Artificial Intelligence and Law (2022) 30:1–17
https://doi.org/10.1007/s10506-021-09286-4

REVIEW ARTICLE

A review of predictive policing from the perspective of fairness

Kiana Alikhademi[1] · Emma Drobina[1] · Diandra Prioleau[1] · Brianna Richardson[1] · Duncan Purves[2] · Juan E. Gilbert[1]

Accepted: 31 March 2021 / Published online: 15 April 2021

Abstract

Machine Learning has become a popular tool in a variety of applications in criminal justice, including sentencing and policing. Media has brought attention to the possibility of predictive policing systems causing disparate impacts and exacerbating social injustices. However, there is little academic research on the importance of fairness in machine learning applications in policing. Although prior research has shown that machine learning models can handle some tasks efficiently, they are susceptible to replicating systemic bias of previous human decision-makers. While there is much research on fair machine learning in general, there is a need to investigate fair machine learning techniques as they pertain to the predictive policing. Therefore, we evaluate the existing publications in the field of fairness in machine learning and predictive policing to arrive at a set of standards for fair predictive policing. We also review the evaluations of ML applications in the area of criminal justice and potential techniques to improve these technologies going forward. We urge that the growing literature on fairness in ML be brought into conversation with the legal and social science concerns being raised about predictive policing. Lastly, in any area, including predictive policing, the pros and cons of the technology need to be evaluated holistically to determine whether and how the technology should be used in policing.

Keywords Fairness · Algorithmic fairness · Predictive policing · AI in criminal justice

✉ Kiana Alikhademi
kalikhademi@ufl.edu

Extended author information available on the last page of the article

1 Introduction

Algorithms are used to automate tasks in a multitude of areas, such as health, finance, and law enforcement. Machine Learning (ML) is a class of algorithms that produces outcomes based on patterns found in data. It can be used for tasks ranging from cancer detection to predicting the likelihood that a parolee will re-offend. ML algorithms are gaining in popularity for tasks that can be time-consuming and cumbersome for humans to do. These algorithms produce results in less time and can overcome and mitigate human prejudices. However, ML methods can reflect and entrench the biases of humans. This threat of bias has produced a new arena of research focused on maintaining fairness within algorithms, often referred to as algorithmic fairness. Mehrabi et al. (2019) defines algorithmic fairness as "absence of any prejudice or favoritism toward an individual or a group based on their inherent or acquired characteristics." Fairness is of concern in algorithmically assisted policing. ML algorithms in law enforcement have the potential to predict future crime based on historical crime data (Perry 2013). Perry (2013) defines predictive policing as the use of quantitative and statistical methods to forecast the individuals who may commit a crime and target areas where crime is likely to occur.

A growing chorus of academics have warned that predictive policing algorithms are susceptible to influence from fallible human decision-making, which can have enormous social consequences. For example, as Lum and Isaac (2016a) have described, communities with higher police presence will naturally have higher arrest rates. This leads to the creation of datasets that appear to reflect higher crime rates, but which really reflect greater police attention. This can be to the disadvantage of communities already burdened by the costs of over-policing. When historical racism and class discrimination are encoded in the outputs of an algorithm, minority and low-income communities might fall victim to a feedback loop of ever greater police attention. It is vital to study fairness in the predictive policing domain to ensure that racial and economically disparate impacts are not perpetuated in new and emerging technologies in this field. Selbst (2017) has described in detail some of the potential sources of bias and threat of disparate impact from the use of predictive policing.

While there is a growing body of research on fairness guidelines in ML (Martinez et al. 2019; Lohia et al. 2019; Calmon et al. 2017; Berk et al. 2018), this body of research has not been brought into conversation with emerging technologies in policing. There thus remains a need for a comprehensive literature review of fairness in ML as it pertains to emerging ML applications in policing. This requires paper offers a first step to achieving this goal by gathering relevant works and identifying the major themes to help researchers conceptualize areas that need improvement and further evaluation. There is a necessity to evaluate the predictive policing from the perspective of fairness This makes evaluations of predictive policing, as it is used by police, one of the most critical areas of need in the ML fairness literature. Conducting evaluations is complicated by the lack of clear standards for fairness techniques, which can and do conflict, as well as intellectual property protections of the source code. Ultimately, ML as a field must critically engage with policing itself.

With this motivation, the present paper reviews recent publications in the field of fairness in machine learning, predictive policing, and criminal justice as follows:

- Section 2 investigates the emerging definitions of fairness and the fairness tools available to ML practitioners
- Section 3 reviews the major expectations about predictive policing from different perspectives such as biased data, accountability and transparency, policy and civil liberties
- Section 4 reviews the literature from criminal justice to provide a fuller picture of existing obstacles and problems in this domain
- Lastly, Sect. 5 examines the existing methods to improve the fairness in predictive policing with the goal to shed more light on the current research in this field

2 General perspectives on fairness

2.1 Fairness definitions

To begin, it will be helpful to understand the growing landscape of fairness in the ML literature. Below we described the dominant fairness definitions that have been suggested by ML researchers (Kusner et al. 2017; Dwork et al. 2012; Corbett-Davies et al. 2017; Hardt et al. 2016; Grgic-Hlaca et al. 2016; Mehrabi et al. 2019; Verma and Rubin 2018). Fairness in the ML literature is defined in terms of predicted outcome, actual outcome, or similarity measures (Verma and Rubin 2018). According to Bellamy et al. (2018), the protected attribute(e.g., race, sex) divides the population into different groups of privileged and disadvantaged individuals. The privileged group has been at a systematic advantage due to the values of protected attributes.

