




Achieving Equity with Predictive Policing Algorithms: A Social Safety Net Perspective

Chun-Ping Yen¹ · Tzu-Wei Hung¹ 

Received: 15 January 2021 / Accepted: 7 May 2021 / Published online: 1 June 2021
© The Author(s), under exclusive licence to Springer Nature B.V. 2021

Abstract

Whereas using artificial intelligence (AI) to predict natural hazards is promising, applying a predictive policing algorithm (PPA) to predict human threats to others continues to be debated. Whereas PPAs were reported to be initially successful in Germany and Japan, the killing of Black Americans by police in the US has sparked a call to dismantle AI in law enforcement. However, although PPAs may statistically associate suspects with economically disadvantaged classes and ethnic minorities, the targeted groups they aim to protect are often vulnerable populations as well (e.g., victims of human trafficking, kidnapping, domestic violence, or drug abuse). Thus, determining how to enhance the benefits of PPA while reducing bias through better management is important. In this paper, we propose a policy schema to address this issue. First, after clarifying relevant concepts, we examine major criticisms of PPAs and argue that some of them should be addressed. If banning AI or making it taboo is an unrealistic solution, we must learn from our errors to improve AI. We next identify additional challenges of PPAs and offer recommendations from a policy viewpoint. We conclude that the employment of PPAs should be merged into broader governance of the social safety net and audited publicly by parliament and civic society so that the unjust social structure that breeds bias can be revised.

Keywords Predictive policing algorithm · Artificial intelligence · Equity · Racism · Social safety net

✉ Tzu-Wei Hung
htw@sinica.edu.tw

Chun-Ping Yen
chunping.yen@gmail.com

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

Introduction

Whereas predictive policing—using artificial intelligence (AI) technologies to predict threats in law enforcement—seems to be initially successful in Germany and Japan (Egbert & Krasmann, 2020; Ohyama & Amemiya, 2018), the killing of Black Americans by police in the US has sparked massive protests, raising doubts about whether police budgets should be redirected from predictive technology to other facilities to serve the community better. Despite its potential in crime prevention, critics and policymakers have questioned the use of predictive policing algorithms (PPAs). They are concerned that the algorithms may replicate or amplify the disparities inherent in police-recorded data and thus potentially lead to flawed or unlawful predictions (Castelvecchi, 2020; Ensign et al., 2018; Heaven, 2020; Morley et al., 2019; Richardson et al., 2019; Sheehy, 2019), which in turn risk perpetuating racist patterns of policing (Angwin et al., 2016; Heaven, 2020; Kusner & Loftus, 2020; Selbst, 2017). Many critics believe that PPAs should be dismantled in law enforcement. Recently, the City Council of Santa Cruz, California, became the first city in the US to pass an ordinance to ban predictive policing because the technologies can be “disproportionately biased against people of color,” as Mayor Justin Cummings said in a Reuters report (Asher-Schapiro, 2020). Not long after, Pittsburgh City Councilors introduced an ordinance to restrict predictive police techniques (Deto, 2020). In the Netherlands and Canada, there are also civil rights activists calling on law enforcement to halt the use of algorithmic systems until legislative safeguards are put in place (Amnesty International, 2020; Roberson et al., 2020).

In contrast, more countries have a cautious but positive attitude towards PPAs. For example, the law enforcement and policymaking community in the UK and Canada have endeavored to develop a new policy framework with standardized processes to regulate police use of PPAs (Babuta & Oswald, 2020; Roberson et al., 2020). In addition, the Japanese government is ready to adopt PPAs at the Tokyo Olympic Games, for which the Kanagawa Prefectural Police is using the Hitachi AI system for crime prediction and prevention (Hitachi Inc, 2019). Data scientists also devote themselves to improving algorithm fairness (Kusner & Loftus, 2020). In other words, there is no consensus on the usage of PPAs. Although most NGO reports cast doubt on PPAs, think tanks and consulting firms’ analysis points out their potential advantages (Babuta & Oswald, 2020; Hollywood et al., 2019; Jenkins & Purves, 2020; Perry et al., 2013). Recently, American mathematicians have boycotted collaborations with brutal police and for-profit contractors (Aougab et al., 2020). Nonetheless, they do not call for banning PPAs but encourage colleagues to audit PPAs with professional knowledge and to work with NGOs (e.g., Data 4 Black Lives and Black in AI) for more transparent PPA applications.

