Synthese (2023) 201:206
https://doi.org/10.1007/s11229-023-04189-0

ORIGINAL RESEARCH

Predictive policing and algorithmic fairness

Tzu-Wei Hung[1,2] · **Chun-Ping Yen**[3]

Received: 26 July 2022 / Accepted: 11 May 2023 / Published online: 5 June 2023

Abstract

This paper examines racial discrimination and algorithmic bias in predictive policing algorithms (PPAs), an emerging technology designed to predict threats and suggest solutions in law enforcement. We first *describe* what discrimination is in a case study of Chicago's PPA. We then *explain* their causes with Broadbent's contrastive model of causation and causal diagrams. Based on the cognitive science literature, we also explain why fairness is not an objective truth discoverable in laboratories but has context-sensitive social meanings that need to be negotiated through democratic processes. With the above analysis, we next *predict* why some recommendations given in the bias reduction literature are not as effective as expected. Unlike the cliché highlighting equal participation for all stakeholders in predictive policing, we emphasize power structures to avoid hermeneutical lacunae. Finally, we aim to *control* PPA discrimination by proposing a governance solution—a framework of a social safety net.

Keywords Predictive policing · Algorithmic bias · Fairness · Discrimination · Social safety net

✉ Chun-Ping Yen
chunping.yen@gmail.com

Tzu-Wei Hung
htw@sinica.edu.tw

[1] Institute of European and American Studies, Academia Sinica, No. 128, Sec. 2, Academia Rd., Nankang District, Taipei 115, Taiwan

[2] Center for Advanced Study in the Behavioral Sciences at Stanford University, 75 Alta Road, Stanford, CA 94305, USA

[3] Department of Philosophy, Soochow University, 111-02 No. 70, Linxi Rd., Shilin Dist., Taipei, Taiwan

1 Introduction

Predictive policing algorithms (PPAs) refer to the use of technologies in data science and artificial intelligence (AI) to predict threats and suggest solutions in law enforcement. Modern-day police are increasingly turning to big data tools to forecast where and when crimes will occur and who might be involved. Although prediction has always been an important part of policing (Berk, 2008), predictive algorithms are considered particularly innovative because they apply AI to datasets previously thought to be too large to analyze (Perry et al., 2013). PPA proponents claim that such initiatives reduce crime, revolutionize public safety, and help underresourced departments better allocate resources; however, critics maintain that they produce self-perpetuating feedback loops of crime prediction, placing historically overpoliced individuals and communities at even greater risk of harm. The main argument is that PPA reproduces the data it is given to learn. When the data that police provide already contain contextual priorities, filtering, and decisions, the results will also reflect these assumptions. Moreover, as police deploy resources based on these predictive results, they produce even more data that confirm what the algorithm has predicted (Richardson et al., 2019; Selbst, 2018).[1]

The literature on the responses to such concerns about algorithmic bias largely uses a technical approach; namely, these concerns are primarily described as an engineering challenge to ensure the output of an algorithm will approximate outcomes required by specific fairness criteria (Dolata et al., 2022; Selbst et al., 2019; Wong, 2020; Žliobaitė, 2017). Various studies have suggested reweighing or filtering and balancing datasets, as well as adapting algorithms, including pre- and post-processing steps (Berk et al., 2021; Kamiran et al., 2013). However, other studies insist that what fairness should mean is unlikely to be answered by a "better" or "fairer" algorithm because the very concept is the criteria for assessing what counts as a better technological solution. Otherwise, circularity occurs. Thus, it is argued that algorithmic fairness is far beyond a technical challenge (Birhane et al., 2022; Dolata et al., 2022; Green, 2022; Huang et al., 2022; Mitchell et al., 2021). Critics of the technical approach hold that definitions of fairness are contestable in many ways, posing an immediate problem to the realization of fairness (Berk et al., 2021; Chouldechova, 2017; Corbett-Davies et al., 2017; Kleinberg et al., 2017). As Wong (2020, p. 231, emphasis original) remarks, it is "not only about designing and implementing algorithms that satisfy some fairness measures but also about *which ideas of 'fairness'* and *what other values* should be considered and accommodated in an algorithm."

Nonetheless, algorithmic fairness is more complicated than a simplified dichotomy of pro- and anti-technical approaches, especially when the controversial PPAs are involved (such that one can agree with an analysis based on one technical approach but disagree with certain others). Recently, Sunstein (2022) argued that human cognition has *bias* and *noise* (i.e., our judgment shows unwanted variability); both can lead to errors. Conversely, while algorithms may inherit human bias, they are noise-free

[1] The European Commission presented the draft of the Artificial Intelligence (AI) Act to the European Parliament on April 21, 2022. The draft includes a ban on using individual risk assessment for predictive policing. Once adopted, the AI act will be immediately applicable in all EU Member States, and its requirements are expected to take effect three years after its implementation.

and thus can help reduce discrimination caused by noise. Thus, "if the goal is to eliminate discrimination, properly constructed algorithms nonetheless have a great deal of promise for administrative agencies" (Sunstein, 2022, p. 1175). While Sunstein (2022) focuses not on PPA but on general AI, his diagnosis of discrimination is consistent with the affirmative view of carefully employed PPA that we have argued elsewhere (Hung & Yen, 2021; Yen & Hung, 2021). However, his prescription of properly constructed algorithms, albeit necessary, is insufficient. For example, except for highly constrained cases, there are trade-offs among different types of fairness.[2] The decision regarding whether a measure of fairness will be acceptable depends on factors beyond the formalized definition of fairness, and it will require balancing fairness with the interests of stakeholders (Huang et al., 2022; Hung & Yen, 2021; Narayanan, 2018). Hence, what matters is not just *what* fairness is but also how to reach an agreement about it. In this sense, algorithmic fairness is more akin to a political matter than merely an engineering or conceptual solution. Thus, if banning PPAs (as argued by Heaven, 2020) is less realistic, it seems that a better governance framework would be desirable.

In this paper, we investigate algorithmic fairness in predictive policing. We first describe racial discrimination reported in Chicago's use of PPAs between 2012 and 2019 (Sect. 2). Based on Broadbent's contrastive model of causation and causal diagrams, we then explain the relationship between the discriminations and factors derived from common criticisms of PPAs (Sect. 3), as well as why fairness is context-sensitive and requires negotiation (Sect. 4). With this analysis, we evaluate several recommendations for bias reduction and predict why some of them may not work (Sect. 5). Finally, we present a governance framework to control the harm of discrimination (Sect. 6). The central proposal of this paper is fourfold: (i) Algorithm revision only has a limited casual role in reducing discrimination in PPAs. (ii) Fairness is not an objective truth to be discovered in a laboratory but has context-sensitive social meanings that need to be negotiated through democratic processes. (iii) Recommendations highlighting "equal participation of all stakeholders" in the PPAs may not work because they fail to notice biased power structures, repeating the same mistake as that of the "All Lives Matter" proponents. (iv) We offer a governance solution based in the social safety net, which can effectively reduce discrimination in PPAs.

[2] As an example, we consider the now paradigmatic COMPAS recidivism algorithm. ProPublica and Northpointe (now Equivant) disagreed over whether the COMPAS recidivism algorithm violated fairness and thus exhibited racial bias based on different understandings of fairness. ProPublica contested that the COMPAS recidivism algorithm was biased, since, for those who did not reoffend, Black defendants were more likely to be incorrectly classified as a higher risk for repeat offence than they actually were, while White defendants were more likely to be incorrectly classified as a lower risk of reoffence than they actually were (Angwin et al., 2016). In response, Northpointe (now Equivant) countered that the algorithm was not biased because the reoffending rate is roughly the same at each COMPAS scale regardless of a defendant's race (Dieterich et al., 2016). Given the different base rates of recidivism for Black and White defendants, it is impossible to satisfy both definitions of fairness simultaneously (Chouldechova, 2017; Kleinberg et al., 2017). There is no "right" definition of fairness. Trying to be fair in one way necessarily means being unfair in another way.

2 Case study: Chicago's PPA

How does a predictive policing algorithm function? What is its controversy? The case of the Chicago Police Department offers a detailed example here.

Chicago was one of the first cities to experiment with a person-based predictive policing strategy in the United States. With funding from the National Institute of Justice, the Chicago Police Department (CPD) developed a PPA in collaboration with the Illinois Institute of Technology's Strategic Subjects List (SSL) in 2012. The SSL used arrest data and crime incident records within the CPD's record management systems to estimate an individual's risk of becoming a victim of such violence over the next eighteen months (Ferguson, 2021). Later, the SSL became the crime and victimization risk model (CVRM). The inputs of the CVRM algorithm include numbers of past shooting victimizations, age at latest arrest, aggravated burglary and assault victimizations, the linear trend of arrests, unauthorized use of weapon arrests, and arrests for violent offenses. The resultant list of subjects and their risk scores were reviewed and then deferred to different police districts to conduct relevant policing interventions, including home visits by police with custom notification letters detailing why these individuals were at risk.[3]

On the one hand, this technology seems promising. It was reported that "among the individuals with the highest CVRM risk scores, approximately 1 in 3 will be involved in a shooting or homicide in the next 18 months" (Illinois Institute of Technology, 2019, p. 3). According to the department's "Violence Reduction Strategy" webpage, the information was reasonably effective in helping prioritize the custom notifications process because "a Chicago resident with no arrests in the past four years has about a 1 in 2300 chance of being a shooting victim [in the next 18 months]" (Chicago Police Department, n.d.b). In a 2019 review (Hollywood et al., 2019, p. 36), the RAND Corporation also concluded that "the CVRM was reasonably effective at identifying a subset of those at a highly elevated risk for being a [party to violence]" and "interventions with the roughly 10,000 people in the highest risk categories could potentially preempt about one-quarter of Chicago's shootings."

On the other hand, this technology seems to be controversial. While the algorithm's inputs do not include variables such as race and gender, SSL and CVRM were criticized as racially biased (Ferguson, 2021). For example, *Chicago Magazine* (Kunichoff & Sier, 2017) profiled the following findings.

1. Fifty-six percent of Black men in the city ages 20 to 29 had a listed score.
2. The data suggested that more people on the list were being arrested than approached for social Par88services.
3. Police say they are not using the list to question or arrest people, but official documents show otherwise.
4. The list was based on arrests rather than convictions.
5. Arrests were concentrated in already heavily policed areas.

Moreover, it was reported that the vast majority of people with the highest score—85 percent—were Black males (Dumke & Main, 2017).

[3] The program was officially discontinued on November 1, 2019.

Why are these findings bringing about racial discrimination concerns?[4] Let us examine these findings in turn. To access Finding 1, we need first to understand what it means for an individual to receive a risk score here. In accordance with the review submitted by the City of Chicago Office of Inspector General (2020), all individuals arrested at least once during the four-year period prior to the start of the Illinois Institute of Technology's calculations in 2012 were assigned a score. Consequently, a person arrested for a nonviolent misdemeanor (such as driving over the speed limit) might have received a risk score, while a victim of a gunshot wound (who was not arrested) would not have been included in the model (The City of Chicago Office of Inspector General, 2020). There are almost 400,000 people in the publicly available SSL dataset; the vast majority have low-risk scores. The fact that a specific group of people had a listed score is not particularly useful in predicting the likelihood of that group being involved in violence. Thus, the list is problematic, and the CPD should make this clear in their public statements when introducing SSL (Hollywood et al., 2019).

