# Maintaining Accountability in a Criminal Justice System that Uses Machine Learning

**A Research Paper submitted to the Department of Engineering and Society**

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Adam Klein
Spring, 2020

On my honor as a University Student, I have neither given nor received
unauthorized aid on this assignment as defined by the Honor Guidelines
for Thesis-Related Assignments

## Algorithms in the Criminal Justice System

The Supreme Court defines probable cause as existing “where the facts and circumstances” are sufficient “to warrant a man of reasonable caution in the belief that an offense has been or is being committed” (*Brinegar v. United States*, 1949). The use of the word “man” and the standard of being “reasonable” in his “belief” imply that human judgment is essential to establishing probable cause. Law enforcement officers are held accountable by standards of proof such as probable cause. However, what would happen if a computer told a police officer that there was a high probability of a crime occurring on their current patrol route? Can the computer’s insight be used as a basis for probable cause? How does a society evaluate whether the computer’s obscured calculations meet the standard of a “reasonable belief”?

Increasingly sophisticated software packages are being used to help make decisions about arrests and paroles in the criminal justice system. As these complex decisions become increasingly influenced by machines, it becomes less clear how to uphold a standard of proof that is defined in terms of human reasoning. This shift in the way decisions are made necessitates a re-evaluation of how officers of the criminal justice system are held accountable for their actions.

## Research Question & Methods

This research investigates the following question: does the use of machine learning in the criminal justice system erode the accountability of law enforcement agencies in the United States? The answer to this question provides insight into how this technology can be ethically utilized in the realm of the criminal justice system. In order to find an answer to this question, this paper first investigates the merits of using algorithmic tools in the criminal justice system. This analysis utilizes case studies from scholars, such as Berk and Mohler, who argue the benefits of using predictive policing software and recidivism forecasters. These sources make claims that this technology is more effective and more objective than human actors, thereby reducing the issue of accountability. Contradictory data that demonstrates bias in predictive policing software, such as PredPol, is then incorporated to help assess the claims of scholars who advocate its usage. Together, these sources are used to reach a conclusion on whether or not algorithmic decision aids make human actors in the criminal justice system more or less accountable for their actions.

Second, this paper conducts a policy analysis on existing legal standards for accountability in the criminal justice system. Since the term accountability has no legal definition, this analysis reasons about how accountability is achieved by looking at policies related to standards of proof for certain legal actions. Specifically, case law related to probable cause and reasonable suspicion is used to gain insight into how the judiciary expects officers of the criminal justice system to act with accountability. By establishing how accountability is established in the current criminal justice system, this policy analysis indicates whether accountability can still be established in a system that employs machine learning.

## Background

This paper focuses on two specific types of machine learning technology being used in the criminal justice system – predictive policing software and recidivism forecasters. Predictive policing software helps police officers determine where crimes are likely to occur. This can be used to predict "hotspots" that inform police patrol routes (Mohler et al, 2015). Recidivism forecasters are used to help make decisions about whether or not to parole prison inmates. A recidivism forecaster makes predictions about how likely an inmate would be to commit another crime if he or she were granted parole (Berk, 2017).

As the technology described above demonstrates, algorithms are being used to aid decision-making in the criminal justice system (Liptak, 2017). However, studies have revealed that these techniques can promote racially discriminative practices (Lum & Isaac, 2016). In the United States, this is occurring in the context of a criminal justice system that already produces racially disproportionate outcomes. For example, statistics show that Black and Hispanic populations are incarcerated at notably higher rates than white populations (Carson, 2018). Accountability is necessary to prevent discriminative practices. However, the use of algorithmic policing technology could make actors in the United States' criminal justice system less accountable by providing them with an opaque mathematical process that makes decisions seem more rational and scientific than they truly are (Ananny & Crawford, 2018).