Classification parity, also known as statistical parity, counts a predictor as fair if it is equally likely to generate a positive classification for members of the privileged groups as it is for members of the disadvantaged groups (Kusner et al. 2017; Dwork et al. 2012). *Conditional statistical parity* proposes that a predictor needs to provide a positive outcome with the same likelihood for both privileged and disadvantaged groups over a specific set of factors (Corbett-Davies et al. 2017).

Calibration (Chouldechova 2017) considers a predictor fair if, for any predicted probability (P), subjects in both privileged and disadvantaged groups have the same likelihood of positive classification. The main difference between calibration and statistical parity is that calibration considers the fraction of positive predictions over the number of all predictions for any probability.

Hardt et al. (2016) propose *equalized odds* and *equal opportunity* to define fairness for binary predictors. According to the *equal opportunity* definition, a fair classifier predicts positive outcomes for members of the positive class (e.g. the group of people on parole who will reoffend) in both privileged and disadvantaged groups with the same likelihood. *Equalized odds* requires that a fair classifier predicts positive outcomes for members in the positive class and negative class with the same likelihood for both privileged and disadvantaged groups.

These two metrics assess the true positive, which is the number of instances correctly predicted as positive, and the false positive, which is the number of instances incorrectly predicted as positive, remains the same across the privileged and disadvantaged groups.

Fairness through awareness considers an algorithm fair if similar individuals, based on some defined metrics, are given similar outcomes (Dwork et al. 2012). On the other hand, *fairness through unawareness*, also known as Anti-classification (Corbett-Davies and Goel 2018), defines fairness as not using protected attributes in the decision-making process in order to avoid any unintentional consequences (Kusner et al. 2017; Grgic-Hlaca et al. 2016).

Lastly, *counterfactual fairness* considers a predictor fair if it treats the individual the same way regardless of which group the individual belongs to in the real world (Kusner et al. 2017).

Fairness definitions fall into two different categories of individual or group fairness. Individual fairness metrics such as *fairness through awareness*, *fairness through unawareness*, *counterfactual fairness* focus on similar outcomes for individuals in the same group(Kusner et al. 2017; Grgic-Hlaca et al. 2016; Dwork et al. 2012; Mehrabi et al. 2019). In contrast, group fairness metrics such as *demographic parity*, *conditional statistical parity*, *equalized odds*, *equal opportunity*, and *calibration* emphasize treating different groups in similar ways (Kusner et al. 2017; Dwork et al. 2012; Corbett-Davies et al. 2017; Hardt et al. 2016; Mehrabi et al. 2019).

Other domains such as economics inspired scholars to develop fairness frameworks for predictive analytic purposes. Heidari et al. (2019) developed a fairness framework using equality of opportunity measure, typically found in economic models. This equal opportunity framework recognizes that individuals' outcomes are determined by a combination of circumstances outside of their control and their own efforts, which might be within their control. This framework emphasizes the importance of personal qualifications and seeks to minimize the impact of circumstances beyond a person's control on individual outcomes. They attempt to show that most of the fairness notions appearing in ML research are attempts to operationalize fairness as equal opportunity.

2.2 Fairness limitations

While algorithms help to automate the decision-making process, they are imperfect tools (Persson and Kavathatzopoulos 2018b). According to Persson and Kavathatzopoulos (2018b), the main limitations of predictive analytics and algorithms are: (1) The use of probabilistic estimations on humans, (2) A singular focus on recognizing the patterns, instead of understanding the underlying cause of the patterns, (3) and the lack of apprehension and judgment. Unfairness and discrimination caused by these algorithms could be rooted in the algorithms' inability to diagnose underlying causes or to exercise judgement. These deficiencies, if unchecked, can exacerbate unfairness and discrimination. The following solutions were proposed to remedy the existing problems regarding explainability, accountability, and transparency:

- Better technological methods to reveal possible actions and provide more insight into the process of decision-making (Explainability)
- Proper laws and regulations to limit the types of activities that can be automated using predictive analytics (Accountability)
- Designing human-in-the-loop systems that require active oversight and interaction for algorihtms used in decision-making (Transparency)

While the fairness metrics discussed previously have the potential to measure fairness and help in identifying the biases, they have some limitations, too. Corbett-Davies and Goel (2018) discussed some of the limitations with using anti-classification, classification parity, and calibration. One limitation these measures share is that they do not treat individuals with the same risk level equally. The limitations for each method are as follows:

- Classification parity: heavily depends on the risk distributions across different groups and has difficulty handling marginal cases. Also, it is ignorant of any correlation between the positive outcome and protected attributes (Shrestha and Yang 2019).
- Anti-classification: unable to control the proxy attributes existing in the data source
- Calibration: unable to ensure that decision outcomes are equal across protected attributes or that risk scores are accurately computed

Corbett-Davies and Goel (2018) observe that factors such as measurement error, sample bias, model explainability, and equilibrium effect make it challenging to design a fair classifier which satisfies the previously mentioned fairness metrics.

Binns (2018) discusses through the lens of moral and political philosophy the definitions of "fair" and "non-discriminatory" as they apply to algorithmic decision-making systems.