Moreover, people's concerns regarding Finding 1 may rest on the fact that the CPD's predictive models generated scores for all individuals arrested, including those ultimately not convicted. Similarly, Finding 4 is problematic because, given that members of groups subject to overpolicing are more likely to be arrested (O'Neil, 2016), failing to consider whether one was actually convicted unavoidably yields inaccurate predictions.[5] In fact, Fogliato et al.'s (2021) survey reveals existing racial bias in arrest data. Therefore, it is helpful to distinguish arrests from convictions whenever possible to adequately protect the rights of individuals who are assessed. The City of Chicago Office of Inspector General (2020) also recommended at least distinguishing an arrest with a conviction from an arrest without a conviction and noting whether the individual who was arrested was ever charged in the first place.

Findings 2, 3, and 5 are related to the operational problems noted in each of the reviews of the CPD's predictive models by RAND (Hollywood et al., 2019) and by The City of Chicago Office of Inspector General (2020). The CPD not only permitted all sworn personnel to access risk scores via its internal dashboards but also failed to provide them with proper training on how to use these risk scores (Hollywood et al., 2019; The City of Chicago Office of Inspector General, 2020). Indeed, there was no supervisor protocol to support compliance with the intended purpose and permissible uses of the predictive models, thereby making misapplication of this information more likely. Accordingly, the CPD is advised to develop protocols guiding the use of information generated by the predictive models, grant access to this information on an as-needed basis, and monitor use. A supervisor protocol supporting compliance will also help the department make timely responses to public concerns, such as the one raised in Finding 3. The silence from the CPD and mayor, however, reinforces people's distrust of the SSL.

Notably, Finding 2 requires more careful reading. Kunichoff and Sire (2017) reported that, in 2016, 1024 custom notifications were attempted by the police; among

[4] To a first approximation, it may be partly because that race is usually conceived as a visible identity attribute of an individual and, as we will elaborate in Sect. 4, is a product of multiple axes of a society's existing systems of power.

[5] Nonetheless, the problematic prediction resulting from biased training samples is not indicative of a problem with the algorithm so much as it is indicative of the skewed sample itself.

them, 558 were completed, and only 26 people attended a call-in meeting. However, the CPD stated that 280 individuals with SSL scores were arrested in four gang raids over six months in the same year. Thus, while Kunichoff and Sire (2017) seem to imply that many people on the list were arrested, it should be clarified that they were not arrested simply for being on the list but for other reasons (e.g., gang involvement). Additionally, regarding Finding 5, Kunichoff and Sire (2017) worried about what they called "a troubling cycle"; police use SSL scores to determine where officers are assigned, which leads to more arrests and higher SSL scores in an already highly monitored area. However, if this cycle is troubling, it seems to be a problem of police actions rather than algorithms (see Sect. 3 for details).

Now we consider that 85 percent of people with the highest score were Black men. Can the CPD's predictive models explain this demographic disparity?[6] To what extent is this demographic disparity discriminatory? To say it is racially discriminatory, for example, is to say that race plays a causal role in determining risk scores. One prominent way a race could play such a role is for it to be an input of the algorithm. Even if race is not included explicitly as such, as in our Chicago case, it could still indirectly determine one's risk score in terms of its proxies, such as ZIP code and family structure (Berk et al., 2021; Calders & Žliobaitė, 2013; Selbst, 2018). Furthermore, Black people and Black men composed 30 and 15 percent of Chicago's population in 2017, respectively, but Black men were victims of approximately 72 percent of homicides in the same year (Chicago Police Department, n.d.a, p. 70). While the demographic disparity of the CPD's predictive list (85 percent of high risk individuals are Black males) does not match the racial composition of the city's population, it should be noted that "shooting victims" also represents one of the attributes used by the CPD's models to generate risk assessment.

We also need to carefully read the racial indicators of the targeted group generated by the predictive models. In a recent analysis of murder trends in Chicago from 1965 to 2020, Sharkey and Marsteller (2022) found that racial and economic segregation has been closely linked to violence over the last five decades in the city. They also found that among Black residents living in majority-Black neighborhoods in Chicago, murder rates are remarkably similar to those for all Black residents, whereas among Black residents living outside majority-Black neighborhoods, murder rates are entirely different.[7] This means that the distribution of violence is not equally distributed across the Black population in Chicago. There is a difference in whether they live in majority-Black neighborhoods, which, due to the set of social, economic, and political forces influencing Chicago over the past five decades, are also the concentrated and persistently disadvantaged neighborhoods in the city. We must take this fact into consideration when analyzing the risk assessment from the predictive models. In the following two sections, we further elaborate on these issues (Sect. 3 for causal analysis and Sect. 4 for structural discrimination).

[6] We use the term "disparity" to refer to the observable difference in outcomes between different groups. The presence of disparity alone does not necessity mean there has been discrimination.

[7] Majority-Black neighborhoods are neighborhoods where at least 50 percent of residents are Black (Sharkey & Marsteller, 2022, p. 360).

3 Where does discrimination come from?

The controversy in the Chicago case lies in its algorithmic output (e.g., 85 percent of people with the highest score were Black males) and crime rate reports (e.g., Chicago's district with the highest crime rate is also a Black community; see below). However, how does the controversy occur?

According to Kleinberg et al. (2018), bias can be decomposed into *algorithmic bias* and *structural bias*. The former fully refers to bias in selecting input variables, selecting output measures, and the training procedure. The latter refers to disparity among social groups that remains after accounting for the three types of algorithmic bias. Based on their distinction, we have four possible sources of bias in the context of the PPAs:

(i) Input: police data that are selectively fed to the PPA
(ii) Output: measures that are used to generate predictions
(iii) Training: police data that are selectively used for machine learning
(iv) Structure: police action (e.g., deployment and decision-making) that systematically reflects social meanings and practices.

However, structural bias may often affect algorithmic bias because construction of the training data and data selection involve specific goals to be achieved. Such goals are value laden (Huang et al., 2022) and often defined by social meanings and practices embedded in power structures that frequently cause discrimination (Haslanger, 2019; Soon, 2020). Moreover, while algorithmic bias is likely to be handled by technology updates, structural bias is not; what matters is existing systematic injustice in society (Haslanger, 2012). Thus, to examine discrimination in the Chicago case, we particularly focus on algorithmic and structural factors that could possibly lead to disproportional distribution results.

We offer a causal explanation of this disparity in the Chicago case (Sect. 3) and explain why fairness is context-sensitive from a cognitive science perspective (Sect. 4). We elaborate on them in turn.

3.1 Causal analysis

We employ Broadbent's (2013, p. 52) contrastive model of causation, which postulates that the right kind of difference making for causation "is a difference between the effect being as it is and the effect being different or absent." Such *effect-led difference making* differs from *cause-led difference making* invoked by the counterfactual approach of causation. Broadbent argues that we often ask 'Why P rather than Q?' rather than simply 'Why P?' because making a difference in the cause-led sense is not sufficient to provide a good causal explanation. Instead, we must mention a causal difference between fact A and foil B. For example, if you ask why a logician arrived late to a lecture rather than on time, it is pointless to mention the presence of oxygen even though it makes a difference to her late arrival in the cause-led sense (e.g., without the presence of oxygen, she would not have arrived at all). Conversely, the fact that her flight was late can explain her late arrival because in the case where she did not arrive

late, the flight arrived on time. This represents a difference from the actual case, while the presence of oxygen does not.[8]

Now, we examine PPA predictions and crime rate reports. First, according to police records, the most arrested suspects are of a relatively young age and are male (Chicago Police Department, n.d.a, p. 82, 84). While it is interesting that the majority are youth, our focus here is on why Black individuals have been targeted.[9] Second, PPA prediction is not based on data on criminal convictions but on arrest records. Therefore, arrest records determine PPA output. In addition, heavily policed areas (i.e., those that have the most police service events) in Chicago also have higher arrest rates. They are not just positively related; police actions such as patrol allocation and deployment affect the number and type of arrests. For instance, O'Neil (2016) indicates that if the police were focused on Part 1 crimes (e.g., homicide and arson), many Part 2 crimes (e.g., drug dealing and aggressive panhandling) would go unrecorded if police were not present to see them. Then, which districts have the most police service and why? According to CPD's *2017 Annual Report* (Chicago Police Department, n.d.a), among 25 police districts, the 11th District staff responded to the most call events (p. 18), including citizen calls for police services, crime responses, and public service activity that police generate while on duty. This is because the 11th District had the most shooting incidents reported and the most homicides in 2017. Additionally, according to the Chicago Department of Public Health (Ann & Robert H. Lurie Children's Hospital of Chicago, 2019), this district's inhabitants have less advantaged socioeconomic and health conditions, such as shorter life expectancy and more opioid-related overdose deaths. It is also reported that its residents are mainly non-Hispanic Black.[10]

Based on these governmental reports, we can derive variables A (arrest records), O (PPA output), I (PPA input and training), D (police actions), C (reported crime rate), and their relations $A \rightarrow I$, $I \rightarrow O$, $D \rightarrow A$, and $C \rightarrow D$. We also know that if an area is predicted by the PPA to be a high-crime area, the police force will also increase, so $O \rightarrow D$. In addition, as mentioned, structural bias (embedded in police actions) could affect algorithmic bias (e.g., tainted data for input and training). Hence, $D \rightarrow I$ and

[8] Broadbent's model offers a better explanation than traditional models. Comparably, a traditional monocausal model may be both too strict and too permissive. It is too strict because some events do not fit it. It is too permissive because the model does not exclude events common to the causal history of many events, such that the Big Bang, using Broadbent's (2013) own example, satisfied as the cause of all events. Since everything is multifactorial in the sense that it takes the operation of multiple causal factors to give rise to it (Broadbent 2012, 2014), a nontrivial causal model cannot simply permit the cataloging of multiple causal factors without discriminating among the causes. Conversely, to preserve the idea that effects resulted from causes in a certain way, Broadbent's model highlights explanatory reasons present in cases of impact but absent in others. In this paper, we choose Broadbent's model to analyze PPA discrimination because it does not appeal to a complete theory of the nature of causation, and we need not give an account of how we can use measures of association to say something about causal facts. Additionally, Broadbent's model is multifactorial, offering a general explanation for the difference between cases with and without the effect resulting from the causes.

[9] Only 29.7 percent of the city's population was Black in 2016. See the United States Census Bureau https://data.census.gov/cedsci/table?q=Chicago%20blacj%20population&g=9700000US3408220&y=2016.

[10] This phenomena are not exclusive to Chicago. For example, Gase et al. (2016, p. 308) examined national data and found that "neighborhood composition (the percent of White residents in the neighborhood)" was the primary driver of racial/ethnic differences in average arrest rates. In addition, racial categories such as "Black" and "non-Hispanic Black" are used in different governmental reports. While acknowledging their distinction, in this paper we use the term "Black" to refer to both for simplicity.

D → O hold. However, we are not sure whether heavy police force deployment leads to a high crime rate (D → C). We are also not sure whether the PPA's prediction of a high-risk area leads to a high reported crime rate in that area (O → C). Thus, further analysis is needed.

