of changes in the forms of crime with which society is concerned, and partly because of

an inappropriate application of its theories [24]. Despite this decline, however, discussion is still alive around Left realism and the contributions that it may offer to contemporary debates [15].

In the following, we will not present an organic revision or modern declination of Left realism for machine learning. This task, while interesting, is beyond the scope of this work. Rather, we appeal to some core ideas of Left realism and investigate how they can help us to better understand the political relevance of the adoption of machine learning systems in criminal justice.

4 Left Realist Critique of Machine Learning Systems for Criminal Justice

In this section we examine the central question of this paper, that is, what is the political relevance of adopting a machine learning system, paying special attention to the field of criminal justice. More precisely, we rely on the theory of Left realism to investigate where the adoption of machine learning systems in criminal justice would fall in the spectrum of criminological approaches ranging from law and order to Left realism. We examine this question by analyzing a set of issues raised and discussed in [25] that are particularly relevant to the adoption of machine learning systems.

4.1 Focus on Effects and Correlations

Left realism. Understanding the causes of crime is a central endeavor of Left realism. According to this system, a serious and successful approach to crime must move beyond the simple apprehension of crime to the uncovering of the causes underlying these behaviors[1]. The aim should be to determine which social factors (e.g.: marginalization, lack of political voice) are causally related to anti-social behaviors, so that effective crime policies may be defined in relation to these aspects[2]. This stance contrasts with more conventional approaches to crime, such as law and order, which prioritize fighting crime in itself and which are often uninterested in determining its actual causes[3]. Law and order approaches often explain away anti-social behaviors through the misuse of sociological categories[4] (such as condemning behaviors as psycho-pathological or under-socialized) or relying on simplistic explanations[5] (such as blaming the criminal as evil or lacking human values). From the perspective of Left realism, law and order tends to vainly struggle with the effects of social factors; that is, it focuses on the crime itself disregarding its origin; Left realism, instead, favors an approach that concentrates on the social factors that are seen as the underlying causes of crime.

[1] [25], p.265
[2] [25], p.74
[3] [25], p.265
[4] [25], p.77
[5] [25], p.95

Machine learning. Supervised machine learning systems such as those defined above are designed to model a direct relationship between a set of inputs and outputs. It is well known that, in statistical terms, standard supervised models learn correlations between inputs and outputs, and not causal relationships. A model $f(\mathbf{x}) = y$ may learn to predict the output y based on features that do not determine it. As long as what matters is a prediction, this may work fine; however, if we were interested in prevention through intervention, then acting on the correlated, but not-necessarily causative, features may not produce the desired result. Standard machine learning models are agnostic of causes; they process static sets of features with no explicit information about causal links. Understanding and analyzing causes is just beyond the concern of standard supervised learning systems.

Also, the common use of binary or discrete labels as outputs seems to comply with a methodology in which understanding subtle causal dynamics is disregarded; discrete categories may be seen as a way to bin cases or individuals into coarse classes; at the extreme, the use of binary output classes may be interpreted as a way to partition cases or behaviors into normal and deviant (or psycho-pathological or under-socialized or evil), with no interest for uncovering deeper dynamics.

Overall, then, supervised learning systems seem to fit better an approach to crime which is more concerned with tackling well-categorized effects by predicting them, instead of engaging and working on the causes of the crimes.

4.2 Focus on Specific Crimes

Left realism. A starting point of Left realist thought is the awareness of the political value undergirding the definition of crimes; deciding which crimes to focus on, and how to delimit them, are choices that are inevitably bound to shape, and to be shaped by, political programs and public opinion[6]. The very decision to focus on violent crimes or crimes against property, instead of highlighting, say, "white-collar" crimes or financial crimes, is a choice that Left realism tries to bring to the forefront[7]. While not arguing against the importance of dealing with violent crimes, Left realism points out that the definition of what constitutes violent crimes has profound effects on public perception and policies. Official definitions and public opinion may not match and, at times, diverge, as shown, for instance, in the gap between official statistics (based on given definitions of crimes) and self-victimization reports (based on the understanding of crime by a victim)[8].