Binns (2018, pp.77) proposes that we understand fairness in terms of "spheres of justice," where each sphere calls for equality in one specific domain such as human resources or education. By understanding fairness in terms of different domains, we can shed light on the most appropriate methods for improving fairness in a context. Verma and Rubin (2018) review the existing notions of fairness and provide one case study for each one. Their work shows how one result might be fair according to specific fairness metric and unfair according to other metrics. The paper emphasizes that indications of fairness are dependent on various factors such as data, classifiers, and outcome space.

2.3 Fairness toolkits

Fairness definitions and frameworks are explained in the previous section. However, there are many researchers in this field without any technical background in Machine Learning. Significant research conducted to remove technical knowledge barriers. These efforts have produced various fairness "toolkits" (Bird et al. 2020;

Bellamy et al. 2018; Friedler et al. 2019; Saleiro et al. 2018; Wexler et al. 2019). Bird et al. (2020) proposed the Fairlearn toolkit that provides an interactive experience where users can evaluate, compare, and mitigate biases in their models using a consortium of mitigation algorithms. Bellamy et al. (2018) introduced the Fairness AI 360 toolkit that allows practitioners to evaluate, detect, and mitigate bias in their datasets and models, while also providing a plethora of support for learning about fairness in AI. Friedler et al. (2019) created a python package that compares models according to a substantial collection of fairness metrics. Saleiro et al. (2018) created the multi-platformed Aequitas tool that includes a command line interface, a web application, and a python package and provides a user-friendly bias report that allows ML and fairness novice and experts alike to gain valuable insight. Wexler et al. (2019)'s fairness evaluation and visualization toolkit provides a multi-functional, interactive experience where users can visualize bias reports and experiment with manipulating fairness constraints and thresholds.

3 Expectations for predictive policing

A burgeoning literature outlines reasonable expectations for the development and implementation of predictive policing technologies. Such work focuses on themes of fairness, privacy, and accountability. Take as a whole, this growing body of literature suggests a code of conduct for the successful implementation of predictive policing. This section surveys the key themes from that body of literature.

3.1 Biased data and fairness

One of the most frequent criticisms of predictive policing centers around the issue of bias. Several works (Richardson et al. 2019; Joh 2017; Degeling and Berendt 2018; Perry et al. 2018; Scantamburlo et al. 2018; Xiang and Raji 2019; Persson and Kavathatzopoulos 2018b; Abdollahi and Nasraoui 2018; Lum and Isaac 2016b; Vestby and Vestby 2019; Reisman et al. 2018; Ferguson 2016; Bakke 2018; Selbst 2017) confront predictive policing with similar complaints about biased data, confirmation biases in development, systematic biases, inductive biases, and institutional biases. Lum and Isaac (2016b) discuss the deeply institutional and historical issues with police data that stem from the biases held by officers. Arrest statistics can be impacted by policing decisions about where to patrol and about the individuals they decide to detain or search. Furthermore, past research has shown that these decisions are highly motivated by race and ethnicity (Lum and Isaac 2016b). Lum and Isaac (2016b) argue that using such data in sophisticated software under the guise of 'fair' and 'bias-free' algorithms legitimizes biased police practices.

Richardson et al. (2019) studied 13 counties that have employed predictive policing technologies. The results of their work strongly suggest that utilizing predictive policing in 'broken' or highly corrupt police departments exacerbates corruption (Richardson et al. 2019).

Furthermore, there is scrutiny around the optimization models used for predictive policing. Degeling and Berendt (2018) discuss inductive biases that exist in tools like PredPol, where the system utilizes the hypothesis that minimizes the number of assumptions for accurate predictions. In such a system, the algorithm would not be able to make accurate predictions if given a test set with instances unlike those used to train the model. Furthermore, there exists maximum conditional independence bias, which assumes that factors work independently instead of contributing to each other (Degeling and Berendt 2018). Vestby and Vestby (2019) also discuss the learning process in the machine learning model and request that non-ML experts be critical of the learning goals, measurements, and optimization decisions.

Perry et al. (2018); Persson and Kavathatzopoulos (2018a) also state concerns about historical bias that can occur from ignoring the fact that as society changes, the motivations, means, and perpetrators change as well. Inevitably the people who commit crimes, the types of crimes that are committed, and the location of criminal activity will change. It is critical that predictive technology is sensitive and responsive to these changing societal conditions. If not, these historical biases will create a feedback loop.

Substantial issues of bias and fairness arise for socially and economically disadvantaged populations. Joh (2017) describes some reasonable expectations for addressing bias and fairness concerns when implementing predictive policing systems:

> Police agencies, communities, and local governments should ask: how can these Al systems address the potential of reproducing and amplifying bias? This should involve not only testing of an Al system before release but also continuous monitoring. Will the company providing the system permit access to researchers to ensure that rigorous and open monitoring will be possible? Will results of findings be provided to those communities that have historically experienced biased policing? (see Joh 2017, p. 1142)

3.2 Privacy and civil liberties

With the era of 'Big Data,' a new wave of technology and, subsequently, demands for political engagement concerning that technology have quickly risen. While many companies buy or utilize user data with little backlash, the use of such data by police departments for public safety induces skepticism. Concerns of fairness are directly linked to the protection of privacy and civil liberties, especially for communities with higher police presence Xiang and Raji (2019).