It is known that police deployment affects arrest records (D → A). We also know that police action (D) positively correlates with the reported crime rate (C). However, does D cause C? There could be two possible scenarios here. First, let us consider the actual crime rate rather than the reported crime rate. We know that socioeconomic disadvantage is associated with clear increases in rates of self-reported and officially recorded crimes (Fergusson et al., 2004). This association is mediated by the adverse effects of prolonged economic and related pressures on family functioning (Rutter et al., 1998), such as poor childrearing (Brody et al., 1994), parental depression (Conger et al., 1992), and parental behavior (Bolger et al., 1995).[11] Moreover, the poverty rate in the USA is twice as high for Blacks as for Whites (Pew Research Center, 2016). Among the poor, Blacks are twice as likely as Whites to live in high-poverty (> 40 percent) neighborhoods (Kneebone & Holmes, 2016), and the average net worth of Black households is only approximately one-seventh that of Whites (Wolff, 2018). We thus suspect that the higher the arrest rate in a Black community is, the lower the alternative chance of not being arrested is (e.g., obtaining good jobs or education). In this sense, arrest records may affect the crime rate (A → C). Now, based on Broadbent's (2013) "effect-led difference making," we suppose that if the crime rate in heavily policed areas were low, the police force would decrease. In such cases, police force deployment may affect the crime rate. For this hypothetical scenario, we can consider two facts: (i) studies show that social resources[12] help reduce crime in disadvantaged areas (e.g., replacing the police force with education, social welfare, and church support) and (ii) Chicago districts with low crime rates have less police deployment. Therefore, if Chicago has limits on its budget, then the areas where heavy police services have already deployed may have fewer resources to intervene with mediators between crime and socioeconomic disadvantage. Disadvantaged Black youth in these areas may have fewer choices than their wealthy counterparts other than crime activities, causing crime incidents to increase. Thus, in this sense, police action can change the crime rate in these areas, where the contingent majority is Black.

Second, however, there is another possible scenario. Suppose that the actual crime rates in heavily policed areas and other areas show no significant differences. However, due to a lack of police forces in other areas, sufficient arrest data cannot be collected. Thus, the crime rate reported in heavily policed areas will be high and that reported in other areas will be low simply because it is easier to detect suspects with more police forces. In this case, police deployment determines crime rate reports as well. Therefore, in either scenario, D → C holds. Additionally, we know that O → D. Through mediator D, O indirectly affects C. Therefore, police action (e.g., decisions for heavy force deployment) is crucial.

[11] For simplicity, we only discuss this association and do not go into detail on these mediators.

[12] These resources include church (Johnson et al., 2000), family (Fergusson et al., 2004; Sampson & Laub, 1993), peer and neighborhood (Case & Katz, 1991), and education and health resources (Morrisroe, 2014; Rosenfield et al., 2006).

3.2 Causal diagram

The causal diagram can be illustrated as follows. Here, in Fig. 1 (left), we can identify a vicious cycle between D, A, and C when the PPA is not considered (e.g., Chicago before employing PPAs). The higher the deployment of heavy police forces is, the higher the arrest rate in these areas becomes (D → A). The higher arrest rate leads to increased reports of crime (A → C), which in turn makes it necessary for Chicago to deploy more police forces into these areas (C → D). Therefore, if bias in C is what we want to reduce, then D is key. Moreover, when the PPA is introduced (Fig. 1 right), there are intense interactions between multiple factors such that A → I, D → O, D → I, I → O, and O → D. If bias in O must be reduced, D is still crucial because D can affect O both directly and indirectly (through I or A and I). Even if we were to ban the PPA (i.e., remove I and O) as recommended by some critics (Heaven, 2020), bias in D would remain. Therefore, D is at the core of discrimination in predictive policing. Our analysis conforms Ferguson's (2021, p. 244) finding that the racial disparity in the PPA in Los Angeles "lies in policing, not the algorithm."

Moreover, as mentioned, PPA prediction is determined by arrest records. Therefore, if the PPA's output is discriminatory, this should be caused by biased records used in training or as input variables. These records could be produced by police actions such as disproportional police deployment and probable abuse of power. In fact, police brutality incidents and racially motivated violence against Black people are not rare in American history. Innocent individuals are sometimes arrested, and people do not trust the police due to racial bias data, police scandals, or power misuse (Morley et al., 2019; Sheehey, 2019; Susser, 2021). According to Broadbent's model, this race-related abuse in police causally explains the PPA's discriminatory list because in the case where race-related abuse is absent, racial discrimination was not reported (e.g., Japan's Kanagawa Prefectural Police also employed a PPA system, but racial discrimination was not reported). Conversely, racial discrimination remains even after Chicago's SSL was updated with the CVRM. Thus, police action (D) is the main cause of discrimination when employing the PPA. This result also conforms to Brantingham

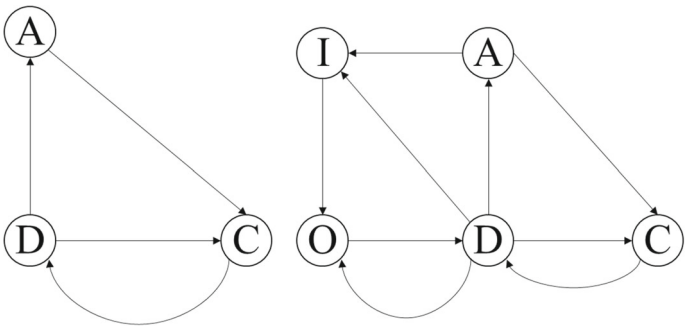

Fig. 1 Causal relationships between A (arrest records), I (PPA's input and training), O (PPA's output), D (police action), and C (reported crime rate). Left: causal relationship before the PPA is introduced. Right: causal relationship after the PPA (i.e., I and O) is introduced

et al.'s (2018) survey of the LAPD showing that employing the PPA does not lead to more biased arrests in any significant sense than not using it. Bias already exists.

3.3 Expanding the analysis to other PPA cases

The above causal analysis applies to other cases in the US. For example, the LAPD uses person-based predictive policing called the Chronic Offender Program (2011–2019), part of Operation LASER (the Los Angeles Strategic Extraction and Restoration Program). It identified "probable offenders" based on a point system built on prior criminal histories, such as arrest records, gang affiliation, probation and parole status, and recent police contacts. If one got stopped by the police or had them knock on the door, for example, it could lead to more points. Police bulletins featured the worst "probable offenders"—those with the most points—with photos and physical descriptions and were posted in the roll call room, distributed to officers during roll call, and uploaded to patrol officers' in-car laptops. While officers cannot detain a person based solely on this information, they are instead instructed to gather intelligence on these chronic offenders during routine patrols (Brayne, 2017).

A recent audit by the LAPD Office of the Inspector General found that the program had been filled with inconsistencies: 44 percent of those with detailed point calculations were listed as having either zero or one arrest for violent crimes; approximately half had no arrest for gun-related crimes, and others were in custody or had been arrested for only nonviolent crimes (Smith, 2019, p. 16). It also noted that the racial/ethnic makeup of chronic offenders on the LAPD's high-point lists in August 2018 roughly approximated the makeup of those arrested for Part I violent crimes from 2012 to 2018 (Smith, 2019, p. 15).[13] Critics have lambasted the program for generating a feedback loop: An individual with a high point value is more likely to be under heightened surveillance and therefore is subject to increased risk of future police contact, which may lead to arrests, as well as further increase the individual's point value (Brayne, 2017; The Stop LAPD Spying Coalition, 2021). In addition, concerns about how suspects were racially identified were raised in the audit (Smith, 2019, p. 15).

In the LAPD case, we again find that police action (D) is the main cause, as it can affect algorithmic prediction (O) through the PPA's input and training data (I) or biased arrest records (A). Likewise, based on Broadbent's model, the LAPD's history of race-related abuse causally explains the PPA's discriminatory tendencies because in the case where there was no race-related abuse, racial discrimination was not reported either. Accordingly, the causal analysis applies to the LAPD case. Two recent studies conform to our analysis that discrimination comes from factors outside algorithms. Mehrotra et al. (2021) reviewed the outcome of using PredPol, an algorithm forecasting areas of likely criminal activities, based on an unsecured cloud database linked from the LAPD's website. The data they found allowed them to analyze the police dispatch for areas where PredPol was known to be used. The analysis showed

[13] Part I violent crimes are murder or non-negligent homicide, rape, robbery, and aggravated assault. The racial/ethnic makeup of chronic offenders on the LAPD's high-points lists in August 2018 was 49.8 percent Hispanic/Latino, 30 percent Black/African American, 12 percent White, and 1.3 percent Other (Smith 2019, p.15).

that PredPol's targeting mirrored existing arrest patterns for the local police. Cruz Cortés et al. (2022) argue that although fair AI applies causal inference interventions to the internal workings of technical objects (i.e., fairness constraints), these interventions remain insufficient to handle discrimination at a systemic level. Instead, they urge detecting bias outside the algorithms and proposing integrated interventions on social dynamics and algorithm design.

Moreover, police departments are often reluctant to disclose complete PPA data for public analysis. The above audit by the LAPD Office of the Inspector General was mainly a product of sustained pressure from the local community. For an audit of an earlier version of CVRM in Chicago, the RAND team had no alternative but to finish the report without the necessary data from the Illinois Institute of Technology (Hollywood et al., 2019). In 2016, the American Civil Liberties Union (ACLU) launched its Community Control Over Police Surveillance (CCOPS) campaign. As of the end of 2022, 22 US jurisdictions have passed CCOPS laws, including New York City, whose council passed the Public Oversight of Surveillance Technology (POST) Act in 2020. Nevertheless, asking police departments to meet the law's disclosure requirements can be challenging.[14] Police have long been criticized for their lack of transparency, which is critical to assess the statistical validity and operational impact of predictive policing systems. Had information about predictive policing systems been more publicly accessible, there would have been more examples that conformed to our analysis.

4 What is fairness?

One may wonder about the implication of causal analysis that police action results in major racial discrimination. Does it mean that PPA cannot help reduce bias? Or does AI help little in improving structural fairness?

Our answers are not wholly pessimistic. First, in our analysis, PPA can still indirectly cause biased arrests via police action. The reason why recent surveys show that Chicago's upgrade (from SSL to CVRM) has had limited effects may be due to not only PPA per se but also how it is employed. The effect may differ if PPA is not simply used to predict person-based risk lists but to detect social inequity that breeds crimes (we shall unpack the relationship between crimes and structural inequity in Sect. 5). In addition, how police react to PPA's output matters. We will argue in Sect. 6 that a better governance framework can improve PPA's efficiency. Second, PPA's difficulty eliminating bias does not imply the failure of other AI technologies. In fact, there has been much progress in identifying various fairness criteria via machine learning, although there are also constraints. This section explores some major limits and possible ways to overcome them.

[14] As US's largest police force, the NYPD strongly opposed disclosing information about the technology used even after passing the POST Act, arguing that any transparency would aid criminals. As required by law, the NYPD published its Impact and Use Policies (IUPs) on April 11, 2021. After reviewing the IUPs, however, the Office of the Inspector General for the NYPD reported on November 3, 2022, that the NYPD used "vague, nonspecific boilerplate language throughout the IUPs" (Strauber & Barrett, 2022, p. 31). The language was so vague that they could not conduct complete analyses to fully inform the public of how the NYPD used surveillance technologies. See also Manis & Cahn (2021) for shortcomings of the NYPD's Implementation of the POST Act.

4.1 Challenges to fairness modeling

To what extent AI technologies can help reduce bias in human society is hotly debated.