Machine learning technology is used under the general assumption that it allows law enforcement officers to make more accurate decisions that are less subject to human error. However, there is also evidence that suggests the contrary. Lum and Isaac demonstrate that the use of PredPol, a leading predictive policing software package, can simultaneously increase arrest rates and increase bias towards certain neighborhoods (Lum & Isaac, 2016). This presents a conflict. Algorithms are presented as a way to help us make wiser decisions. However, research suggests that they can actually influence us to systematically make harmful decisions. Accountability is essential if we are to curtail this institutionalization of bias.

## STS Framework

This research is organized through Actor-Network Theory, which helps characterize the social relationships that arise from the interactions between technology, people, and ideas (Whittle & Spicer, 2008). By observing the social structures that arise between actors, we can analyze how a technology could impact stakeholders. One of the originators of Actor-Network Theory is Bruno Latour. Latour argues that Actor-Network theory has the advantage of extending our analysis of a society to non-human actors. An actor-network is all-encompassing – "there is nothing but networks, there is nothing in between them," (Latour, 1996). The fact that an actor-network encompasses non-human components is particularly beneficial in a discussion of how algorithmic tools impact the criminal justice system. By definition, accountability is about holding humans responsible for their actions. However, technological components, such as software and datasets inform human actors. Thus, actor-network theory provides a way to capture the influence that algorithms have on human decisions.

Actor-Network Theory is important to this discussion because it helps capture the complex relationships between different pieces of the criminal justice system. The criminal justice system already consists of many actors that range from humans to organizations to technologies. The use of predictive software introduces new actors into this system. These include both the algorithms that form the basis of the software and the data sets that serve as inputs to them. Data sets themselves are potential sources of bias since they are curated by humans (Kirkpatrick, 2017). Thus, the process of data collection and maintenance is an important component that can be modeled by this network. This paper looks at the actor-network of the criminal justice system as it has historically existed without algorithmic tools. It then introduces these new actors to determine whether or not this changes the standard for accountability.

## Results & Discussion

The results of this research find that the use of machine learning to aid in law enforcement activities erodes the accountability of law enforcement officers in the criminal justice system. The Supreme Court of the United States has defined in many instances that law enforcement officers must be able to justify their actions through the articulation of facts and inferences. Machine learning tools that inform these actions are a black box in a decision-making process. There must be a mechanism by which a piece of software is forced to explain how it reached a conclusion. Without this, actors in the criminal justice system will be capable of committing actions without needing to articulate the entire chain of reasoning that justified them. These findings necessitate the development of a regulatory framework that ensures algorithms are not used to make decisions about arrests and incarcerations without their effects being properly understood first.

### Bias in Predictive Policing Software and Recidivism Forecasters

To begin, it is necessary to refute the idea that machine learning produces omniscient, unbiased decisions. Proponents of this technology claim that it is more effective and more objective than human actors, thereby reducing the issue of accountability. In contrast, this paper argues that any human biases percolate into these tools due to the existence of an interdependence between humans and technology in the actor-network of the criminal justice system. The presence of these biases leads to an erosion of accountability, as algorithmic actors provide human actors pseudoscientific justification for unjust practices.

Proponents of predictive policing software and recidivism forecasters argue that they allow the criminal justice system to make more accurate decisions that are less subject to human bias and error. Mohler demonstrates that patrolling crime “hotspots” produced by predictive policing algorithms reduces crime rates (Mohler et al., 2015). Similarly, Berk argues that recidivism forecasters decrease the number of inmates who are arrested again by allowing parole boards to make better decisions about nonviolent offenders (Berk, 2017). These are statistics that are cited to support the idea that the decisions made by this technology are more accurate and impartial than human decisions.

However, the notions of increased accuracy proposed in the previous paragraph are based on statistics such as higher arrest rates, which are not necessarily indicative of a wiser, more impartial criminal justice system. Lum and Isaac demonstrate that the use of PredPol, a leading predictive policing software package, can simultaneously increase arrest rates and increase bias towards certain neighborhoods (Lum & Isaac, 2016). Specifically, they demonstrate that drug-related arrests would increase in less-affluent Oakland neighborhoods, even though drug offenses are projected to be evenly distributed throughout the city (Lum & Isaac, 2016). The key insight from this is that arrest rates are an incomplete picture. Higher arrest rates do not imply "better" decisions, and they certainly do not imply a lack of bias.