Degeling and Berendt (2018) discuss several predictive policing practices that give rise to civil liberties concerns, including data collection through surveillance, questioning of people and neighborhoods, and intrusion into private physical or virtual spaces. As was revealed by Edward Snowden's leaked internal NSA documents, surveillance was conducted on a multitude of people, including those not under suspicion, to collect enough data to train a classifier. Degeling and Berendt (2018) argue that transparent data collection must be implemented to ensure citizens that similar privacy invasions are not used for predictive policing (Degeling and Berendt

2018). Robertson et al. (2020) have recently provided an in-depth critical evaluation of a number of data-driven policing systems in use in Canadian law enforcement in terms of international human rights law. They identify threats to privacy, freedom of expression, assembly, and association, freedom from discrimination, arbitrary detention, and due process (Robertson et al. 2020).

The concern about civil liberties and data gathering comes from the fact that the legal framework for personal data collection is ill-defined (Perry et al. 2018). This problem is two-fold: (1) lawmakers and political scientists are not active participants in the development life cycle of these technologies (Persson and Kavathatzopoulos 2018b; Perry et al. 2018; Xiang and Raji 2019) and (2) laws have not been updated to cover current technologies (Perry et al. 2018). Xiang and Raji (2019) discusses the huge misalignment that exists between computer science and law as it concerns fairness in AI. The writers suggest that it is imperative that collaboration occur between the legal and technological side to effectively make policy for this technology. Perry et al. (2018)'s report on NIJ funded projects for assessing predictive policing includes details of a Predictive Policing Symposium that was held in Los Angeles (LA) in November of 2009. According to this report, many participants in the symposium felt as though privacy and civil liberties were critical to the future of predictive policing and felt as through privacy advocates should be involved in the development of these technologies (Perry et al. 2018).

The concluding remarks by Perrot (2017) reflect the major expectations held by the public for the future of predictive policing: it is critical that the potential of AI in this domain be limited by a respect for privacy.

3.3 Accountability and transparency

Transparency is a further concern about predictive policing due to the frequently proprietary software being used.A requirement of transparency can promote fairness and accountability by forcing organizations to remain cognizant of the functionality and impact of their tools, and by supporting citizens attempts to understand and question tools and their impact on the community (Scantamburlo et al. 2018).

Many scholars also emphasize the importance of transparency when building trust between the police and their respective communities (Reisman et al. 2018; Abdollahi and Nasraoui 2018; Perry et al. 2018; Asaro 2019; Scantamburlo et al. 2018; Joh 2017; Ferguson 2016; Bakke 2018; Persson and Kavathatzopoulos 2018a).

Participants at the Predictive Policing Symposium, introduced in the previous section, insisted that transparency was critical to establishing community trust (Perry et al. 2018). Since trust is a key component for the success of these systems, Scantamburlo et al. (2018) propose four benchmarks for assessing their predictive policing technology, one of which includes transparency and accountability.

Furthermore, it is critical that there be predefined roles of accountability (Joh 2017; Persson and Kavathatzopoulos 2018a; Bennett Moses and Chan 2018). Bennett Moses and Chan (2018) emphasizes that police must continue to be held accountable for their decisions, even when influenced by predictive software.

Furthermore, it must be clear who is responsible when it comes to system failures and misuse of such technology, whether it be the police department, the auditors, or the developers of the technology.

3.4 Development, implementation, and assessment of policing AI

Many scholars in this area discuss methods for developing, implementing, and assessing predictive policing technologies.

Perry et al. (2018) stated that the development of predictive policing technology should focus on tactical utility instead of high accuracy, where algorithms consider the officers available, the landscape of the suspected crime, and the means for de-escalating or preventing a crime. Furthermore, Persson and Kavathatzopoulos (2018a) and Ridgeway (2013) state the importance of ensuring that these tools are supportive of decision-making and do not supersede the judgement of the user.

Additionally, when it comes to implementing this technology, much needs to be done to ensure the product is being used correctly. It is critical that users understand what the system can and cannot do and that they are properly trained to use the system as a supplement to their responsibilities (Asaro 2019; Bennett Moses and Chan 2018; Ridgeway 2013). As Santos (2019) points out, officers are in many cases given little or no instruction about what to do when they arrive at a designated high-risk area. "What is an officer to do in a 500-by-500-foot area, especially when there is no actionable intelligence[?]" (Santos 2019, pp. 384).

Perry et al. (2018) observes a lack of emphasis on assessing and evaluating predictive policing technologies by developers and police departments. Degeling and Berendt (2018) discuss a three-part test that predictive policing technology should satisfy. This includes:

(1) a suitability test to evaluate effectiveness of technology; (2) a necessity test to determine if there are less intrusive means; and (3) the proportionality test that measures the balance of interests where benefits are limited to the scope of police responsibility. Once a technology is in use, it is critical that frequent assessment and correction be done (Reisman et al. 2018). Assessments should not only evaluate the software design, but how it is being used and the impact it is having on its community (Reisman et al. 2018). Furthermore, it is critical that individuals that perform the assessments are not limited by claims of trade secrecy (Reisman et al. 2018). Ferguson (2016) criticizes the current system in which policing technology is invented, adopted, and then only later assessed. He urges that assessment be involved at the beginning of the process (Ferguson 2016). Furthermore, Ferguson (2016) emphasizes that if the vulnerabilities of a system cannot be effectively managed by the jurisdiction opting to deploy them, the jurisdiction will not be able to do so responsibly and shall remain open to criticism and challenges (Ferguson 2016).