On the one hand, many studies employ causal or machine-learning models to identify optimal definitions (or criteria) for the calculus of fairness (Grgic-Hlaca et al., 2016; Hardt et al., 2016; Kusner & Loftus, 2020; Zafar et al., 2017). For instance, Kusner et al. (2017) model fairness using tools from causal inference and present the definition of counterfactual fairness, capturing the intuition that a decision is fair for a person if it holds in both the actual world and a counterfactual world where the person has a different demographic background. These studies focus on algorithmic fairness, but they may also apply to structural fairness because the proposed criteria help evaluate whether a social bias is discriminatory. On the other hand, researchers show that it is mathematically impossible to simultaneously satisfy all criteria or definitions of fairness (Berk et al., 2021; Chouldechova, 2017; Hedden, 2021; Kleinberg et al., 2017), indicating that the algorithm's potential to promote equality as a matter of public policy is fundamentally constrained (Berk et al., 2021; Green, 2022). For example, Berk et al. (2021) argue that neither maximizing accuracy and fairness simultaneously nor satisfying all fairness criteria is possible. They examine cases of criminal justice risk assessments and identify six different concepts of fairness; however, these concepts may be incompatible with one another and with accuracy. Thus, as Berk et al. (2021) argue, conflicts between formal definitions of fairness are inevitable, as are the trade-offs between diverse base rates across different legally protected groups.

Recently, two approaches have been proposed to solve this incompatibility: one normative and the other methodological. For example, Holm (2022) examines four fairness criteria commonly used to evaluate the performance of a predictive algorithm (i.e., equal false-positive rate, equal false-negative rate, equal positive predictive value, and equal negative predictive value). He argues that the four are reducible to Broome's (1990) moral principle that "fairness in the distribution of a good between people consists in the proportional satisfaction of their claims to the good." In such a view, the conflict between the criteria is merely about what grounds we have to claim a good being distributed by AI. Thus, the conflict between the four criteria can be solved. In contrast, Green (2022) suggests a methodological turn from focusing on formal algorithmic fairness (mathematical modeling) to substantive algorithmic fairness (evaluating algorithms in social practices). He argues that mathematical modeling relies on a narrow frame of analysis restricted to isolated decision-making procedures, which often worsen existing oppression and legitimize unjust institutions (Green, 2020, 2022). A more feasible way is not to incorporate substantive equality into a mathematical model but to extend the analysis to encompass the relational and structural factors surrounding particular decision points. Thus, he does not reject formal fairness but proposes to expand it. Green (2022) also acknowledges that while substantive algorithmic fairness helps mitigate the low odds of fairness, it cannot avoid all normative conflict in structural fairness.

To a first approximation, Green's (2022) approach seems to be more adequate for reducing PPA bias because it neither evaluates abstract principles alone nor attempts to create compatibility by sacrificing accuracy. We present two arguments from cognitive

science (Sect. 4.2) and political philosophy (Sect. 4.3) for our claim that AI can help detect bias, but this needs to be done with nonalgorithmic implementation. Thus, an integrated social safety net will be proposed (Sect. 6).

4.2 Lessons from cognitive science

Examining cognitive scientific literature about the descriptive sense of fairness (how it actually works) helps clarify the normative idea of fairness (what it ideally should be). The two notions are different but closely interactive; human intuitions/practices about fairness may turn into moral or legal norms of fairness (slavery or women's suffrage), and the norms, in turn, shape our intuitions/practices about fairness. In addition, since "ought implies can," or that a proposed moral obligation should not go beyond the human capacity to realize it, cognitive sciences help clarify the boundary of the normative criteria of fairness.

On the one hand, cognitive scientists do report that the human brain already has developed mechanisms to detect unfair behavior by early infancy (McAuliffe et al., 2017; Sommerville, 2022), and the cheater detection mechanism in adults developed to automatically and unconsciously identify who owes whom (Van Lier et al., 2013). For instance, Sloane et al. (2012) report that 21-month-old children expected the experimenter to reward them when they all completed assigned tasks, not when one did all the work while the other played. In another experiment, Shaw and Olson (2012) discover that 6- to 8-year-old children tend to distribute items in fair (e.g., equal pay for equal work) and efficient (e.g., maximize available resources) manners. However, when conflict occurs, such as when an unequal amount of a resource must be divided between two equally deserving recipients, children will throw away the rewards (erasers in this case) to avoid unfair distribution. Accordingly, at the descriptive level, the pursuit of fairness seems to be a universal human instinct.

However, on the other hand, what should count as equity remains a debate (Kusner & Loftus, 2020) and often reflects cultural or social differences (Subramanian, 2019). Berman et al. (1985) asked Indian and US subjects to allocate resources in the fairest way. They found that Indians preferred to give resources to the needy over the meritorious more than Americans. Schäfer et al. (2015) discover that children in sampled African societies share resources more equally than their Western counterparts. Bolton et al. (2010) identify consumer perceptions of differences in price fairness; Chinese consumers are more sensitive to in-group versus out-group changes than Americans. Likewise, Strimling and Frey (2020) examine how social contracts about resource distribution converge within communities of online multiplayer games (e.g., World of Warcraft) and discover a wide-ranging diversity in the norms that communities consider fair. Moreover, fairness could evolve in the same society; an exchange agreement for slaves in nineteenth century Yunnan (Gros, 2008) would no longer be considered a "fair trade" in China today. Thus, fairness changes with space and time. It is more like a context-dependent social construct.[15]

[15] Two recent AI studies support this view as well. Schwöbel and Remmers (2022) hold that the fairness of procedures and distributions should not be static; otherwise, structural preconditions and the downstream effects of interventions could be ignored. Chien et al. (2022) hold that machine learning researchers seek to

The view that human fairness is a social construct can be analogous to another sociocognitive skill, language. While we all evolved with common linguistic genes (e.g., Foxp2), this fact does not imply a shared universal grammar or abstract syntax (Christiansen & Chater, 2015; Evans & Levinson, 2009; Hung, 2015, 2019). Humans develop thousands of natural languages to describe surrounding environments, which in turn shape our languages differently. In this sense, language is more like a cultural product than a biological substrate. Likewise, our common capacity to detect unfair behaviors by no means indicates a universal criterion for fairness. This capacity helps us survive diverse scenarios for cooperation in social contexts, which also shapes our ideas differently about what fairness should be. As a result, fairness is more like a social construct than an objective truth (e.g., Pythagorean theorem).

Furthermore, human ideas of fairness are subject to limited cognition. The bounded rationality literature has shown that people make inconsistent gain–loss calculations due to cognitive limits and environmental complexity (Kahneman et al., 1982, 2021). On the one hand, humans can hardly achieve coherence in moral intuitions because of the dynamics of indignation. Moral intuitions depend upon emotions, beliefs, and response tendencies that define indignation, namely, affections such as anger, disgust, and contempt (Kahneman & Sunstein, 2005). On the other hand, our inconsistent judgment results from uncertainty, variability, and complexity in the surrounding world. While algorithms may help us overcome cognitive limits such as noise (Kahneman et al., 2016), the environmental factor remains. This factor is not just a challenge to be overcome but essential to our judgment. For example, the brain evolved to deal with actual environmental changes rather than abstract principles (Cosmides & Tooby, 1992). People's reasoning often violates the formal rules of deductive logic in some selection tasks (e.g., Modus Ponens, $p \rightarrow q, p, \therefore q$. Wason, 1968). However, if the tasks are translated into a more specific situation (e.g., if she drinks beer, she must be over the legal drinking age. She drinks beer; therefore, she must be over the legal drinking age. Cox & Griggs, 1982), people reason significantly better. This is because the brain requires concrete situations to trigger sociocognitive mechanisms like emotions (e.g., loss aversion), and emotion is central to moral/political judgments (Haidt, 2012; Hung & Hung, 2022). Principally, a person should save not one but five lives in the classic trolley problem. However, her decision may be different if the one to be sacrificed is her beloved daughter. Saving one against five is irrational or imperfect regarding maximum utility in saved lives, but this relational-emotional (and somehow selfish) decision is how the mind works; it is human nature. Hence, to bridge the gap between descriptive reality and normative proposal, formal principles must be expanded into concrete situations for stakeholders to negotiate tradeoffs.

All these examples show that (a) fairness is a context-sensitive social construct. Its criteria should be determined in reality by societies rather than as an objective truth to be discovered in an isolated laboratory.[16] (b) While AI may help us overcome

Footnote 15 continued
address equity in different applications, but many overlook that fairness is context-dependent and domain-specific.

[16] Reader et al. (2022) have a similar diagnosis. They find that while many studies of algorithmic fairness focus on closed systems with a specific decision-maker and particular engagement, real societies are not closed systems, and there is no singular decision-maker or defined agent behavior rules. Nonetheless, Reader et al.'s (2022) prescription differs from ours.

certain cognitive limits (e.g., noise), others may be essential to humanity. If human-centered AI is desirable, not all our imperfections should be eliminated; some need to be acknowledged and affirmed in normative solutions.

4.3 Lessons from political philosophy

The literature on political philosophy has shown that fairness is not just about distribution (e.g., equal opportunity and demographic parity) or gain–loss calculation.

In addition to social structure, many elements of fairness have been explored, including self-identity (Young, 1990), personal need (Marx, 1875), social relationships (Anderson, 1999), and citizens' moral powers (Rawls, 1963, 2005). For example, Anderson (1999) maintains that the point of fairness is not distributional but relational, and equal relationships among citizens should be prioritized. While Rawls (1963, 2005) highlights fairness in distributive justice, he also values each citizen's moral powers for the sense of justice and good. Those elements (e.g., the sense of justice and civic relations between individuals) are unlikely to be quantified in nonreductive ways; we are not living in Disney World, and neither Black Lives Matter nor PPA's racial issues can be simplified into conflicting of formal criteria for fairness. Furthermore, even if we can reduce the criteria of fairness into notions of distributive justice, there are always trade-offs. In reality, which criterion should outweigh another often involves who should benefit less than others; it thus needs dialogue among stakeholders. For example, given limited budgets, should a government buy influenza vaccines for everyone or expensive orphan drugs for the economically disadvantaged? Which policy is more just? There is neither a universal nor a "right" answer. It requires citizens of different societies to reach agreements that accommodate the challenges they face, given the resources they have. Therefore, in this sense, fairness is more about politics than sciences.

We return to the incompatibility issue. Green's (2022) substantive fairness better fits our goal of reducing PPA bias than Holm's (2022) definition because while Holm resolves the incompatibility of the four criteria, there are always other criteria (Kusner & Loftus, 2020; Narayanan, 2018). That the four terms can all hold does not imply that other unstated terms do too. In addition, fairness reaches far beyond Holm's distribution of goods, and the solution only handles consequential disparity (inequality) instead of initial disparity (inequity). Additionally, as cognitive science shows, moral intuitions are closer to emotion than logic. Hence, we need to expand formal fairness to substantive evaluation and to settle various tradeoffs through social agreement.

Accordingly, in jural implementation, PPA fairness should be a social construct that requires democratic processes. In liberal societies, consensus relies on political consultation and negotiation among the people (e.g., registration representatives and advocacy groups), the government, and the jurisprudence system. As each society has its own cultural and historical legacies and thus its own challenges, a universal guideline or definition of fairness is less likely to be desirable. For example, while racial discrimination is a key issue in US law enforcement (Aougab et al., 2020), most factors affecting public cooperation with the police in Japan are not racial (Tsushima & Hamai, 2015); instead, Japanese society has long been confronted with gender

inequity (Belarmino & Roberts, 2019). Even in the US and EU, sexism, racism, and other bigotries will manifest differently (Wachter et al., 2021). Therefore, an agreement on fairness should be reached by respective societies in individual democracies. The democratic process ensures transparency and audit mechanisms and can be held accountable if something goes wrong. Hence, what algorithmic fairness is or the normative goal to be achieved should not be determined merely by engineering but through public negotiation. In short, while algorithms have the potential to improve structural fairness, this potential cannot be realized without democratic procedures.