The debate is even more complicated in multiple aspects due to other factors. First, strictly speaking, banning predictive algorithms in law enforcement and dismantling predictive policing are not the same. The former refers to halting AI prediction technologies in police, courts, and corrections (e.g., risk assessment in parole). The latter refers to halting the use of non-AI statistical analysis to predict

possible risk in policing as well.¹ Second, given law enforcement's diverse tasks, ranging from criminal investigation, highway patrol, child protection, and foreign affairs to counterterrorism, predictive algorithms may have broad applications in many of these tasks. Predictive algorithms could be questioned in one assignment but not in another. Third, the technologies involved in predictive policing could be different, in which some may cause huge controversy (e.g., face recognition), but others do not (e.g., data mining). Fourth, there are types of applications that should be distinguished (Hung & Yen, 2020). *Area-based* policing focuses on the time and place in which crimes are more likely to occur. *Event-based* policing focuses on the type of activity that is more likely to occur. *Person-based* policing, the most controversial one, focuses on the individual who is more likely to be involved in criminal acts. As such, the debate should not be whether we should utterly ban or implement the technology; instead, a subtler assessment of specific tasks, technologies, and applications should be the focus. Hence, to avoid confusion, this article primarily focuses on whether we should ban AI technologies²—person-based predictive algorithms—in community crime prevention. This article aims to explore to what extent we should use and develop PPAs and how to use them.

To this end, the “[Criticisms And Analysis](#)” section will first classify and examine common objections to PPAs. We will clarify which are the genuine challenges and which are merely rhetoric. The “[Explainability, Accountability, and Communication](#)” section will discuss other problems of PPAs, which are also common problems faced by most, if not all, AI applications. The “[Policy Schema of Social Safety Net](#)” section will propose a solution to reduce the issues discussed (e.g., distrust, bias prediction, transparency, and social inequality). The final section, “[Conclusions](#)”, discusses further questions regarding PPAs (e.g., whether the same algorithms can be used to detect implicit vulnerable groups or unjust social structures). This article concludes with the view that banning PPAs helps little to solve the problems; instead, integrating PPAs into a broader social safety net can help reduce racism.

Criticisms and Analysis

This section categorizes and reviews major criticisms against the development and employment of predictive policing algorithms. We discover that some complaints, albeit intuitive and capturing human fear of AI, lack clarity to a certain degree. For example, they articulate neither which type of predictive policing nor what kind of technologies (e.g., real-time face-recognition scanning or criminal data mining) is the concern. To avoid confusion, we primarily focus on person-based predictive

¹ For example, the RAND Safety and Justice Program defines predictive policing as “the application of analytical techniques—particularly quantitative techniques—to identify likely targets for police intervention and crime prevention or solve past crimes by making statistical predictions” (Perry et al., 2013, pp. 1–2). It also notes that “[t]he use of statistical and geospatial analyses to forecast crime levels has been around for decades.” “[A] surge of interest in analytical tools that draw on very large data sets to make predictions in support of crime prevention” occurs only recently (Perry et al., 2013, p. 2).

² For example, the applications include PredPol, PreCobs, Hunchlab and Crime Anticipation System, etc. (Hardyns & Rummens, 2018).

policing, while other types may also be discussed wherever needed. We also distinguish three common types of reasons against PPAs (Section “[Categorizations of the Criticism of PPAs](#)”) and clarify which criticisms are real challenges and which are less realistic (“[Replies to the Criticisms](#)”).

Categorizations of the Criticism of PPAs

Current criticisms of employing predictive policing algorithms can be generally sorted into three groups by the targets of critics, including those against *organizational abuse of power* (e.g., by police administrators and service providers), against *tools per se* (algorithms and training data), and against *possible remedy* (both technical and ethical solutions).

First, many critics of PPAs are against organizational abuse of power. They do not trust the police due to racial bias data, police scandals, or misuse of police power; for example, in American history (Morley et al., 2019; Sheehey, 2019; (Susser 2021), forthcoming). This distrust was aggravated by recent police brutality incidents and racially motivated violence against Black people in the US. Additionally, there seems to be a lack of checks and balance in some US police departments, not only because the local legislature is often unaware of the police use of PPA but also because AI companies do not share technical details due to trade secrets (Castelvecchi, 2020; Heaven, 2020). This latter issue also raises whether PPAs are compatible with the current legal system: Are there already effective avenues for appeal and accountability, or do we need a new set of regulations? Likewise, there are problems with preemptive profiling. When the police mark individuals as suspicious, regardless of any overtly suspicious behavior, the PPA robs people of their right to be presumed innocent or fails to respect individuals as full moral subjects (Hosein, 2018; Kerr & Earle, 2013; Shapiro, 2017).