Remarks by Richardson et al. (2019) about the expectations for predictive policing emphasize how important it is that the public is able to know, assess and reject such systems. If residents are unsatisfied with the results of assessments, if they feel like their civil rights are being encroached upon, or if they feel unsafe or threatened

by this technology, it is critical that police, as public servants, always consider adopting alternatives to the technology.

3.5 Other approaches to predictive policing

Some scholars proposed different approaches to predictive policing that did not involve predicting crime or assessing individuals, but instead involve removing motivations for crime. Ferguson (2016) refers to such technology as Predictive Policing 3.0. This approach acknowledges that certain individuals face challenges that increase their propensity for violence and it recommends generating a public health model for identifying these people and their needs.

Nissan (2017) discusses the capabilities for AI to assist police in understanding the context in which they work. Asaro (2019) proposes an ethical framework for considering and adopting predictive policing AI called the 'AI Ethics of Care' approach. This approach is modeled from the police tenants of "Duty to Protect" and "Duty to Care," and the guiding aim of which is to benefit everyone who participates in the system (Asaro 2019). This approach requires training, guidance, and direction in using the system so that users understand the capacities of the system to promote this guiding aim. It also requires that domain experts be involved in the design of the system so that there is plenty of prototyping and testing, and so that the algorithm, data, and practices are transparent (Asaro 2019).

4 Evaluations of predictive policing systems

A vital component in the use of intelligent policing systems is system evaluation. However, in spite of the burgeoning literature on fairness, transparency, and civil liberties concerns related to predictive policing, there are few published evaluations of machine learning systems in connection with criminology and policing. This is possibly due to the lack of access to predictive policing systems by independent researchers, which was discussed in Sect. 3.3. Campedelli (2019) conducted a systematic literature review on the intersection of AI and crimes and found that most research in this area has been concentrated to cyber-related crimes. In addition, there is a lack of research focusing on algorithmic discrimination, bias, and ethics within this domain (Campedelli 2019). Santos (2019) confirms the absence of peer reviewed studies of predictive policing systems, finding only one accuracy study published in a peer reviewed journal (Santos 2019) (Mohler et al. 2015).

In line with these earlier findings, we found only a few papers that evaluated the accuracy or fairness of outcomes produced by predictive policing systems.

4.1 Predictive policing systems

Marda and Narayan (2020) discussed the use of a predictive policing system called COMAPS (Crime Mapping Analytics and Predictive System) within the capital of India, Delhi, which used AI capabilities for spatial hot-spot mapping of potential

crime areas, criminal behavior patterns, and suspect analysis. They evaluated the effects of this system and highlighted the biases within the data collection process. Their evaluation was conducted via interviews and observations within limited divisions at the Delhi Police Headquarters. It was found that initial data collection and creation lacked standard operating procedures and auditing mechanisms. In addition, they discussed three areas of bias found within the use of this system: (1) historical bias, (2) representation bias, and (3) measurement bias. It was discovered, although unsurprisingly, that individuals of higher socioeconomic status were underrepresented in the data. This can cause the system to inherently overlook crimes that do occur within this demographic group and may lead to over-policing in other areas that historically are already targeted or have negative interactions with the police. Another interesting finding was measurement bias, in which people who called the dispatcher to report a crime may have been unable to provide their address due to living in temporary settlements or being isolated to a small radius within their community. This isolation was typically associated with the individual being a housewife and not having the need to know or remember their home address. This could also be a similar issue for those who are homeless and therefore may have difficulties reporting where a crime took place. As a part of measurement bias, it was discovered by the authors that how a crime was categorized was reliant on the interpretation of the officer or call taker. It then is easy to see how biases can be aggregated into the system and can ultimately affect marginalized and vulnerable populations. The authors also described a lack of transparency in the creation of the system and the decisions that were being made.

Mohler et al. (2015) and Brantingham et al. (2018) are the two peer reviewed studies of the PredPol predictive policing system, which was until spring 2020 used by the Los Angeles Police Department in allocating police resources to deter vehicular theft and theft from a vehicle. The 2015 study assessed PredPol's accuracy in comparison with human crime analysts, finding that it was as much as twice as accurate as the control method at predicting the location and timing of crime. The 2018 paper reports the results of a randomized controlled trial of the predictive policing software PredPol in order to test the claim of some critics that predictive policing will lead to racially biased arrests. Brantingham et al. found that PredPol did not lead to racially biased arrests when compared with the control method of allocation. The results of this trial, of course, do not show that PredPol improves on any bias in the control method, though it does provide some evidence that PredPol does not exacerbate those biases.

5 Techniques for improving predictive policing

In response to critical evaluations of predictive policing, AI researchers have developed a number of techniques for data manipulation and machine learning. These techniques can be grouped based on the point in the development process they are implemented: (1) pre-processing, (2) algorithm design, and (3) post-processing. While these techniques do not eliminate the need for holistic evaluations of predictive policing tools or social programs to address the root of criminal behavior, they

can alleviate problems that arise from poor data preparation or thoughtless model construction.