Nonetheless, what are the implications of using algorithms to support fair policy reform if fairness is context-sensitive? More specifically, one might wonder that given that fairness is not fixed, how can we ever get to a situation where PPAs can be used as part of a solution to discrimination and bias in criminal justice and law enforcement?

We hold that PPAs remain beneficial if properly integrated with nonalgorithmic implementation. When designing algorithms, developers need to acknowledge lived social realities and be aware of the real-world impacts of algorithms. For example, in 2019, Black Americans comprised 36 percent of arrests for violent crimes, and White Americans comprised 59 percent. Blacks represented only 13 percent of the US population that year, while Whites represented 60 percent.[17] Crime results partly from social and economic forces. The disproportionate rate should not be surprising, given that Blacks in the US are more likely to experience concentrated urban poverty. An accurate PPA trained on these data will reflect the group disparities in the data, a result similar to the above Chicago case. This, in turn, would subject those already burdened with structural discrimination to adverse outcomes and further exacerbate social inequality. We cannot treat the status quo as neutral (Green, 2022; Wachter et al., 2021). With respect to technical work, the usage of fairness metrics is not a neutral but a normative decision. Instead of evaluating fairness metrics on the basis of mathematical tests alone, developers can acknowledge social realities and then play a more active role in dismantling social inequality by choosing appropriate fairness metrics.[18] If unjustified disparities are detected, they may change decision criteria, add variables, or adjust the weights of existing variables whenever applicable.

Of course, algorithms alone cannot remedy all social problems. We have proposed implementing predictive algorithms in the framework of a larger social safety net (Hung & Yen, 2021; Yen & Hung, 2021). This human-centered approach to governance expands the scope of analysis to include structural bias and discrimination so that it provides guidance for using PPAs to promote equitable social changes. Accordingly, PPAs can be used to detect systematic inequality by revealing disparities in crime prediction, as the statistics show in Sect. 2. The group disparities in the output of the SSL and CVRM in the Chicago case are indications and consequences of contingent social arrangements, including police actions (Sect. 3.1).

[17] Crime in the United States, Table 43A: Arrests by Race and Ethnicity, 2019, United States Department of Justice, Federal Bureau of Investigation Uniform Crime reporting Program, https://ucr.fbi.gov/crime-in-the-u.s/2019/crime-in-the-u.s.-2019/topic-pages/tables/table-43.

[18] For a recent development on this front, see Wachter et al. (2021). They provide a classification scheme for fairness metrics based on their treatment of historical social bias to help developers choose appropriate fairness metrics.

Before exposing how PPAs can help advance equality from the perspective of the social safety net, we will first diagnose the inequality involved and then suggest some potential reforms to mitigate the existing problems in the next section.

5 Structural bias and discrimination in the PPA

The distribution of violence is not equally distributed across racial groups or even the Black population in Chicago, as noted in Sect. 2. These disparities are not neutral facts about the world nor merely reflect measurement bias. The risk of being a party to violence is not an intrinsic and neutral attribute of individuals. Rather, it is the product of structural discrimination that generates disparities in social and material resources. In this section, we first propose a structural explanation of the disparities in the PPA's predictive outputs and then discuss its indications for the direction and steps of effective police intervention to reduce violent crime.

5.1 Explaining disparities in violence

Past and present discrimination has created social conditions in the city of Chicago in which Black people are empirically at higher risk of being involved in violent crimes. While racial disparities in violence may reflect biases in the way criminal justice institutions treat different groups based on their race and ethnicity rather than based on differences in actual offences, other aspects of social and economic disadvantage, such as poverty, segregation, and unemployment, also contribute to the high rates of violence (Lauritsen et al., 2018; Sharkey & Marsteller, 2022; Wilson, 2012). The effects of such intersectional discrimination belong to structural injustice, reflecting multiple and overlapping forms of oppression (Crenshaw, 2015).[19] After World War II, loans guaranteed by the Federal Housing Administration and the Department of Veteran Affairs opened up the possibility of homeownership to millions of American households. However, these loan programs were explicitly structured to exclude Black people and favor the newly minted suburbs (Sharkey & Sampson, 2015). These processes collectively confined Black city residents to neighborhoods overlooked and underserved by local governments, financial institutions, and private developers. As a result, neighborhoods were divided with profound differences in employment opportunities, poverty rates, education quality, access to health care, crime exposure, and more. Increases in inner city violence led to migrations of families from city centers and further amplified racial, ethnic, and economic segregation. The United States remains highly residentially segregated by race despite improvements made since the 1960s (Cheon et al., 2020; Lauritsen et al., 2018; Sharkey & Marsteller, 2022). The household wealth of Black families is systematically lower than that of White families, including lower home values. In addition, majority-Black neighborhoods are less

[19] Intersectionality is a viewpoint to examine how a person's multiple identities (e.g., middle-class Black female) may result in different discriminatory effects. The importance of intersectionality has been noticed in recent algorithmic studies, such as employing intersectional critical concepts to AI design (Klumbytė et al., 2022), using intersectional constraints to reduce implicit bias (Mehrotra et al., 2022), analyzing feminicide counterdata (Suresh et al., 2022), and applying intersectionality to XAI (Huang et al., 2022).

likely to enjoy political influence and to receive public or private investment. The lack of institutional resources creates neighborhoods of concentrated disadvantage that are vulnerable to violence (Sharkey & Marsteller, 2022). In Chicago, a large city with "rigid segregation by race, ethnicity, and income," regions with the highest violence rates are in low-income areas with larger populations of Blacks (Sharkey & Marsteller, 2022, p. 351; Walker et al., 2016). The reduction of violent crime rates and racial disparities may require more than easing poverty because even among neighborhoods of the same socioeconomic status, residential segregation may put Black individuals at higher risk of being shot (Cheon et al., 2020; Sharkey, 2014). Moreover, while concentrated poverty is directly associated with lower collective efficacy, understood as "the combination of shared expectations for informal social control and social cohesion," communities exhibiting higher levels of collective efficacy are characterized by lower violence rates, regardless of their demographic and economic composition (Sharkey & Sampson, 2015, p. 327). In other words, there are structural power dynamics between the police and marginalized communities.

5.2 Evaluation and prediction

The above analysis of causality (Sect. 3) and structural discrimination (Sect. 5.1) helps predict that some recommendations made in the bias reduction literature may not be as effective as expected. For example, Sunstein (2022) insightfully argues that properly constructed algorithms can avoid cognitive noise and the discrimination it causes. However, the PPA's discrimination, at least in the Chicago case, is primarily caused by human bias (i.e., police action) resulting from unbalanced power and unjust social structures instead of by problematic algorithms. Since the point of predictive policing is not to make arrests but to reduce the number of targeted crimes from happening by alleviating criminogenic conditions for the targeted groups, Sunstein's account is insufficient to engineer injustice out of algorithms without a supplement from a broader human-centered perspective. Only a human-centered approach to analyzing the scope of social relationships and institutional arrangements can address and challenge the necessary conditions that breed discrimination in policing.

Similarly, recent advocates for equal participation of all stakeholders in AI are insufficient to reduce discrimination in PPA. Stakeholders refer to people whose interests and rights could be impacted by PPAs. While their interests and rights vary, stakeholders could include natural persons (e.g., citizens and police) and legal persons (NGOs and PPA service providers). On the one hand, stakeholders are often regarded as a key element for improving fairness (and transparency). Recent studies have argued that PPA, or AI more broadly, needs to include stakeholders and encourage diverse participation in different aspects. These studies highlight the need to ensure that all stakeholders are involved in the technology's research, design, employment, and explanation, and the policy-making process (Seele, 2017; Macnish et al., 2020; Biderman & Scheirer, 2020; Cohen & Graver, 2021; Langer et al., 2021). Such a requirement could safeguard stakeholder rights and increase the legitimacy of using such technologies

as well as improve the design and development of AI. Although defining who stakeholders are remains to be clarified, many studies have emphasized the importance of equal participation.

However, on the other hand, as police action is the main cause of discrimination in PPAs, it would be helpful to improve the checks and balances of police power and manage predictive algorithms in a larger governance framework of a social safety net (see Sect. 6). While stakeholder involvement is to check power, it could backfire because equal participation may replicate or amplify an unjust social structure. These accounts fail to account for existing patterns of injustice and leave the causal link between discrimination and oppressive structure intact. Just as "All Lives Matter" proponents fail to understand the systematic injustice against Black Americans (Lebron, 2017), emphasizing equal participation for all stakeholders may confront the same difficulty. Thus, we need to give more resources to underrepresented stakeholders in their participation to ensure that their voice is well represented in public policies and social arrangements. For example, the government can grant more seats for local youth and NGOs (e.g., Data for Black Lives) on the committee for employing predictive technology in law enforcement. This remedy better avoids hermeneutical lacuna—the lack of proper understanding and linguistic expression (among the PPA's decision-makers) of a disadvantaged population's experience with discrimination (Fricker, 2006; Haslanger, 2019).[20]

Moreover, as noted in Sect. 5.1, it may not be possible to produce a sustained power rebalance among the various population groups without addressing the spatial concentration of advantages and disadvantages. Andrew Ferguson (2021, p. 283), for example, recently advanced a tyrant test model of regulatory constraint that aims to shift power from technology companies, police departments, and government institutions into marginalized communities and initiate democratic community control over policing by creating "a legislatively enacted but community-based power structure."[21] Accordingly, technologies such as PPAs would not be allowed to operate unless approved by a group of technology-informed experts and juries summoned from the residents of impacted communities. Ferguson's model has the merit of acknowledging structural power dynamics between the police and marginalized communities, but this model overlooks the persistent inequality of neighborhoods and its effects. For Ferguson's local oversight jury to well represent community interests, it requires juries to have shared expectations for the social control and social cohesion of their communities. Such collective efficacy, however, varies from neighborhood to neighborhood. Collective efficacy is negatively associated with neighborhood violence (Sampson, 2012; Sharkey & Marsteller, 2022; Sharkey & Sampson, 2015). On the one hand, communities with higher levels of collective efficacy exhibit lower rates of violence. On the other hand, collective efficacy also predicts future variations in violence across neighborhoods. Using Cooper's (2018) distinction, what matters here is not equality (offering the same treatment to each person without discrimination) but equity (not only offering people the same treatment without discrimination but also considering

[20] Linguistic injustice is another example of lacking proper understanding of a disadvantaged population's experience with discrimination. Please see Yen and Hung (2019) and Yen (2021).

[21] The idea is to presume that technologies such as PPAs will be abused by a metaphorical tyrant and then focus on how to limit the potential harms of the use of these technologies.

their needs, differences, and situations). The problem of Ferguson's account is an example of how easy it is to underestimate the scarcity of critical resources needed for those who live in concentrated disadvantage. Granting more resources to the underrepresented helps balance power and reduce abuse that could lead to discrimination. For neighborhoods of resource deprivation across multiple domains, investments in organization-based resources[22] are a key priority in establishing community control over policing.