Second, some criticisms focus on the technology per se, either the *algorithms* or the training *data*. For example, the data input may be problematic, and algorithms could replicate the existing human biases inherent in the dataset or even amplify the existing biases by creating a feedback loop (Amnesty International, 2020; Schuilenburg, 2021; Sheehey, 2019). Similarly, the algorithms may unexpectedly produce biased output. Even with good intentions, software engineers may sometimes design PPAs that reinforce societal injustices by imposing disproportionate impacts on specific groups of people, which results in the issue of what counts as algorithm fairness (Angwin et al., 2016; Heaven, 2020; Kusner & Loftus, 2020; Selbst, 2017). In addition, there are difficulties regarding responsibility and accountability: it is difficult to detect the harm and find its cause with PPA. It is thus difficult to assign responsibility for any harm caused by the deployment of PPA (Castelvecchi, 2020; Richardson et al., 2019; Zarsky, 2013). Opponents also criticize the false beliefs presumed by advocates that PPAs cost less and are more objective than humans (Heaven, 2020). Although algorithms, unlike humans, have no intrinsic preference for discrimination and no ulterior motives (Kleinberg et al., 2018; Goel et al., 2018), they seem to

make discrimination easier. In other words, the above worries altogether cast significant doubt on the effectiveness of the algorithms: Do they truly work?

Third, opponents also doubt the adequacy of the possible remedy. On the one hand, it seems that technology will not fix the problems. Studies show that bias prediction could be reduced by taking protected features such as race and gender explicitly into account (Skeem & Lowenkamp, 2020); an algorithm accessing race can help achieve racial justice as it could maximize the positive predictive value and, at the same time, minimize racial imbalance in error rates.³ However, this kind of remedy is thought to be controversial and discriminatory because it applies different standards to ethnic groups (Heaven, 2020). On the other hand, it also seems that ethics will not fix the problems either because genuine action often gets replaced by superficial promises and abstract guidelines, leading to the trap of ethics-washing. According to Hao (2019), the AI ethics of many organizations are still vague and difficult to implement. Few companies have made significant changes to the way AI products and services are evaluated and approved. Actual actions are replaced by superficial promises. For example, Google has established a nominal AI ethics committee, which has no actual veto power over problematic projects, and the addition of several members has caused controversy. Strong opposition immediately led to its dissolution. Thus, it seems that both technological and ethical remedies help little.⁴

Replies to the Criticisms

While the above criticisms sound intuitive and capture human fear of AI, they need to be carefully examined. In this section, we discuss six major criticisms across the above three categories to identify the genuine challenges of PPAs.

First, regarding the criticism against organizational abuse of power, what happens in the US by the police against Black Americans is intolerable. Reducing racial prejudice in law enforcement is a top priority, but the question is, how? In fact, racial inequality can breed various biases and stereotypes and undermine mutual trust among the police and people in the US. In other countries, the police and people do not necessarily have this racial tension⁵ but have other issues (e.g., gender bias). Thus, to reduce discrimination, we should focus on *the unequal structure* breeding bias in each society. This unequal structure could lead to tensions between

³ See also Kleinberg et al. (2018) and Skeem and Lowenkamp (2016). If the inclusion of feature-specific factors helps advance justice, it may be permissible to do so. For example, recently, in *Wisconsin v. Loomis* (2016), the Wisconsin Supreme Court noted that women are typically less likely to participate in crime and held that a trial court's use of an algorithmic risk assessment that took gender into account served the nondiscriminatory purpose of promoting accuracy.

⁴ As one of our anonymous reviewers pointed out, the problem of ethics washing is usually the failure to track if the ethics guidelines are actually implemented in practice. AI ethics guidelines have told us "the 'what' of AI ethics," but still, there is "the 'how' [question] of AI ethics"—how to translate these guidelines into practices (Morley et al., 2019). Ethics washing is a pointer to the need to bridge the gap between the ethics discourses on the one hand and the technical ones on the other.