Machine learning. By definition, supervised machine learning is very dependent on data. Limitations on the available data sets may severely restrict the subdomains of criminal justice to which machine learning may be applied. Following

[6] [25], pp.11, 68
[7] [25], p.65
[8] [25], p.17

the actual concerns of criminal policies and the definitions of crime previously approved, more data may be available about certain crimes than others (e.g., more data is generally available about violent crimes than financial crimes), and this may improperly justify the adoption of machine learning in those particular sub-domains. What should be a political decision may be implicitly (and fallaciously) justified by the actual availability of data, which may have been determined by entrenched definitions of crime. This, in turn, leads to generation of more data in those specific fields that are the concern of the policymaker.

It is important to remember that algorithms cannot learn but what is provided to them through the data; a given definition of crime, encoded in the pairing of inputs and outputs, will necessarily inform all the results produced by a machine learning system. In their working, supervised learning systems conform and reiterate given definitions. While their accuracy may be quantitatively measured and celebrated, it is always an accuracy with respect to definitions that are often implicitly hidden in the preparation of the data or in the setup of the algorithm. Disagreement or misunderstanding of these definitions may once again give rise to a gap between machine learning results and other reports such as self-victimization reports.

In conclusion, supervised learning systems may feed a self-reinforcing loop in which pre-existing ideas and definitions of crimes are constantly re-affirmed by algorithms trained on the same concepts, making it progressively harder to discuss and challenge the existing and efficient "working" definitions.

4.3 Sensitivity to Data Interpretation

Left realism. Related to the issue of the dependence of statistics on the definition of crimes, Left realism also argues against simplistic readings of statistics. While statistics are precious resources for studying crime, they should never be taken as hard facts; instead, they should be interpreted with special care for understanding the assumptions and the conditions under which such statistics were generated[9]. Data and results are often characterized by complex behaviors; for instance, special attention should be devoted to aggregated statistics, as they may hide highly biased or skewed distributions with respect to sensitive parameters, such as gender or race[10]. Superficial readings may be responsible for inadequate decisions or may be used opportunistically to justify political choices.

Machine learning. Results produced by supervised machine learning systems are also non-trivial to analyze. The outcome of the learning process critically depends on the data provided and on the statistical assumptions defining the behavior of the system. All machine learning algorithms have limitations in their modeling power. Their capacity to deal with complex data, such as highly-skewed data or multi-modal data, strictly depends on the specific implementation adopted. A proper model choice would require a careful study of the data at hand, as

[9] [25], p.12
[10] [25], p.28

well as a deep understanding of the assumptions and the limits of the selected supervised algorithm. Unfortunately, supervised learning algorithms are often applied as black-boxes to data that do not conform to their assumptions (as may be in the case of highly biased criminal justice data sets), and this may lead to grossly approximated conclusions that hide, instead of uncover, the subtlety of the data.

To make things worse, supervised learning systems tend to return non-transparent shallow results in the form of a numeric value that can be readily used for decision-making. Such a simple output, however, combined with the opaqueness of many modern supervised learning systems, tends to hinder the possibility of properly understanding and interpreting the result. After carrying out the hard and menial work of crunching data for us, a supervised learning system often does not provide a transparent or justified output. Despite the job performed by the machine, the crucial and sensitive part of interpreting the data and taking a responsible decision cannot be delegated to a machine. Unfortunately, though, full understanding of the results of a supervised system is often not possible.

A supervised learning system may then be easily exploited and presented as an oracle able to perform a complete and reliable statistical analysis, inviting a superficial acceptance of its output instead of stimulating an engaged interpretation.

4.4 Tool for Military Policing

Left realism. A central concern for criminology and criminal justice relates to policing. In its critique, Left realism identifies two abstract and opposite types of policing. The form of *consensus policing* it advocates is based on a strict and beneficial cooperation between police forces and the community; in this setting, information is voluntarily provided by the community, and police may act with the support and in the interest of the locals[11]. On the opposite end, the form of *military policing* enacted by law and order approaches is based on a unilateral enforcement of law by the police in a context where the cooperation with the community is reduced or has been severed. Deprived of information sources within the community, police have to carry out complex and costly investigations and then act on the uncertain conclusions achieved without local support[12]. According to this perspective, mistakes and prejudices lead to a self-reinforcing loop of antagonizing the masses at large, mobilization of neutral bystanders, and alienation within the community, all of which, in turn, progressively isolate police forces[13].