5.1 Pre-processing

Data pre-processing is the process of transforming "raw" data into a usable form. Calmon et al. (2017) determined a way to transform both training and test datasets to prevent discrimination by reducing the output's dependence on variables that have been identified as discriminatory (e.g. race, sex, etc.) while simultaneously ensuring that the resulting output is not too different from the original dataset. This process requires subject matter experts and stakeholders to determine thresholds for the constraints and the transformations, which will lead to problems if developers do not consult experts. Furthermore, few applications explicitly rely on data about sex, race, or other protected attributes, even when they produce outcomes that, in practice, discriminate against members of a protected group. Calmon's method therefore has limited applicability to many machine learning applications.

5.2 Algorithm design

Methods to improve criminal justice algorithms from academia can be split into two main groups: qualitative ways to think about algorithm design in advance of bias being detected and programmatic interventions in the algorithm to detect and correct bias. From the first category, Altman et al. (2018) proposed a method of counterfactual analysis where designers would identify a major choice in algorithm design that had the potential to substantially affect the well-being of the targets (e.g. what would the results look like if race was excluded from the training data?; what would they look like if race was exclude from the information of an individual input?).

In the second category, Ensign et al. (2017) investigated modifications of the input to predictive policing algorithm PredPol, used until recently by the LAPD, to avoid feedback loops. Predictive policing algorithms can be designed to learn from new crime data gathered by officers on patrol so police can adapt patrolling decisions to trends in crime. However, by feeding the crime data gathered by police back into the algorithm, the algorithm might send police back to the site they just patrolled, since it now has more data showing that as a high-crime area. This makes for a vicious cycle: the more police are sent to an area, the more crime they can see. So, more police officers are sent to the area, where they discover more crime. Ensign's proposed solution to this problem alters how data is added so that the more likely it is that police are sent to a given district, the less likely it is that we should incorporate those discovered incidents. Alarmingly, without this addition, Ensign reports that PredPol's predictions do not converge to the true crime rate.

Benthall and Haynes (2019) argue that protected categories are so embedded in a societal and structural system of discrimination that treating them as simple personal identifiers will not help. Instead, they advocate for learning proxy features from the training data that correlate with minority status and using them as input to machine learning fairness interventions in the algorithm. This, they argue, will help

identify the root causes of racial injustice without replicating racial categorization wholesale.

5.3 Post-processing

Lohia et al. (2019) propose a method for post-processing the results of an algorithm to make them respect both group and individual fairness. Results are split into two groups: privileged and unprivileged. All samples from the unprivileged group are tested for individual bias based on a method the authors developed in a previous paper; if they score highly, they are assigned to the outcome they would have received if they were part of the privileged group. All other samples are left unchanged. The authors found their method consistently reduced bias and disparate impact while maintaining accuracy, but it is a simple algorithm that only seems to work for binary categories.

5.4 Analyzing results

Analyzing the output of an ML algorithm is how we answer the question we are most interested in: is a given algorithm biased? This is not an easy question to answer—different definitions of fairness and bias conflict, and as Chouldechova (2016) observed in her 2016 paper, not all statistical fairness criteria can be simultaneously satisfied. Despite this, and as we saw in Sect. 2, researchers have proposed a wide range of statistical measures to evaluate the fairness of outcomes for groups. Other measures include the examination by Khademi and Honavar (2019) of the causal effect of the protected attribute on the outcome using a measure called Fair on Average Causal Effect on the Treated (FACT). Speicher et al. (2018) adapt existing measures of economic inequality to quantify the unfairness of an algorithm. Speicher's measures allow results to be decomposed into between-group unfairness and within-group unfairness for any set of non- overlapping groups, which shows that a reduction in between-group unfairness can lead to an increase in within-group unfairness. Wang et al. (2019) proposed avoiding explicit fairness metrics entirely and training a second classifier on human judgements, so judgements made by AI can be classified as similar or not similar to human judgement. Ultimately, as Chouldechova (2016) reminds us, there is no simple solution; fairness and disparate impact are social and ethical concepts, not statistical ones, but human-made decisions may be just as biased as AI-made ones.

6 Conclusion and future works

The growing prominence of predictive policing has led to increased interest from researchers and lawmakers. However, research access to active predictive policing systems is limited due to their proprietary nature. Lack of access to code has a ripple effect throughout the body of literature: evaluations of criminology software (like ProPublica's COMPAS analysis) must first simulate the algorithms before they

can evaluate them; researchers designing fairness metrics have extremely limited sources for data to rely on to judge the effectiveness of the tools they design; and ethicists and legal scholars who want to address the moral and legal ramifications of predictive policing have difficulty finding clear information on the capabilities and implementation of criminology software. This makes evaluations of predictive policing, as it is used by police, one of the most critical areas of need in the ML fairness literature. Conducting evaluations is complicated by the lack of clear standards for fairness techniques, which can and do conflict, as well as intellectual property protections of the source code. Ultimately, ML as a field must critically engage with policing itself.