The argument that predictive policing software and recidivism forecasters are more accountable because they follow an objective, mathematical process is fundamentally flawed due to the fact that this technology is informed by human-curated datasets. Kirkpatrick demonstrates how biased data can cause predictive policing and recidivism forecasters to produce biased results (Kirkpatrick, 2017). Lum and Isaac's simulation is an example that demonstrates that predictive policing enhances unfair bias towards certain neighborhoods due to existing data on arrests in these neighborhoods (Lum & Isaac, 2016). The technology used by law enforcement agencies therefore cannot be viewed as an unbiased, scientific process, but rather as a technological actor that is inherently influenced by the values and biases of human actors.

The notion of machine learning as being more accurate, and therefore more accountable, than traditional human analysis is further complicated by the fact that conflicting social values inform the different metrics used to evaluate predictive policing and recidivism forecasting. Mohler argues in favor of predictive policing by showing higher arrest rates (Mohler et al., 2015). Lum and Isaac argue against predictive policing by demonstrating arrest rates that are less equitable (Lum & Isaac, 2016). Looking at what these authors choose to focus on shows a tension between social values, such as public safety and fairness. Along the same lines, Berk supports recidivism forecasting by arguing that it reduces the number of inmates who are arrested again (Berk, 2017). This metric is similar to the one used by Mohler. While the number of inmates arrested again may go down, this analysis tells us nothing about the mix of people that receive parole, and could still fail when using the metrics employed by Lum and Isaac.

The values that inform these metrics are important actors in the network of the criminal justice system, as they indicate how the public expects the system to function. The notion of accountability is inherently tied to these values, since a society uses its value system to define standards to which people should be held accountable. Introducing a technological actor, such as predictive policing software, can cause people to believe these values are no longer relevant under the false assumption that a more objective, scientific process is being followed.

Figure 1 below shows a basic model of the criminal justice system without considering technological actors. The public, the judiciary, and law enforcement are considered as groups of human actors. Accountability is considered an actor itself, as the notion of accountability encapsulates the values and mechanisms by which all three groups hold one another responsible for their actions.

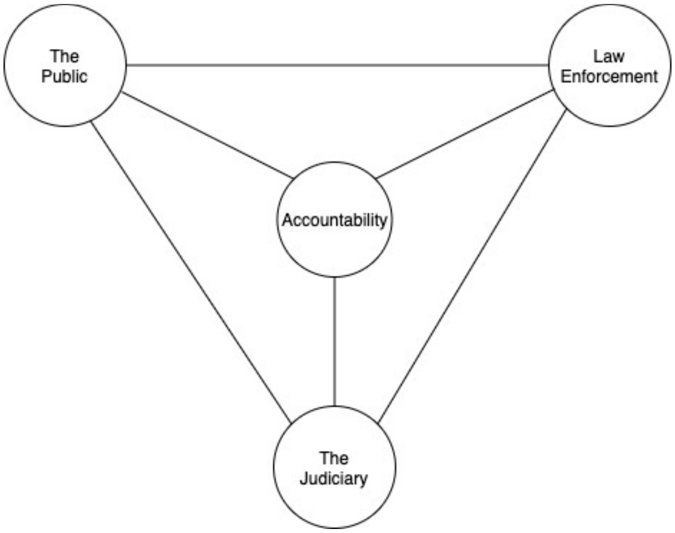

*Figure 1: Actor-Network of the Criminal Justice System without Machine Learning*

Figure 2 models the impact of introducing machine learning tools to the criminal justice system. These tools undermine the existing relationships in the network and obfuscate the actions of law enforcement by reducing the ability to directly scrutinize law enforcement actions. Likewise, data sets are an additional actor that cannot be easily scrutinized for the purposes of accountability. The potential instability created by these new relationships is analyzed in the following sections of this paper.