⁵ In Japan, for example, most factors affecting public cooperation with the police are not racial (Tsushima & Hamai, 2015). Also, in Germany and Japan, albeit they still have room to improve, area-based predictive policing is reported to be initially successful (Egbert & Krasmann, 2020; Ohya & Amemiya, 2018).

epistemology and ethics. Epistemologically, statistics indicate that crime rates are positively correlated with economic inequality, and the more inequality in the economy, the higher the crime rate (Fajnzylber et al., 2002). This correlation may cause the police in the US to incorrectly associate a particular race with certain crimes. However, ethically, it is wrong to presume the relationship between specific ethnic groups and crime rates. In other words, eliminating the discrimination caused by unequal social structure is the key to reducing racial bias in the US. In contrast, banning PPA helps little to solve these historical problems.

Critics have also argued that by marking individuals as suspicious people, regardless of any overtly suspicious behavior, predictive profiling technologies rob them of the principle of presumed innocence (Shapiro, 2017; (Susser 2021), forthcoming). We think this criticism is too quick. There are two possible replies here. Firstly, it is unclear why predictive profiling is identical to a violation of presumed innocence. Predictive algorithms typically target the individuals who are more likely to be involved in criminal acts, either as victims or perpetrators. The PPAs cannot identify who the potential suspects are. The underlying idea of predictive profiling is to assist the police in making better decisions and distribute police resources more effectively. Thus, whether a targeted individual is considered a suspect is not determined by algorithms but by human police. Analogically, in public health, imposing a rapid test of asymptomatic passengers does not imply that they are positive, but the fact that medical professionals do not know who are virus carriers; that is why screening is needed. Likewise, predictive profiling does not mean that the investigated people are suspects but merely that the police do not know who the suspect is; thus, profiling is needed. Secondly, had predictive profiling been equal to violating presumed innocence, it would remain unclear why this principle cannot be violated for reasonable suspicion. People who fit some general profiles can be stopped and questioned at airports, train stations, buses, or other public transportation. Indeed, in these cases, the law enforcement authorities need to justify their move. It nonetheless shows that predictive profiling can be justified. The key is not on the non-infringement of the right to be presumed innocent but the justification of the infringement. Also, the fact that some AI guidelines are too abstract does not thus render them ethics washing; it could reflect the need to transform the guidelines into concrete contexts of practices. Abstract guidelines are ethics washing only if they fail to include ways to track, such as transparency about outcomes of evaluations and actions taken on the basis of the guidelines. Each society needs to develop its own enforcement rules as there are diverse ways to carry out universal values. Hence, it is unclear that the above criticism is sufficient to reject PPAs. Instead, improving the deployment and design of PPAs could be a more important way to resolve the problem.

Second, let us examine the criticism against the tools per se. Although PPAs are regarded as ineffective by some critics, their evaluation involves complicated factors, such as tasks (e.g., community or border security), types (e.g., person-based or area-based), technologies (e.g., with or without facial recognition), goals (risk assessment or reduction), social opinion towards the police, and the business interests of the service provider. These factors vary across and among countries. For example, a primary focus of risk assessment (RA) is algorithmic prediction accuracy, which can

be improved by data scientists. Conversely, risk reduction (RR) focuses on whether the crime rate is decreased by a police department equipped with PPAs, whose advancement depends on the department's efficiency and its integration with novel technology. However, RA and RR can be dissociative. A highly accurate PPA may be exploited bureaucratically and perform poorly in RR. Therefore, citing the mere fact of the limited decline in the crime rate to reject PPAs could omit confounders and cause fallacy. Thus, more detailed comparative studies of PPA employment are necessary.⁶

Similarly, the claim that PPAs should be dismantled unless the bias problems are fixed may warrant a second thought. Bias results from the interplay of the social environment, cognitive limitations, and human thoughts (Haslanger, 2015, 2017; Zheng, 2018). Humans evolve with bias because it beneficially reduces the cognitive processing cost by simplifying the world and offering easy solutions (e.g., outgroup alert) to survive challenges (e.g., avoid plunder). The brain sorts people according to explicit characteristics, such as ethnicity, sex, and language. This social classification, albeit reducing the cognitive workload (e.g., memory) and facilitating generalization, is the source of stereotypes and discrimination. The literature on heuristics and bias has already shown common deficits in this fast but inaccurate processing (Gilovich et al., 2002). Thus, we need to acknowledge that bias is part of human nature and human bias is a significant cause of data bias. According to “ought implies can,” critics’ moral requirement to dismantle PPAs unless the bias problems are fixed is rather strong because some cognitive biases are simply not eliminable. Besides, while introducing AI helps overcome some of the human biases and stereotypes, the need for normative bias—preferring an ideology to another—in deciding which fairness measure to use remains unavoidable. A person must admit that there are objective standards determining that certain biases are better than others (e.g., egalitarianism versus enslavement) to avoid the bias paradox; for example, if a feminist denying all objective standards (as set by men only) will fall into subjectivism and thus lose a standpoint to criticize the biased values of males (Engqvist, 2020). Hence, claiming that PPAs should be dismantled unless the bias problems are fixed needs a second thought.⁷ A more fundamental solution is to choose a relatively good