Machine learning. Machine learning systems can be powerful tools for the definition of security policies and for policing. Referring to the policing spectrum identified by Left realism, supervised learning systems seem to lean towards one of

[11] [25], p.169
[12] [25], p.172
[13] [25], p.182

the two extremes. Modern learning systems seem weakly related to the paradigm of consensual information gathering from the community: learning systems are not designed to ease the relationship between locals and police, nor they are apt to integrate varied data acquired from heterogeneous data sources; instead, they are meant to process uniform data that are acquired in a standardized fashion with or without explicit consent. As tools developed to process data and improve decision-making, supervised learning systems seem more suited at enhancing the investigative expertise of a police force that has been reduced to work on data it gathered by itself. Supervised learning systems can indeed become efficient tools to improve the accuracy and the precision of military policing; at the same time, though, because of the mistakes they are bound to commit, they may cause an increase in the distance between police forces and local communities.

Between the two alternative approaches to policing presented above, current supervised learning systems seem more fit to a military policing approach than a consensus one.

4.5 Issues of Accountability

Left realism. Connected to the issue of policing is the problem of trust and accountability of law enforcement. Left realism asserts that transparent policies are a necessary requirement to guarantee a democratic overview and control of the activities of police forces[14]. Through accountability, a sense of mutual trust and confidence can be built between communities and state representatives. In contrast to Left realism, other approaches often present several arguments, including the technical nature of decisions related to policing or the necessary secrecy of some operations, in order to justify the opaqueness and the autonomy of police forces[15]. The reduction of policing to a technical question of efficiency tends in turn to overshadow the question of accountability and the role that politics should have in defining policing[16].

Machine learning. Machine learning raises new and challenging questions about accountability and trust. Most of the current successful supervised learning systems are opaque, behaving like *black boxes* that, provided with an input, return an output without any explanation or justification for it. The result is often the product of an optimization process with respect to a simple and quantifiable definition of accuracy or efficiency. The problem of interpreting the dynamics and the outputs of supervised learning is a problem relevant to many fields beyond criminal justice and is now a very active area of research in machine learning (see, for instance, [26] for a discussion of the definition of interpretability of machine learning models). Without a way to explain results, supervised learning systems may end up diluting accountability, making it hard to trace responsibility through its opaque internals. Currently, trust is not built upon an

[14] [25], p.269
[15] [25], p.233
[16] [25], p.257

understanding of the systems, but over a technical confidence in their efficiency. Such trust is, however, bound to fall if the decisions of the system were to be questioned or put into discussion.

The adoption of current black-box supervised learning systems poses a strong challenge to any form of democratic oversight: unless such systems are carefully validated in their assumptions and their definitions, their results may be hard to assess. This may turn machine learning into another technical tool that can be used to justify policing decisions not being disclosed and discussed within the community.

4.6 Analogy with CCTV

Left realism. Expanding on the topic of policing within criminal justice, a telling parallel may be drawn between the adoption of close-camera television (CCTV) for surveillance in the 1980s and supervised learning systems in the current time. Interestingly, in its analysis of the causes of a rise in military policing in the UK in the 1980s, Left realism identified, beyond an increase in street lifestyle and a rise in prejudices, the widespread adoption of new technologies[17]. New technological resources, such as CCTV, promoted among police forces a "fire-brigade" mentality: instead of being present among the community, an officer could monitor its neighborhood from afar and intervene only when and where necessary[18]. Thanks to the simplicity of interacting with these surveillance devices, CCTV often became the source information of choice, thus favoring the development of a distant and reactive model to policing instead of an integrated and proactive model based on a constant presence among the community. Ahead of times, it was foreseen that this attitude would lead to the development of technologies, such as computerized preventive tools, that would rely on collecting and storing vast amount of data about citizens[19] and which would naturally raise ethical, legal, and political questions.

Machine learning. Supervised machine learning algorithms are one of the most prominent modern technologies currently deployed in criminal justice. Like CCTVs, these systems generally foster a "fire-brigade" mentality: they offer the possibility of understanding and controlling the community remotely; and, like any technological innovation, they promise unprecedented accuracy and success whenever intervention is necessary. However, the side effect of this development, as it was in the case of CCTVs in the 1980s, is that criminal policies may end up relying more and more on machine-processed data instead of information volunteered by locals, thus further deepening the rift with the community.