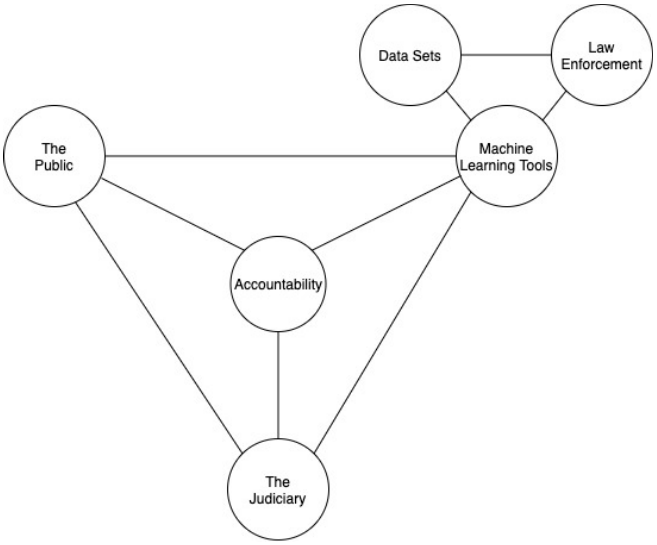

*Figure 2: Actor-Network of the Criminal Justice System with the Introduction of Machine Learning*

## Accountability in the Actor-Network of the Criminal Justice System

At this point, this research has established that technical processes, such as predictive policing and recidivism forecasting, do not erase the issue of accountability. This research now further argues that making decisions inside of an algorithmic black box fundamentally conflicts with existing legal standards for accountability in the criminal justice system. First, the relationship between human actors and accountability in the current network of the criminal justice system must be established. We will then introduce machine-learning tools as actors, and demonstrate how these undermine the relationships that uphold the value of accountability.

The value of accountability in the United States' criminal justice system is defined in a way that inseparably ties it to the logic of human decision processes. Accountability is not a term that has a legal definition, but one can reason about how accountability is achieved by looking at standards of proof for certain legal actions. For example, the Supreme Court defines probable cause to exist "where the facts and circumstances" are sufficient "to warrant a man of reasonable caution in the belief that an offense has been or is being committed" (*Brinegar v. United States*, 1949). The use of the word "man" and the standard of being "reasonable" in his "belief" imply that probable cause is defined in a way that is inextricably tied to human judgment. Similarly, reasonable suspicion is a weaker legal standard of proof that allows a police officer to conduct a search of a person (*Terry v. Ohio*, 1967). The basis for reasonable suspicion is defined as "specific and articulable facts" in conjunction with "rational inference from those facts" (*Terry v. Ohio*, 1967). The words "articulable" and "rational inference" demonstrate that accountability in the context of the Fourth Amendment inherently depends on human explanation. This case law describes a network in which actors in law enforcement verbally justify their actions, and actors in the judiciary hold them accountable by deciding whether their justification meets defined standards.

The introduction of machine learning tools into the criminal justice system inherently undermines this model of accountability by allowing law enforcement to black box part of its decision process. Illinois v. Wardlow expands on the definition of reasonable suspicion by concluding that being in a "high-crime area" is relevant context to inform a police officer's decision to search someone (*Illinois v. Wardlow*, 1999). Opponents of predictive policing accuse it of creating "feedback loops" that are biased towards certain communities, since greater police presence leads to more arrests and the designation of a "high-crime area" (Bennett Moses & Chan, 2018). Using predictive policing tools can thus undermine the standard of an "articulable" suspicion (*Terry v. Ohio*, 1967), since one of the factors that may inform a police officer's conduct is a black box that follows a potentially incomprehensible decision process. The network that we previously established now has a weakened value of accountability, as the justifications that law enforcement agencies provide to the judiciary can be filtered through algorithmic actors that obfuscate them.