6 Solutions

We have proposed a policy schema of the social safety net for predictive policing that seeks to identify the sources of discrimination and remedy the resultant social harms by providing an alternative understanding of the meaning and distribution of risk (Hung & Yen, 2021; Yen & Hung, 2021). It suggests integrating PPA within a broader social safety net. Our proposal shares the assumptions with many predictive policing programs that physical and social environments may encourage predictable acts of criminal wrongdoing and that interfering with that environment would deter would-be crimes. Accordingly, for predictive policing programs to succeed in reducing crime, "[g]enerating predictions is just half of [the business] process; taking actions to interdict crimes is the other half" (Perry et al., 2013, p. xxii). Change requires cross-departmental collaboration within the government to better identify interventions the targeted individuals or communities need. The police department is only one of those segments.[23]

One common challenge of the current predictive policing practices among law enforcement agencies, however, lies in the need for more specific guidance on what these actions are (Ferguson, 2017; Hollywood et al., 2019; Perry et al., 2013; Saunders et al., 2016; Smith, 2019; The City of Chicago Office of Inspector General, 2020). The schema of the social safety net provides more robust guidance on the program's overall goal and intervention practices. It reminds us not to overlook the consequences of the policies that the algorithm facilitates (The Stop LAPD Spying Coalition, 2021). When a risk assessment labels an individual "high risk," there are consequences that the individual bears. It is crucial to reduce the stigma and discrimination associated with the "high risk" label by renovating the meaning of the label. In the schema of the social safety net, PPAs are used to predict immediate risks and socially vulnerable individuals, enabling subsequent assistance and support. This schema's strategy is to

[22] A list of such resources includes: "community newspaper, neighborhood watch, block group or tenant association, crime prevention program, alcohol/drug treatment program, family planning clinic, mental health center, youth center, afterschool recreational programs for youth, counseling or mentoring services (e.g., Big Brother), crisis intervention center, and mental health clinics for children" (Sampson, 2012, p. 191). See also Invest-divest, Movement for Black Lives, https://m4bl.org/policy-platforms/invest-divest/.

[23] For example, in Canada, such social networks are called "hub models" or "situation tables" of tracking risk. This kind of model involves systematic information sharing between service providers from various sectors (education, addictions, social work, and mental health, for example) and law enforcement agencies. (Roberson et al., 2020). For US-based examples, see von Ulmenstein & Sultan (2011) and Kennedy & Friedrich (2014). For UK-based examples, see Babuta & Oswald (2020) and Crawford & Evans (2012).

intervene in the causal link between an unequal social structure and crime, which helps reduce discrimination as well. Specifically, as crime often links up with socioeconomic disadvantages, it is crucial in crime-fighting operations to offer help (e.g., job training, education, job placement, and health services) in terms of improving welfare instead of arresting or sending warnings to anyone on the PPA's list. One merit of a properly employed PPA is that it helps detect systematic inequity by revealing disparities in crime prediction among city areas. Disparities often reflect the unequal distribution of social resources such as opportunities and wealth. For example, in 2017 Chicago, areas with more homicides were also reported to have more dental-related emergency room visits. This correlation could indicate a common factor of economic disadvantages that needed to be seriously handled in these areas. Although Chicago's police records systems can help identify needed interventions for the targeted group (area-based PPA) or individuals (person-based PPA), this requires lateral communication among government departments to better understand what services and interventions are needed (Hollywood et al., 2019). This is why this schema stresses community resources (e.g., the role of the neighborhood) to assist the disadvantaged, and this bottom-up participation benefits the separation of powers and responsibilities in crime prevention. This social safety net schema is decentralized; it empowers the local community to develop a mutual aid network to better understand the needs, differences, and situations of its members so that more proper distribution of social resources can be enabled to achieve equity (and not just equality). In fact, Sharkey and Marsteller (2022) point out that in neighborhoods with concentrated disadvantages (e.g., high-poverty, majority-Black neighborhood; low collective efficacy; etc.), the proportion of violent offenders and victims is higher than in other districts. Not only does prevalent violence lead to disinvestment in communities, but the threat of victimization also changes young people's behavior and network formation. Hence, this schema reverses the concentrated disadvantage to reduce crime.

Changing the unjust social structure also helps reduce racial discrimination in PPA. Police action (e.g., deployment and decision-making) systematically reflects social meanings and practices. If they are discriminatory, so is police action. Akbar's (2018, pp. 449–450) analysis reaches a similar conclusion that the police are central to the devaluation of Black individuals. As she puts it, "The rise of mass incarceration, overcriminalization, and zero-tolerance or broken windows policing is seen as an evolution of the regime of control, exclusion, and exploitation that began with slavery, convict leasing, the Black Codes, and segregation." Accordingly, by targeting the biased social structure in which we (and the police) are embedded, this schema can reduce the foundation of discriminatory police action.

Of course, PPAs within this schema should be watched and reviewed by the public because a large proportion of public fear and distrust of PPAs is due to poor communication between police and communities. The public audit also needs to ensure that any individual whose rights are violated shall have an effective remedy, which requires collaboration from multidisciplinary researchers, policy-makers, citizens, developers, and designers in the endeavor.

In summary, we conclude that (i) discrimination in PPAs is not primarily caused by problematic algorithms but by biased police actions reflecting our unjust social structures. Thus, what should be addressed is not algorithms but structures. (ii) Normatively,

fairness is not an objective truth to be discovered in laboratories but has context-sensitive social meanings that need to be negotiated through democratic processes. (iii) Crimes are caused by multiple unequal conditions (e.g., poverty and insufficient support for health, education, and family), and attributing crimes to a specific race is a categorization error. However, emphasizing the equal participation of all stakeholders has a limited effect in terms of changing an unjust structure. (iv) The social safety net schema aims to better support the underrepresented so that discrimination in a PPA can be reduced by intervening in systematic injustice and balancing resources.

Acknowledgements Sincere thanks to the editors and two anonymous reviewers for their insightful comments. This research was supported by the Stanford-Taiwan Social Science Fellowship Program and the National Science Council, Taiwan.

Funding This research is sponsored by the Ministry of Science and Technology Taiwan (110-2628-H-001-005-MY2).

Declarations

Conflict of interest We declare that this research involves no conflict of interest. Both authors confirm that this research involves no data archiving and sharing (no data are available). This research involves neither human nor animal experiments and requires no ethical approval. This research involves no experimental subjects and requires no patient consent. This research does not involve reproducing material from other sources. In short, I confirm that the manuscript adheres to ethical guidelines specified in the APA Code of Conduct as well as my national ethics guidelines.

References

Akbar, A. A. (2018). Toward a radical imagination of law. *New York University Law Review, 93*, 405–479.

Anderson, E. S. (1999). What is the point of equality? *Ethics, 109*(2), 287–337. https://doi.org/10.1086/233897

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. *ProPublica*, May 23, 2016. Retrieved March 17, 2022, from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Ann & Robert, H. Lurie Children's Hospital of Chicago. (2019). Community Health Needs Assessment for Chicago youth, adolescents and families. Retrieved July 24, 2022, from https://www.luriechildrens.org/globalassets/documents/luriechildrens.org/community/community-health-needs-assessment/chna-2019.pdf

Aougab, T., Ardila, F., Athreya, J., Goins, E., Hofman, C., Kent, A., Khadjavi, L., O'Neil, C., Patel, P., & Wehrheim, K. (2020). Boycott collaboration with police. *The Notices of the American Mathematical Society, 67*(9), 1293.

Babuta, A., & Oswald, M. (2020). *Data analytics and algorithms in policing in England and Wales: Towards a new policy framework*. Royal United Services Institute for Defence and Security Studies. Retrieved February 8, 2023, from https://static.rusi.org/rusi_pub_165_2020_01_algorithmic_policing_babuta_final_web_copy.pdf

Belarmino, M., & Roberts, M. R. (2019). Japanese gender role expectations and attitudes: A qualitative analysis of gender inequality. *Journal of International Women's Studies, 20*(7), 272–288.

Berk, R. (2008). Forecasting methods in crime and justice. *The Annual Review of Law and Social Science, 4*, 219–238. https://doi.org/10.1146/annurev.lawsocsci.3.081806.112812

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2021). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research, 50*(1), 3–44. https://doi.org/10.1177/0049124118782533

Berman, J. J., Murphy-Berman, V., & Singh, P. (1985). Cross-cultural similarities and differences in perceptions of fairness. *Journal of Cross-Cultural Psychology, 16*(1), 55–67. https://doi.org/10.1177/0022002185016001005

Biderman, S., & Scheirer, W. J. (2020). Pitfalls in machine learning research: Reexamining the development cycle. https://doi.org/10.48550/arXiv.2011.02832

Birhane, A., Ruane, E., Laurent, T., Brown, M. S., Flowers, J., Ventresque, A. & Dancy, C. L. (2022). The forgotten margins of AI ethics. *FAccT '22, June 21–24, 2022, Seoul, Republic of Korea*, 948–958. https://doi.org/10.1145/3531146.3533157

Bolger, K. E., Patterson, C. J., Thompson, W. W., & Kupersmidt, J. B. (1995). Psychosocial adjustment among children experiencing persistent and intermittent family economic hardship. *Child Development, 66*, 1107–1129. https://doi.org/10.1111/j.1467-8624.1995.tb00926.x

Bolton, L. E., Keh, H. T., & Alba, J. W. (2010). How do price fairness perceptions differ across culture? *Journal of Marketing Research, 47*(3), 564–576. https://doi.org/10.1509/jmkr.47.3.564

Brantingham, P. J., Valasik, M., & Mohler, G. O. (2018). Does predictive policing lead to biased arrests? Results from a randomized controlled trial. *Statistics and Public Policy, 5*(1), 1–6. https://doi.org/10.1080/2330443X.2018.1438940

Brayne, S. (2017). Big data surveillance: The case of policing. *American Sociological Review, 82*(5), 977–1008. https://doi.org/10.1177/0003122417725865

Broadbent, A. (2012). Causes of causes. *Philosophical Studies, 158*, 457–476. https://doi.org/10.1007/s11098-010-9683-0

Broadbent, A. (2013). Philosophy of epidemiology. *Palgrave Macmillan*. https://doi.org/10.1057/9781137315601

Braodbent, A. (2014). Disease as a theoretical concept: The case of "HPV-it is." *Studies in History and Philosophy of Biological and Biomedical Sciences, 48*, 250–257. https://doi.org/10.1016/j.shpsc.2014.07.010

Brody, G. H., Stoneman, Z., Flor, D., McCrary, C., Hastings, L., & Conyers, O. (1994). Financial resources, parent psychological functioning, parent co-caregiving, and early adolescent competence in rural two-parent African-American families. *Child Development, 65*, 590–605. https://doi.org/10.2307/1131403

Broome, J. (1990). Fairness. *Proceedings of the Aristotelian Society, 91*, 87–101.

Calders, T., & Žliobaitė, I., et al. (2013). Why unbiased computational processes can lead to discriminative decision procedures. In B. Custers (Ed.), *Discrimination & privacy in the information society: Data mining and profiling in large databases* (pp. 43–57). Springer.