Now more than ever, it is critical that predictive policing be developed and implemented in ways that are sensitive to its social impacts. One may also question whether ML and AI should be used in socially sensitive areas, such as policing, due to the negative consequences that can arise from inaccuracy or bias. For instance, being arrested - even if charges are dropped - can cause an individual to lose their job, be evicted, and struggle with background checks throughout their life. Moreover, police encounters can be life-threatening, particularly for Black and Indigenous individuals. This is what we must acknowledge as we see predictive policing and other criminology data tools proliferate, recidivism prediction and police facial recognition being among them. As of 2013, over 150 police departments used some form of predictive policing tools (Bond-Graham and Winston 2013); as of 2016, approximately one in four departments had access to police facial recognition systems (Garvie 2016). Not only can skewed data lead to police over patrolling in poor and minority neighborhoods (Ensign et al. 2017; Marda and Narayan 2020), but any corruption in police departments can also lead to predictive policing being used as a tool of abuse (Richardson et al. 2019). Therefore, current policing systems should be evaluated for their potential to address and overcome systemic racism, discrimination, and prejudice within policing alongside of their potential to automate the processes of law enforcement. Technology is only as objective and fair as the practice it is being used to automate. And city officials must remain open to the possibility that some social issues currently addressed by law enforcement would be better handled through reforms to public education, mental health, and employment. Therefore, as technological systems are designed to help police address certain social issues, we must also acknowledge that technology may not be the most plausible avenue through which to address these issues insofar as policing may not be the best method for addressing them in first place.

Funding This material is based upon work supported by the National Science Foundation under Grant No. 1917712.

References

Abdollahi B, Nasraoui O (2018) Transparency in fair machine learning: the case of explainable recommender systems. In: Human and Machine Learning. Springer, pp 21–35

Altman M, Wood A, Vayena E (2018) A harm-reduction framework for algorithmic fairness. IEEE Secur Privacy 16(3):34–45

Asaro PM (2019) Ai ethics in predictive policing: from models of threat to an ethics of care. IEEE Technol Soc Mag 38(2):40–53
Bakke E (2018) Predictive policing: the argument for public transparency. NYU Ann Surv Am L 74:131
Bellamy RK, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, Lohia P, Martino J, Mehta S, Mojsilovic A, et al. (2018) Ai fairness 360: an extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv:181001943
Bennett Moses L, Chan J (2018) Algorithmic prediction in policing: assumptions, evaluation, and accountability. Policing Soc 28(7):806–822. https://doi.org/10.1080/10439463.2016.1253695
Benthall S, Haynes BD (2019) Racial categories in machine learning. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, pp 289–298
Berk R, Heidari H, Jabbari S, Kearns M, Roth A (2018) Fairness in criminal justice risk assessments: the state of the art. Sociol Methods Res 0049124118782533
Binns R (2018) What can political philosophy teach us about algorithmic fairness? IEEE Secur Privacy 16(3):73–80
Bird S, Dudík M, Edgar R, Horn B, Lutz R, Milan V, Sameki M, Wallach H, Walker K (2020) Fairlearn: a toolkit for assessing and improving fairness in ai. Tech. Rep. MSR-TR-2020-32, Microsoft, https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/
Bond-Graham D, Winston A (2013) All tomorrow's crimes: the future of policing looks a lot like good branding. SF Weekly News https://archives.sfweekly.com/sanfrancisco/all-tomorrows-crimes-the-future-of-policing-looks-a-lot-like-good-branding/Content?oid=2827968&showFullText=true
Brantingham PJ, Valasik M, Mohler GO (2018) Does predictive policing lead to biased arrests? Results from a randomized controlled trial. Stat Public Policy 5(1):1–6
Calmon FP, Wei D, Ramamurthy KN, Varshney KR (2017) Optimized data pre-processing for discrimination prevention. arXiv:170403354
Campedelli GM (2019) Where are we? Using scopus to map the literature at the intersection between artificial intelligence and crime. arXiv:1912.11084
Chouldechova A (2016) Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. arXiv:1610.07524
Chouldechova A (2017) Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. Big Data 5(2):153–163
Corbett-Davies S, Goel S (2018) The measure and mismeasure of fairness: a critical review of fair machine learning. arXiv:180800023
Corbett-Davies S, Pierson E, Feller A, Goel S, Huq A (2017) 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
Degeling M, Berendt B (2018) What is wrong about robocops as consultants? A technology-centric critique of predictive policing. AI & Soc 33(3):347–356
Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R (2012) Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference, pp 214–226
Ensign D, Friedler SA, Neville S, Scheidegger C, Venkatasubramanian S (2017) Runaway feedback loops in predictive policing. arXiv:1706.09847
Ferguson AG (2016) Policing predictive policing. Wash UL Rev 94:1109
Friedler SA, Scheidegger C, Venkatasubramanian S, Choudhary S, Hamilton EP, Roth D (2019) A comparative study of fairness-enhancing interventions in machine learning. In: Proceedings of the conference on fairness, accountability, and transparency, pp 329–338
Garvie C (2016) The perpetual line-up: unregulated police face recognition in America. Georgetown Law, Center on Privacy & Technology
Grgic-Hlaca N, Zafar MB, Gummadi KP, Weller A (2016) The case for process fairness in learning: feature selection for fair decision making. In: NIPS symposium on machine learning and the law, vol 1, p 2
Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. In: Advances in neural information processing systems, pp 3315–3323
Heidari H, Loi M, Gummadi KP, Krause A (2019) A moral framework for understanding fair ml through economic models of equality of opportunity. In: Proceedings of the conference on fairness, accountability, and transparency, pp 181–190
Joh EE (2017) Artificial intelligence and policing: first questions. Seattle UL Rev 41:1139