**The Public in the Actor-Network of the Criminal Justice System**

We have argued that applying machine learning to the criminal justice system undermines accountability in terms of the legal standards defined by the judiciary. We now argue about the role of the general public as an important actor in upholding the accountability of both law enforcement agencies and the judiciary. The central way in which the public holds a system accountable is by keeping that system transparent. However, in systems that involve algorithmic components, transparency is difficult to achieve. For this reason, we argue that the use of machine learning directly undermines the primary framework by which the public is able to maintain accountability in the criminal justice system.

Transparency is an essential value that allows the public to keep systems accountable for upholding their values. Ananny and Crawford describe historical and philosophical interpretations of how transparency is intended to yield accountability (Ananny & Crawford, 2018). Floyd v. New York City serves as a case study for how the public uses transparency to hold both law enforcement agencies and the judiciary accountable. Floyd v. New York City was brought by plaintiffs who claimed that New York City's stop-and-frisk policy was discriminatory (*Floyd v. City of New York*, 2013). By gaining access to forms filled out by officers to justify their searches, the lawyers in the case were able to demonstrate that the searches were often on questionable legal ground (*Floyd v. City of New York*, 2013). In this case, the public was able to leverage transparency to present the judiciary with evidence that it needed to raise its standards for holding law enforcement accountable. Thus, the public is an important part of the network that can hold other actors accountable when given transparent documentation of their actions.

Predictive policing software directly undermines transparency and makes it significantly more difficult for the public to maintain accountability in the criminal justice system. To demonstrate this, we can consider the effect that algorithmic actors might have on the Floyd v. City of New York case study. 55% of forms that officers filled out to justify their reasonable suspicion checked a "High Crime Area" box as part of their reasoning (*Floyd v. City of New York*, 2013). The attorneys in this case found this factor to be "problematic" due to officers interpreting it too broadly (*Floyd v. City of New York*, 2013). However, the debate over this box could become much more of a challenge if the designation were to be informed by predictive policing software. As we established initially, the technical nature of these algorithms does not insulate them from being biased or incorrect. Thus, we return to the problem of how to evaluate them. Ananny and Crawford point out that even if the public had access to source code, this wouldn’t necessarily yield an understanding of the decision process employed by an algorithm, as even software engineers often cannot explain how their own machine learning code works (Ananny & Crawford, 2018). Even if we accept Kirkpatrick’s argument that the primary source of concern is not an algorithm itself, but rather biased datasets, (Kirkpatrick, 2017), there is still an issue of providing transparency into immense, growing bodies of data. Losing this transparency takes power away from the general public in the actor-network of the criminal justice system, as the addition of algorithmic actors that obfuscate law enforcement practices makes the task of challenging these practices more difficult.

**Limitations and Future Work**

This paper establishes how machine learning tools fundamentally undermine the mechanisms that are used to maintain accountability in the criminal justice system. However, this analysis is limited to case law from the United States. This limits how far these results can be generalized, as the criminal justice system in other countries certainly manifests as a different actor-network. While notions of accountability and transparency exist in all countries, the burdens of proof that are analyzed in this paper are defined by the United States Supreme Court. Thus, it is not possible to use this analysis to make an argument about what specific legal conflicts may arise when machine learning is applied to criminal justice systems outside of the United States.

Future research should redress this limitation by analyzing international case law to find common themes on how accountability is achieved in different legal systems. This has the potential to provide a greater body of evidence to support the assertion that machine learning fundamentally conflicts with the way facts are traditionally established and argued in a courtroom. Beyond this, it is necessary to develop a regulatory framework that addresses some of the issues outlined in this paper. This framework must, at a minimum, have a mechanism for evaluating the models used by software. Since accountability is so closely tied to human decision processes, there must be a method of evaluating whether the decisions being made inside an algorithm are sound. It is also necessary to regulate the usage of data sets. There needs to be an understanding of how human biases influence data and what inadvertent consequences this can create when it is used as input to a machine learning algorithm. Finally, there need to be an established set of metrics on which the performance of these algorithms can be evaluated. Without a consensus on the desired impact of using machine learning in the criminal justice system, there is no way to evaluate the outcomes it produces.