Case, A., & Katz, L. F. (1991). The company you keep: The effects of family and neighborhood on disadvantaged youths. *NBER Working Paper No. w3705*. https://doi.org/10.3386/w3705

Cheon, C., Lin, Y., Harding, D. J., Wang, W., & Small, D. S. (2020). Neighborhood racial composition and gun homicides. *JAMA Network Open, 3*(11), e2027591. https://doi.org/10.1001/jamanetworkopen.2020.27591

Chicago Police Department. (n.d.a). *The 2017 annual report*. Retrieved May 4, 2023, from https://home.chicagopolice.org/wp-content/uploads/2017-Annual-Report.pdf

Chicago Police Department. (n.d.b). Violence reduction strategy. Retrieved March 17, 2022, from https://home.chicagopolice.org/information/violence-reduction-strategy-vrs/

Chien, I., Deliu, N., Turner, R., Weller, A., Villar, S., & Kilbertus, N. (2022, June). Multi-disciplinary fairness considerations in machine learning for clinical trials. In *2022 ACM conference on fairness, accountability, and transparency* (pp. 906–924). https://doi.org/10.1145/3531146.3533154

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data, 5*(2), 153–163. https://doi.org/10.1089/big.2016.0047

Christiansen, M. H., & Chater, N. (2015). The language faculty that wasn't: A usage-based account of natural language recursion. *Frontiers in Psychology, 6*, 1182. https://doi.org/10.3389/fpsyg.2015.01182
Cohen, I. G., & Graver, H. (2021). What big data in health care can teach us about predictive policing. In J. McDaniel & K. Pease (Eds.), *Predictive policing and artificial intelligence* (pp. 111–131). Routledge.
Conger, R. D., Conger, K. J., Elder, G. H., Lorenz, F. O., Simons, R. L., & Whitbeck, L. B. (1992). A family process model of economic hardship and adjustment of early adolescent boys. *Child Development, 63*, 526–541. https://doi.org/10.1111/j.1467-8624.1992.tb01644.x
Cooper, B. (2018). *Eloquent rage: A Black feminist discovers her superpower*. Picador.
Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. *Proceedings of KDD '17*, 797–806. https://doi.org/10.1145/3097983.3098095
Cosmides, L., & Tooby, J. (1992). Cognitive adaptations for social exchange. In J. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture* (pp. 163–228). Oxford University Press.
Cox, J. R., & Griggs, R. A. (1982). The effects of experience on performance in Wason's selection task. *Memory & Cognition, 10*, 496–502. https://doi.org/10.3758/BF03197653
Crawford, A., & Evans, K. (2012). Crime prevention and community safety. In A. Liebling, S. Maruna, & L. McAra (Eds.), *The Oxford handbook of criminology* (5th ed., pp. 769–805). Oxford University Press.
Crenshaw, K. (2015, September 24). Why intersection can't wait. *The Washington Post*. Retrieved July 24, 2022, from https://www.washingtonpost.com/news/in-theory/wp/2015/09/24/why-intersectionality-cant-wait/
Cruz Cortés, E., Rajtmajer, S., & Ghosh, D. (2022, June). Locality of technical objects and the role of structural interventions for systemic change. In *2022 ACM conference on fairness, accountability, and transparency* (pp. 2327–2341). https://doi.org/10.1145/3531146.3534646
Dieterich, W., Mendoza, C., & Brennan, T. (2016). COMPAS risk scales: demonstrating accuracy equity and predictive parity. Northpoint Inc. Retrieved March 17, 2022, from http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf
Dolata, M., Feuerriegel, S., & Schwabe, G. (2022). A sociotechnical view of algorithmic fairness. *Information Systems Journal, 32*(4), 754–818. https://doi.org/10.1111/isj.12370
Dumke, M., & Main, F. (2017, May 18). A look inside the watch list Chicago police fought to keep secret. *Chicago Sun-Times*. Retrieved March 17, 2022, from https://chicago.suntimes.com/news/what-gets-people-on-watch-list-chicago-police-fought-to-keep-secret-watchdogs
Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences, 32*(5), 429–448. https://doi.org/10.1017/S0140525X0999094X
Ferguson, A. G. (2017). Policing predictive policing. *Washington University Law Review, 94*(5), 1109–1189.
Ferguson, A. G. (2021). Surveillance and the tyrant test. *The Georgetown Law Journal, 110*, 205–290.
Fergusson, D., Swain-Campbell, N., & Horwood, J. (2004). How does childhood economic disadvantage lead to crime? *Journal of Child Psychology and Psychiatry, 45*(5), 956–966. https://doi.org/10.1111/j.1469-7610.2004.t01-1-00288.x
Fogliato, R., Xiang, A., Lipton, Z., & Chouldechova, A. (2021). On the validity of arrest as a proxy for offense: Race and the likelihood of arrest for violent crimes. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 100–111. https://doi.org/10.1145/3461702.3462538
Fricker, M. (2006). Powerlessness and social interpretation. *Episteme, 3*(1–2), 96–108. https://doi.org/10.3366/epi.2006.3.1-2.96
Gase, L. N., Gleen, B. A., Gomez, L. M., Kuo, T., Inkelas, M., & Ponce, N. A. (2016). Understanding racial and ethnic disparities in arrest: The role of individual, home, school, and community characteristics. *Race and Social Problems, 8*, 296–312. https://doi.org/10.1007/s12552-016-9183-8
Green, B. (2020). The false promise of risk assessments: Epistemic reform and the limits of fairness. *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 594–606). https://doi.org/10.1145/3351095.3372869
Green, B. (2022). Escaping the impossibility of fairness: From formal to substantive algorithmic fairness. *Philosophy & Technology, 35*, 90. https://doi.org/10.1007/s13347-022-00584-6
Grgic-Hlaca, N. et al. (2016). The case for process fairness in learning: Feature selection for fair decision making. *NeurIPS Symposium on Machine Learning and the Law*.
Gros, Stéphane, (2008) The Salt, the ox and the slave: Exchange and politics in northwest Yunnan, 19th~20th centuries. In *Luobu Jiangcun & X. Zhao (Eds.), Kang-Zang yanjiu xin silu: wenhua, lishi*

yu jingji fazhan (New directions in Tibeto-Kham studies: Culture, history, and economic development, pp. 107–115). Minzu chubanshe.
Haidt, J. (2012). *The righteous mind: Why hood people are divided by politics and religion*. Vintage.
Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems, 29*, 3315–3323.
Haslanger, S. (2012). Resisting reality: Social Construction and Social Critique. *Oxford University Press*. https://doi.org/10.1093/acprof:oso/9780199892631.001.0001
Haslanger, S. (2019). Cognition as a social skill. *Australasian Philosophical Review, 3*(1), 5–25. https://doi.org/10.1080/24740500.2019.1705229
Heaven, W. D. (2020, July 17). Predictive policing algorithms are racist. They need to be dismantled. *MIT Technology Review*. Retrieved March 10, 2022, from https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/
Hedden, B. (2021). On statistical criteria of algorithmic fairness. *Philosophy and Public Affairs, 49*(2), 209–231. https://doi.org/10.1111/papa.12189
Hollywood, J. S., Mckay, K. N., Woods, D., & Agniel, D. (2019). *Real-time Crime Centers in Chicago*. Rand Corporation. Retrieved March 17, 2022, from https://www.rand.org/content/dam/rand/pubs/research_reports/RR3200/RR3242/RAND_RR3242.pdf
Holm, S. (2022). The fairness in algorithmic fairness. *Res Publica*. https://doi.org/10.1007/s11158-022-09546-3
Huang, L. T. L., Chen, H. Y., Lin, Y. T., Huang, T. R., & Hung, T. W. (2022). Ameliorating algorithmic bias, or why explainable AI needs feminist philosophy. *Feminist Philosophy Quarterly, 8*(3/4), 2.
Hung, T. W. (2015). How sensorimotor interactions enable sentence imitation. *Minds and Machines, 25*(4), 321–338. https://doi.org/10.1007/s11023-015-9384-8
Hung, T.-W. (2019). How did language evolve? Some reflections on the language parasite debate. *Biological Theory, 14*(4), 214–223. https://doi.org/10.1007/s13752-019-00321-x
Hung, T.-W., & Yen, C.-P. (2021). On the person-based predictive policing of AI. *Ethics and Information Technology, 23*, 165–176. https://doi.org/10.1007/s10676-020-09539-x
Hung, T. C., & Hung, T. W. (2022). How China's cognitive warfare works: A frontline perspective of Taiwan's anti-disinformation wars. *Journal of Global Security Studie*s, 7(4), ogac016. https://doi.org/10.1093/jogss/ogac016
Illinois Institute of Technology. (2019). Crime and victimization risk model (CVRM) fact sheet. Retrieved March 17, 2022, from https://home.chicagopolice.org/wp-content/uploads/2019/01/FACT-SHEET-Crime-and-Victimization-Risk-Model-1.pdf
Johnson, B. R., Larson, D. B., De Li, S., & Jang, S. J. (2000). Escaping from the crime of inner cities: Church attendance and religious salience among disadvantaged youth. *Justice Quarterly, 17*(2), 377–391. https://doi.org/10.1080/07418820000096371
Kahneman, D., Rosenfield, A. M., Gandhi, L., & Blaser, T. (2016). Noise: How to overcome the high, hidden cost of inconsistent decision making. *Harvard Business Review* (October, 2016), 36–43.
Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A flaw in human judgment*. Little, Brown Spark.
Kahneman, D., Slovic, S. P., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press.
Kahneman, D., & Sunstein, C. R. (2005). Cognitive psychology of moral intuitions. In J.-P. Changeux, A. R. Damasio, W. Singer, & Y. Christen (Eds.), *Neurobiology of human values* (pp. 91–105). Springer.
Kamiran, F., Calders, T., & Pechenizkiy, M. (2013). Techniques for discrimination-free predictive models. In B. Custers, T. Calders, B. Schermer, & T. Zarsky (Eds.), *Discrimination and privacy in the information society: Data mining and profiling in large databases* (pp. 223–239). Springer.
Kennedy, D. M., & Friedrich, M. (2014). *Custom notifications: Individualized communication in the group violence intervention. Office of Community Oriented Policing Services*. Retrieved February 8, 2023, from https://nnscommunities.org/wp-content/uploads/2017/10/GVI_Custom_Notifications_Guide.pdf
Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2018). Discrimination in the age of algorithms. *Journal of Legal Analysis, 10*, 113–174. https://doi.org/10.1093/jla/laz001
Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017) Inherent trade-offs in the fair determination of risk scores. *Proceedings of 8th Innovations in Theoretical Computer Science Conference*. https://doi.org/10.4230/LIPIcs.ITCS.2017.43

Klumbytė, G., Draude, C., & Taylor, A. S. (2022, June). Critical tools for machine learning: Working with intersectional critical concepts in machine learning systems design. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1528–1541). https://doi.org/10.1145/3531146.3533207

Kneebone, E., & Natalie, H. (2016, March 31). U.S. concentrated poverty in the wake of the great recession. Brookings Institute. Retrieved July 24, 2022, from https://www.brookings.edu/research/u-s-concentrated-poverty-in-the-wake-of-the-great-recession/

Kunichoff, Y., & Sier, P. (2017, Aug. 21). The contradictions of Chicago police's secretive list. Chicago Magazine. Retrieved March 17, 2022, from https://perma.cc/2PDQ-53FW

Kusner, M. J., & Loftus, J. R. (2020). The long road to fairer algorithms. *Nature, 578*, 34–36.

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. *Advances in Neural Information Processing Systems, 30*.

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & Baum, K. (2021). What do we want from Explainable Artificial Intelligence (XAI)?—A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence, 296*, 103473. https://doi.org/10.1016/j.artint.2021.103473

Lauritsen, J., Heimer, K., & Lang, J. B. (2018). The enduring significance of race and ethnic disparities in male violent victimization: An analysis of NCVS micro-data, 1973–2010. *Du Bois Review: Social Science Research on Race, 15*(1), 69–87. https://doi.org/10.1017/S1742058X18000097

Lebron, C. J. (2017). *The making of black lives matter: A brief history of an idea*. Oxford University Press.

Macnish, K., Wright, D., & Jiya, T. (2020). Predictive policing in 2025: A scenario. In H. Jahankhani, B. Akhgar, P. Cochrane, & M. Dastbaz (Eds.), *Policing in the era of AI and smart societies* (pp. 199–215). Springer.