It is also clear that modern supervised learning systems meet the prediction about the craving for data. They frequently need to acquire large amounts of data to be trained, to the point that often the term *big data* systems is just used

[17] [25], p.179
[18] [25], p.181
[19] [25], p.243

as a synonym for many contemporary machine learning systems. As foreseen, this hunger for data is the source of ethical and political debates, concerning, for instance, the scope, the transparency and the accountability of scoring systems [11]. Debate about the privacy of users and the extent of legal use of their data constitute the topic of relevant and current discussion in the political and economical arena.

In conclusion, there are significant similarities between the adoption of CCTV in the past and the current trend of the adoption of supervised learning systems in the present days. The doubts and the questions about trust, accountability, and control raised by the deployment of CCTV should be asked for supervised learning systems as well. The reflections and the answers about the political values (such as, privacy and security) raised in the debate about CCTV may enlighten similar evaluations on the political implications of adopting supervised learning systems today.

5 Discussion

In the previous sections we saw how a Left realist critique may be used to analyze the decision of adopting machine learning from a political standpoint. Several features of supervised learning algorithms (focus on effect, restriction to certain crimes, favoring investigation over cooperation with the community) seem to align their adoption in the field of criminal justice to a law and order political view. Table 1 offers a simplified and essential overview of the connections we drew between assumptions and features of supervised learning systems in machine learning and their potential political value or meaning interpreted using the theory of Left realism. This analysis of the political significance of supervised learning systems in criminal justice is, of course, far from being exhaustive. We focused our analysis on those aspects that have an overt parallel with observations and critiques offered by Left realism in [25]. However, other more technical aspects of supervised learning systems, such as the assumption of stationarity of data, the definition of a loss function with its terms and constraints, or the assumption of independence of the data samples, could also be investigated for their political implications.

In general, a supervised learning system constitutes an abstract representation of a phenomenon, in this case a criminal justice process. The reduction of a complex reality to a mathematical model often requires strong assumptions and coarse simplifications. Modeling in a socially sensitive context opens the space for political debate. Even if the primary motivation in modeling is practicality and efficiency, the decision of considering or omitting certain aspects of the problem has a political relevance. In contexts such as criminal justice, any choice, from the decision to collect data in a certain environments to the definition of output classes, may be interpreted in a political light.

Despite its limits, this study reveals a potential implicit political bias in the decision of adopting supervised learning systems in criminal justice. Paraphrasing the infraethics position [19], this bias may follow from the fact that machine

Table 1. Summary of the political value or meaning of some of the features of supervised learning systems, as they were analyzed in this paper.

ML assumption or feature	***Political value or meaning***	***Section***
Working on correlations	Disregard of actual cause-effect links	4.1
Coarse categorical outputs	Oversimplification of actual sociological/criminological explanations	4.1
Availability of data for limited problems	Restricted political concern with only certain types of crime	4.2
Dependence on given labelling	Implicit enforcement of certain definitions of crime	4.2
Sensitivity to data interpretation	Possible instrumental misinterpretation of the data	4.3
Automatic support for decision making	Possible complete delegation of decision to a legalistic algorithm	4.3
Functional relationship input-output	Better support for military policing rather than consensus policing	4.4
Lack of interpretability	Possible promotion of certain policies solely on the ground of efficiency	4.5
Opaque internals	Dilution of responsibility for the choices of the algorithm	4.5
Remote fast processing of data	Better support for enforcement of "fire-brigade" mentality	4.6
Reliance on big data	Justification for collection of large amounts of data	4.6

learning and the decision of adopting supervised learning systems cannot be a purely politically-neutral choice; instead, as with every technical decision, it inevitably embeds, even in a minimal way, certain values, and thus favors certain choices, becoming a tool to promote political agendas.

An awareness of this reality is important in order to make critical decisions about the adoption of machine learning in sensitive fields like criminal justice. This understanding would allow us to reflect more clearly on the issue of using supervised learning systems by helping us to avoid at least two mistakes.