Khademi A, Honavar V (2019) Algorithmic bias in recidivism prediction: a causal perspective. arXiv:1911.10640

Kusner MJ, Loftus J, Russell C, Silva R (2017) Counterfactual fairness. In: Advances in neural information processing systems, pp 4066–4076

Lohia PK, Ramamurthy KN, Bhide M, Saha D, Varshney KR, Puri R (2019) Bias mitigation post-processing for individual and group fairness. In: Icassp 2019–2019 ieee international conference on acoustics, speech and signal processing (icassp), IEEE, pp 2847–2851

Lum K, Isaac W (2016a) Predictive policing reinforces police bias. Human Rights Data Anal Group

Lum K, Isaac W (2016b) To predict and serve? Significance 13(5):14–19

Marda V, Narayan S (2020) Data in new delhi's predictive policing system. In: Proceedings of the 2020 conference on fairness, accountability, and transparency, pp 317–324

Martinez N, Bertran M, Sapiro G (2019) Fairness with minimal harm: a pareto-optimal approach for healthcare. arXiv:191106935

Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2019) A survey on bias and fairness in machine learning. arXiv:190809635

Mohler GO, Short MB, Malinowski S, Johnson M, Tita GE, Bertozzi AL, Brantingham PJ (2015) Randomized controlled field trials of predictive policing. J Am Stat Assoc 110(512):1399–1411

Nissan E (2017) Digital technologies and artificial intelligence's present and foreseeable impact on lawyering, judging, policing and law enforcement. Ai & Soc 32(3):441–464

Perrot P (2017) What about ai in criminal intelligence? From predictive policing to ai perspectives. Eur Law Enforc Res Bull 16:65–75

Perry W, McInnis B, Price C, Smith S, Hollywood J (2018) Predictive Policing: the role of crime forecasting in law enforcement operations. RAND Corporation, Tech. rep

Perry WL (2013) Predictive policing: the role of crime forecasting in law enforcement operations. Rand Corporation, Santa Monica

Persson A, Kavathatzopoulos I (2018a) How to make decisions with algorithms. ACM SIGCAS Comput Soc 47(4):122–133

Persson A, Kavathatzopoulos I (2018b) How to make decisions with algorithms: ethical decision-making using algorithms within predictive analytics. ACM SIGCAS Comput Soc 47(4):122–133

Reisman D, Schultz J, Crawford K, Whittaker M (2018) Algorithmic impact assessments: a practical framework for public agency accountability. Tech. rep., AI Now Institute

Richardson R, Schultz J, Crawford K (2019) Dirty data, bad predictions: how civil rights violations impact police data, predictive policing systems, and justice. New York University Law Review Online, Forthcoming

Ridgeway G (2013) The pitfalls of preduction. NIJ J 271:34–40

Robertson K, Khoo C, Song Y (2020) To surveil and predict: a human rights analysis of algorithmic policing in Canada. https://ihrp.law.utoronto.ca/

Saleiro P, Kuester B, Hinkson L, London J, Stevens A, Anisfeld A, Rodolfa KT, Ghani R (2018) Aequitas: a bias and fairness audit toolkit. arXiv:181105577

Santos RB (2019) Predictive policing: where's the evidence? In: Police innovation: contrasting perspectives. Cambridge University Press, p 366

Scantamburlo T, Charlesworth A, Cristianini N (2018) Machine decisions and human consequences. arXiv:181106747

Selbst AD (2017) Disparate impact in big data policing. Ga L Rev 52:109

Shrestha YR, Yang Y (2019) Fairness in algorithmic decision-making: applications in multi-winner voting, machine learning, and recommender systems. Algorithms 12(9):199

Speicher T, Heidari H, Grgic-Hlaca N, Gummadi KP, Singla A, Weller A, Zafar MB (2018) A unified approach to quantifying algorithmic unfairness: measuring individual&group unfairness via inequality indices. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, pp 2239–2248

Verma S, Rubin J (2018) Fairness definitions explained. In: 2018 IEEE/ACM international workshop on software fairness (FairWare), IEEE, pp 1–7

Vestby A, Vestby J (2019) Machine learning and the police: asking the right questions. Policing J Policy Pract

Wang H, Grgic-Hlaca N, Lahoti P, Gummadi KP, Weller A (2019) An empirical study on learning fairness metrics for compas data with human supervision. arXiv:1910.10255

Wexler J, Pushkarna M, Bolukbasi T, Wattenberg M, Viégas F, Wilson J (2019) The what-if tool: interactive probing of machine learning models. IEEE Trans Visual Comput Graph 26(1):56–65
Xiang A, Raji ID (2019) On the legal compatibility of fairness definitions. arXiv:191200761

Authors and Affiliations

Kiana Alikhademi[1] · Emma Drobina[1] · Diandra Prioleau[1] · Brianna Richardson[1] · Duncan Purves[2] · Juan E. Gilbert[1]

Emma Drobina
edrobina@ufl.edu

Diandra Prioleau
dprioleau@ufl.edu

Brianna Richardson
richardsonb@ufl.edu

Duncan Purves
dpurves@ufl.edu

Juan E. Gilbert
juan@ufl.edu

[1] Computer & Information Sciences & Engineering Department, College of Engineering, University of Florida, Gainesville, FL, USA

[2] Department of Philosophy, College of Liberal Arts and Sciences, University of Florida, Gainesville, FL, USA