Manis, E., & Cahn, A. F. (2021). Above the law?: NYPD violations of the Public Oversight of Surveillance Technology (POST) Act. Retrieved 6 Feb., 2023, from https://static1.squarespace.com/static/5c1bfc7eee175995a4ceb638/t/615df7245561b315e7289cee/1633548068620/2021.10.7_Above+the+Law_Research+Report.pdf

Marx, K. (1875). In *Karl Marx and Frederick Engels: Selected works* (Vol. 3, pp. 13–30). Progress Publishers

McAuliffe, K., Blake, P. R., Steinbeis, N., & Warneken, F. (2017). The developmental foundations of human fairness. *Nature Human Behaviour, 1*(2), 1–9. https://doi.org/10.1038/s41562-016-0042

Mehrotra, A., Pradelski, B. S., & Vishnoi, N. K. (2022, June). Selection in the presence of implicit bias: the advantage of intersectional constraints. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 599–609). https://doi.org/10.1145/3531146.3533124

Mehrotra, D., Mattu, S., Gilbertson, A., & Sankin, A. (2021). How we determined predictive policing software disproportionately targeted low-income, black, and Latino neighborhoods: A trove of unsecured data allowed the first-ever independent analysis of actual crime predictions across the U.S. by the self-described software leader, PredPol. *Gizmodo*. Retrieved 6 Feb., 2023, from https://gizmodo.com/how-we-determined-predictive-policing-software-dispropo-1848139456

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application, 8*, 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2019). From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and Engineering Ethics, 26*, 2141–2168. https://doi.org/10.1007/s11948-019-00165-5

Morrisroe, J. (2014). *Literacy changes lives 2014: A new perspective on health, employment and crime*. National Literacy Trust. Retrieved May 4, 2023, from https://files.eric.ed.gov/fulltext/ED560667.pdf

Narayanan, A. (2018). 21 fairness definitions and their politics. In Tutorial presented at the Conference on Fairness, Accountability, and Transparency. Retrieved March 15, 2022, from https://www.youtube.com/watch?v=jIXIuYdnyyk

O'Neil, C. (2016). *Weapons of math destruction*. Crown Books.

Perry, W. L., McInnis, B., Price, C. C., Smith, S. C., & Hollywood, J. S. (2013). *Predictive policing: The role of crime forecasting in law enforcement operations*. Rand Corporation. Retrieved March 15, 2022, from https://www.rand.org/content/dam/rand/pubs/research_reports/RR200/RR233/RAND_RR233.pdf

Pew Research Center. (2016, June 27). On views of race and inequality, blacks and whites are worlds apart. Retrieved March 15, 2022, from http://www.pewsocialtrends.org/2016/06/27/1-demographic-trends-and-economic-well-being/.

Rawls, J. (1963). The sense of justice. *The Philosophical Review, 72*(3), 281–305.

Rawls, J. (2005). *Political liberalism*. Columbia University Press.
Reader, L., Nokhiz, P., Power, C., Patwari, N., Venkatasubramanian, S., & Friedler, S. (2022, June). Models for understanding and quantifying feedback in societal systems. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1765–1775). https://doi.org/10.1145/3531146.3533230
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, 94*, 15–55.
Roberson, K., Khoo, C., & Song, Y. (2020). *To surveil and predict: A human rights analysis of algorithmic policing in Canada*. Retrieved February 8, 2023, from https://citizenlab.ca/wp-content/uploads/2020/09/To-Surveil-and-Predict.pdf
Rosenfield, S., Phillips, J., & White, H. (2006). Gender, race, and the self in mental health and crime. *Social Problems, 53*(2), 161–185. https://doi.org/10.1525/sp.2006.53.2.161
Rutter, M., Giller, H., & Hagell, A. (1998). *Antisocial behavior by young people*. Cambridge University Press.
Sampson, R. (2012). *Great American City: Chicago and the enduring neighborhood effect*. University of Chicago Press.
Sampson, R. J., & Laub, J. H. (1993). *Crime in the making: Pathways and turning points through life*. Harvard University Press.
Saunders, J., Hunt, P., & Hollywood, J. S. (2016). Predictions put into practices: A quasi-experimental evaluation of Chicago's predictive policing pilot. *Journal of Experimental Criminology, 12*(3), 347–371. https://doi.org/10.1007/s11292-016-9272-0
Schäfer, M., Haun, D. B., & Tomasello, M. (2015). Fair is not fair everywhere. *Psychological Science, 26*(8), 1252–1260. https://doi.org/10.1177/0956797615586188
Schwöbel, P., & Remmers, P. (2022, June). The long arc of fairness: Formalisations and ethical discourse. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 2179–2188). https://doi.org/10.1145/3531146.3534635
Seele, P. (2017). Predictive sustainability control: A review assessing the potential to transfer big data driven 'predictive policing' to corporate sustainability management. *Journal of Cleaner Production, 153*, 673–686. https://doi.org/10.1016/j.jclepro.2016.10.175
Selbst, A. D. (2018). Disparate impact in big data policing. *Georgia Law Review, 52*(1), 109–195.
Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68. https://doi.org/10.1145/3287560.3287598
Sharkey, P. (2014). Spatial segmentation and the Black middle class. *American Journal of Sociology, 119*(4), 903–954. https://doi.org/10.1086/674561
Sharkey, P., & Marsteller, A. (2022). Neighborhood inequality and violence in Chicago, 1965–2020. *University of Chicago Law Review, 89*(2), 349–381.
Sharkey, P., & Sampson, R. (2015). Violence, cognition, and neighborhood inequality in America. In R. K. Schutt (Ed.), *Social neuroscience: Brain, mind, and society* (pp. 320–339). Harvard University Press.
Shaw, A., & Olson, K. R. (2012). Children discard a resource to avoid inequity. *Journal of Experimental Psychology: General, 141*(2), 382. https://doi.org/10.1037/a0025907
Sheehey, B. (2019). Algorithmic paranoia: The temporal governmentality of predictive policing. *Ethics and Information Technology, 21*, 49–58. https://doi.org/10.1007/s10676-018-9489-x
Sloane, S., Baillargeon, R., & Premack, D. (2012). Do infants have a sense of fairness? *Psychological Science, 23*(2), 196–204. https://doi.org/10.1177/0956797611422072
Smith, M. P. (2019). Review of selected Los Angeles Police Department data-driven policing strategies (Los Angeles: Office of the Inspector General, Los Angeles Police Department, March 12, 2019). Retrieved 6 Feb., 2023, from https://www.oig.lacity.org/_files/ugd/b2dd23_21f6fe20f1b84c179abf440d4c049219.pdf
Sommerville, J. A. (2022). Developing an early awareness of fairness. In In M. Killen, & J. G. Smetana (Eds.), *The handbook of moral development* (3rd ed., pp. 153–167). Routledge.
Soon, V. (2020). Implicit bias and social schema: A transactive memory approach. *Philosophical Studies, 177*, 1857–1877. https://doi.org/10.1007/s11098-019-01288-y
Strauber, J., & Barrett, J. (2022). An assessment of NYPD's response to the POST Act. (Office of the Inspector General for the NYPD, November 3, 2022) Retrieved 6 Feb, 2023, from https://www.nyc.gov/assets/doi/reports/pdf/2022/POSTActReport_Final_11032022.pdf

Strimling, P., & Frey, S. (2020). Emergent cultural differences in online communities' norms of fairness. *Games and Culture, 15*(4), 394–410. https://doi.org/10.1177/1555412018800650
Subramanian, K. R. (2019). Cultural differences and perception of fairness in organizations. *International Journal of Research in IT and Management, 9*(1), 8–17.
Sunstein, C. R. (2022). Governing by algorithm? No noise and (potentially) less bias. *Duke Law Journal, 71*(6), 1175–1205.
Suresh, H., Movva, R., Dogan, A. L., Bhargava, R., Cruxen, I., Cuba, Á. M., Taurino, G., So, W., & D'Ignazio, C. (2022, June). Towards intersectional feminist and participatory ML: A case study in supporting Feminicide Counterdata Collection. In *2022 ACM conference on fairness, accountability, and transparency* (pp. 667–678). https://doi.org/10.1145/3531146.3533132
Susser, D. (2021). Predictive policing and the ethics of preemption. In B. Jones & E. Mendieta (Eds.), *The ethics of policing: An interdisciplinary perspective* (pp. 268–292). New York University Press.
The City of Chicago Office of Inspector General. (2020). Advisory concerning the Chicago Police Department's predictive risk models.
The Stop LAPD Spying Coalition. (2021). *Automating banishment: The surveillance and policing of looted land*. Retrieved 6 Feb, 2023, from https://automatingbanishment.org/assets/AUTOMATING-BANISHMENT.pdf
Tsushima, M., & Hamai, K. (2015). Public cooperation with the police in Japan: Testing the legitimacy model. *Journal of Contemporary Criminal Justice, 31*(2), 212–228. https://doi.org/10.1177/1043986214568836
Van Lier, J., Revlin, R., & De Neys, W. (2013). Detecting cheaters without thinking: Testing the automaticity of the cheater detection module. *PLoS ONE, 8*(1), e53827. https://doi.org/10.1371/journal.pone.0053827
von Ulmenstein, S., & Sultan, B. (2011). *Group violence reduction strategy: Four case studies of swift and meaningful law enforcement responses*. U.S. Department of Justice. Retrieved February 8, 2023, from https://nnscommunities.org/wp-content/uploads/2017/10/LE_Case_Studies.pdf
Wachter, S., Mittelstadt, B., & Russell, C. (2021). Bias prevention in machine learning: The legality of fairness metrics under EU Non-discrimination Law. *West Virginia Law Review, 123*(3), 735–790.
Walker, G. N., McLone, S., Mason, M., & Sheehan, K. (2016). Rates of firearm homicide by Chicago region, age, sex, and race/ethnicity, 2005–2010. *Journal of Trauma and Acute Care Surgery, 81*(4), S48–S53. https://doi.org/10.1097/TA.0000000000001176
Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology, 20*(3), 273–281. https://doi.org/10.1080/14640746808400161
Wilson, W. J. (2012). *The truly disadvantaged: The inner city, the underclass, and public policy* (2nd ed.). University of Chicago Press.
Wolff, E. N. (2018). The decline of African-American and Hispanic wealth since the Great Recession. *Working Paper No. 25198*. https://doi.org/10.3386/w25198
Wong, P.-H. (2020). Democratizing algorithmic fairness. *Philosophy & Technology, 33*, 225–244. https://doi.org/10.1007/s13347-019-00355-w
Yen, C. P. (2021). Linguistic diversity in philosophy. In D. Ludwig, I. Koskinen, Z. Mncube, L. Poliseli, & L. Reyes-Galindo (Eds.), *Global epistemologies and philosophies of science* (pp. 26–38). Routledge.
Yen, C. P., & Hung, T. W. (2019). New data on the linguistic diversity of authorship in philosophy journals. *Erkenntnis, 84*, 953–974. https://doi.org/10.1007/s10670-018-9989-4
Yen, C.-P., & Hung, T.-W. (2021). Achieving equity with predictive policing algorithms: A social safety net perspective. *Science and Engineering Ethics, 27*, 36. https://doi.org/10.1007/s11948-021-00312-x
Young, I. M. (1990). *Justice and the politics of difference*. Princeton University Press.
Zafar, M. B., Valera, I., Rogriguez, M. G. & Gummadi, K. P. (2017). Fairness constraints: Mechanisms for fair classification. In *Proceedings of the 20th international conference on artificial intelligence and statistics, PMLR* (Vol. 54, pp. 962–970). http://proceedings.mlr.press/v54/zafar17a.html
Žliobaitė, I. (2017). Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery, 31*(4), 1060–1089. https://doi.org/10.1007/s10618-017-0506-1