(i) By uncovering the political value of adopting machine learning systems, it would prevent us from making the *naive* mistake of adopting these systems simply on the ground of efficiency and enthusiasm.

(ii) It would prevent the mistake of accepting the *instrumental* use of technical arguments as justifications to make certain political decisions in the arena of criminal justice more acceptable. Indeed, a stance like law and order may be dangerously and fallaciously defended by appealing to the technicalities we have discussed: focus on crimes instead of their causes may be justified by the nature of most supervised systems ("efficient machine learning systems can only deal with correlations, not causes"), restriction to certain crimes may be defended in terms of data availability ("we can tackle only those crimes for which we have data"), the use of simplified definitions and statistics may be explained in terms of historical data ("we can tackle only crimes as we have observed them until now"), a reactive stance may be motivated over consensus policing by the features of predictive systems ("machine learning systems are designed to improve

police investigation"), and opaque decisions may attributed to the intrinsic non-transparency of supervised systems ("results are effective but we cannot explain them").

It is important, though, to underline the limits of the applicability of these considerations. The observations of political value made here apply particularly well to supervised machine learning as we have defined it. However, although supervised learning systems are by far the most widely adopted, these statements can be hardly extended to machine learning in general. Different approaches may be susceptible only to some of the critiques presented in this paper.

For instance, the use of machine learning systems based on the theory of *causality* [31, 32] may be immune to the criticism of focusing only on correlations and effects (see Section 4.1). While agreement on the definition and the identification of causes may be debatable as there may be disagreement over the causal models to consider or the causal assumptions to accept, a causality-based system would allow not only to work with effects and predictions, but also with causes and policies. Arguments in favor of adopting a causal inference framework, such as the one proposed in the insightful analysis of [3], may be indeed read as addressing some of the political concerns of Left realism.

Other forms of machine learning may be seen as addressing other concerns expressed in this paper. *Bayesian machine learning* [4] models outputs and their uncertainty in the form of probability distributions, thus tackling, in part, the problem of working with discrete outputs (see Section 4.1); *transfer learning* [30] studies how to exploit data in a source domain in order to learn in a target domain, thus opening the possibility of deploying learning systems in domain where data are scarce (see Section 4.2); similarly, improvements in *statistical modeling* may, for instance, be used to avoid the narrow focus on few specific crimes following from positive feedback effects [17] (see Section 4.2); *interpretable machine learning* [29, 26], comprising simple understandable algorithms or methods aimed at opening black-boxes, may potentially offer a future solution to the problem of opaqueness and lack of trust (see Section 4.5).

All these technical efforts need to be analyzed more closely and deeply in order to assess their contributions and their value from a political viewpoint.

6 Conclusions

In this paper we offered an analysis of the choice of adopting supervised learning systems in criminal justice as a political decision, reviewed through the lens and the categories of Left realism. Our aim was not to promote a particular political stance, but, rather, to raise awareness about the political value of a choice that, at first sight, may look purely technical and apolitical. An informed debate about the opportunity to adopt supervised learning systems in criminal justice should then revolve not only around the question of their efficiency and fairness in a specific setting, but also on the question of which sort of political project they

endorse more generally. Adopting machine learning systems in sensitive fields should be not just a question of social fairness, but also of political values.

Already in the 1980s, [25] observed that technological developments did not solve the problems posed by crime; instead they made decisions about the adoption and the use of technology *more political*[20]. The adoption of machine learning, in the present days, in the field of criminal justice and beyond, deserves to be also considered and analyzed in a political light. In any social field, the choice of using a supervised learning system is, in itself, already a political decision.