## Conclusion

This research has argued that machine learning algorithms are imperfect and capable of perpetuating and possibly exacerbating existing biased police practices. Beyond this, it has been demonstrated that their usage conflicts with legal standards of proof that the judiciary uses to uphold the value of accountability in the actor-network of the criminal justice system. Furthermore, this technology inherently obfuscates decision processes, thereby undermining the transparency that gives the general public the power to hold actors in both law enforcement and the judiciary accountable. The addition of black-boxed algorithmic actors represents a shift of power in the network, taking away the regulatory power of the public. Our existing standard for accountability relies on being able to explain a decision process. There are no existing mechanisms for forcing a machine-learning algorithm to explain itself, thus accountability cannot be readily achieved in a system that uses machine learning to inform its decisions.

Given that these tools are fundamentally in conflict with the way our society maintains accountability, how can they be applied to the criminal justice system, where accountability is essential? This question must be addressed in future research. A regulatory framework must be established as a check against algorithms that obfuscate the pursuit of justice. A faulty criminal justice system can have a catastrophic impact on individual lives. For this reason, there is no way for machine learning to be ethically applied to the criminal justice system without first addressing these concerns.

## Works Cited

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, *20*(3), 973–989. https://doi.org/10.1177/1461444816676645

Bennett Moses, L., & Chan, J. (2018). Algorithmic prediction in policing: assumptions, evaluation, and accountability. *Policing and Society*, *28*(7), 806–822. https://doi.org/10.1080/10439463.2016.1253695

Berk, R. (2017). An impact assessment of machine learning risk forecasts on parole board decisions and recidivism. *Journal of Experimental Criminology*, *13*(2), 193–216. https://doi.org/10.1007/s11292-017-9286-2

Carson, E. A. (2018). Prisoners in 2016 (No. NCJ 251149). Retrieved from Bureau of Justice Statistics website: https://www.bjs.gov/index.cfm?ty=pbdetail&iid=6187

Brinegar v. United States. , 338 US 160 (Supreme Court 1949).

Floyd v. City of New York. , 959 F. Supp. 2d 540 (Dist. Court 2013).

Illinois v. Wardlow. , 528 US 119 (Supreme Court 1999).

Kirkpatrick, K. (2017). It's not the algorithm, it's the data. *Communications of the ACM*, *60*(2), 21–23. https://doi.org/10.1145/3022181

Latour, B. (1996). On actor-network theory: A few clarifications. Soziale Welt, 47(4), 369-381. Retrieved January 31, 2020, from www.jstor.org/stable/40878163

Liptak, A. (2017, December 22). Sent to Prison by a Software Program's Secret Algorithms. The New York Times. Retrieved from https://www.nytimes.com/2017/05/01/us/politics/sent-to-prison-by-a-software- programs-secret-algorithms.html

Lum, K., & Isaac, W. (2016). To predict and serve? *Significance*, *13*(5), 14–19.

https://doi.org/10.1111/j.1740-9713.2016.00960.x

Mohler, G. O., Short, M. B., Malinowski, S., Johnson, M., Tita, G. E., Bertozzi, A. L., & Brantingham, P. J. (2015). Randomized Controlled Field Trials of Predictive Policing. *Journal of the American Statistical Association*, *110*(512), 1399–1411. https://doi.org/10.1080/01621459.2015.1077710

Terry v. Ohio. , 392 US 1 (Supreme Court 1967).

Whittle, A., & Spicer, A. (2008). Is Actor Network Theory Critique? Organization Studies, 29(4), 611–629. https://doi.org/10.1177/0170840607082223