References

1. Andrews, D.A., Bonta, J., Wormith, J.S.: The recent past and near future of risk and/or need assessment. Crime & Delinquency **52**(1), 7–27 (2006)
2. Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine bias. ProPublica, May **23** (2016)
3. Barabas, C., Dinakar, K., Virza, J.I., Zittrain, J., et al.: Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. arXiv preprint arXiv:1712.08238 (2017)
4. Barber, D.: Bayesian reasoning and machine learning. Cambridge University Press (2012)
5. Berk, R.: Criminal justice forecasts of risk: A machine learning approach. Springer Science & Business Media (2012)
6. Berk, R., Heidari, H., Jabbari, S., Kearns, M., Roth, A.: Fairness in criminal justice risk assessments: the state of the art. arXiv preprint arXiv:1703.09207 (2017)
7. Bishop, C.M.: Pattern recognition and machine learning. springer (2006)
8. Brennan, T., Dieterich, W., Ehret, B.: Evaluating the predictive validity of the compas risk and needs assessment system. Criminal Justice and Behavior **36**(1), 21–40 (2009)
9. Brennan, T., Oliver, W.L.: The emergence of machine learning techniques in criminology: Implications of complexity in our data and in research questions. Criminology & Public Policy **12**(3), 551–562 (2013)
10. Chouldechova, A.: Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data **5**(2), 153–163 (2017)
11. Citron, D.K., Pasquale, F.: The scored society: due process for automated predictions. Wash. L. Rev. **89**, 1 (2014)
12. Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., Huq, A.: Algorithmic decision making and the cost of fairness. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 797–806. ACM (2017)
13. Courtland, R.: Bias detectives: the researchers striving to make algorithms fair. (2018)
14. Darwiche, A.: Human-level intelligence or animal-like abilities? arXiv preprint arXiv:1707.04327 (2017)
15. DeKeseredy, W.S., Donnermeyer, J.F.: Contemporary issues in left realism. International Journal for Crime, Justice and Social Democracy **5**(3), 12–26 (2016)
16. Deng, L.: A tutorial survey of architectures, algorithms, and applications for deep learning. APSIPA Transactions on Signal and Information Processing **3** (2014)

[20] [25], p.242

17. Ensign, D., Friedler, S.A., Neville, S., Scheidegger, C., Venkatasubramanian, S.: Runaway feedback loops in predictive policing. arXiv preprint arXiv:1706.09847 (2017)
18. Floridi, L.: The Philosophy of Information. OUP Oxford (2013), https://books.google.it/books?id=l8RoAgAAQBAJ
19. Floridi, L.: Infraethics–on the conditions of possibility of morality. Philosophy & Technology **30**(4), 391–394 (2017)
20. Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A.r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Kingsbury, B., et al.: Deep neural networks for acoustic modeling in speech recognition. IEEE Signal processing magazine **29** (2012)
21. Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., Mullainathan, S.: Human decisions and machine predictions. The quarterly journal of economics **133**(1), 237–293 (2017)
22. Kleinberg, J., Mullainathan, S., Raghavan, M.: Inherent trade-offs in the fair determination of risk scores. arXiv preprint arXiv:1609.05807 (2016)
23. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
24. Lea, J.: Left realism: A radical criminology for the current crisis. International Journal for Crime, Justice and Social Democracy **5**(3), 53–65 (2016)
25. Lea, J., Young, J., et al.: What is to be done about law and order? (1984)
26. Lipton, Z.C.: The mythos of model interpretability. arXiv preprint arXiv:1606.03490 (2016)
27. Liu, L.T., Dean, S., Rolf, E., Simchowitz, M., Hardt, M.: Delayed impact of fair machine learning. arXiv preprint arXiv:1803.04383 (2018)
28. MacKay, D.J.: Information theory, inference, and learning algorithms. Cambridge University Press (2003)
29. Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., Yu, B.: Interpretable machine learning: definitions, methods, and applications. arXiv preprint arXiv:1901.04592 (2019)
30. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on knowledge and data engineering **22**(10), 1345–1359 (2009)
31. Pearl, J.: Causality. Cambridge university press (2009)
32. Peters, J., Janzing, D., Schölkopf, B.: Elements of causal inference: foundations and learning algorithms. MIT press (2017)
33. Richardson, R., Schultz, J., Crawford, K.: Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. New York University Law Review Online, Forthcoming (2019)
34. Wick, M., Tristan, J.B., et al.: Unlocking fairness: a trade-off revisited. In: Advances in Neural Information Processing Systems. pp. 8780–8789 (2019)
35. Wu, Y., Zhang, L., Wu, X., Tong, H.: Pc-fairness: A unified framework for measuring causality-based fairness. In: Advances in Neural Information Processing Systems. pp. 3399–3409 (2019)
36. Zliobaite, I.: A survey on measuring indirect discrimination in machine learning. arXiv preprint arXiv:1511.00148 (2015)