⁶ As one of our anonymous reviewers pointed out, a lack of proof of the lack of usefulness of PPAs is not required to stop supporting it. The burden of argument should be on those who want to support PPAs. It is a fair point. We argue that in order to measure the effectiveness of PPAs, PPAs must be conceived in the broader context of law enforcement operations. From this perspective, the usefulness of PPAs is to be integrated into broader systems. For example, in Chicago, predictive algorithms are part of the police department's strategic decision support centers (SDSCs). We can take the following passage from a recent report of SDSCs as an upbeat assessment of the deployment of PPAs (Hollywood et al., 2019, p. 70):

As a result, policing decisions can be made with a much higher level of quality—timelier, more complete, and more accurate—than was typically possible before.... More broadly, we see SDSCs as a promising model for improving law enforcement agencies’ awareness of their communities, improving their decisionmaking, and carrying out more effective and more efficient operations that lead to crime reductions and other policing benefits.

⁷ Besides, police departments adopting the technologies must acknowledge these tools’ vulnerabilities and the following limitations of the conclusions drawn to make room for auditing and improving these technologies’ performance. For example, we may establish auditing mechanisms to check the quality of inputs from the algorithms. While predictive policing programs are not completely bias-free, it is not a sufficient reason to dismantle PPAs.

bias in both machine learning (Berthold, 2020) and norms in a society, as well as do our best to change the unjust structures breeding biased data in that society.

Third, regarding the criticism against possible remedies, data correction does not imply discrimination. Whereas some computer scientists show that bias prediction is reducible by taking race explicitly into account (Goel et al., 2018), others hold that this remedy may be racist, as it applies different standards to ethnic groups (Heaven, 2020). At first glance, this remedy seems controversial because the same evaluation processes should be universally applied to all social members; otherwise, discrimination may occur. It is thus important to apply the same standard, as well as to offer the same support, to everyone. Nonetheless, people are inherently unequal (e.g., health and social classes), and the same standard and support could widen this inequality. Hence, it is also crucial to apply different standards and offer different supports to empower the disadvantaged to have equal access to social resources. These two considerations reflect different but valuable traditions in ethics. Therefore, treating ethnic groups differently does not necessarily imply discrimination; it could be moral compensation to fix the groups' unjust situation.

Moreover, we need to acknowledge that there are diverse ways of realizing fairness in law and policymaking. Fairness is a universal value having a common psychological base in human beings (Sloane et al., 2012), but this fact does not mean that there is only one way to implement fairness in policy. Societies can translate the policy goal of algorithmic fairness into actions in different ways. Thus, encouraging each society, via the process of deliberative democracy, to have its own way of carrying out the universal value is important.

In summary, as we have seen, while some criticisms are not as tenable as at first glance, others may pose challenges to PPA development and employment (i.e., distrust, efficiency, racism, and social equity).⁸ In the next section, we will focus on the PPA's additional problems faced by most, if not all, AI applications (e.g., explanation, transparency, and communication). We then offer a solution to address these challenges in the “[Policy Schema of Social Safety Net](#)” section.

Explainability, Accountability, and Communication

Explainability and transparency are often discussed as technical challenges in designing PPAs. Many worry about the lack of interpretability and transparency around how PPAs work, such as the data they collect, how they analyze the data, and so on. It is argued that without proper comprehension of the processing of data implemented in the system, it is difficult for us to control, monitor, and correct the system's performance. It is thus challenging to assign responsibility to any harm caused by the deployment of PPAs (Castelvecchi, 2020).

⁸ There are other challenges, such as *privacy*, that we have not discussed in this article but have elsewhere. For example, a dilemma is that if the algorithmic prediction is accurate, it must be trained on a massive amount of biometric data that risk privacy, but if it is not accurate, the false positives threaten the human rights of misidentified targets. Please see Hung (2020), Hung and Yen (2020), and Lin et al. (2020).

The reality, however, is that some of the most severe challenges are regarding the accountability of organizations using the algorithms in their decision-making (Coyle and Weller, 2020; Ferguson, 2017). As noted in the “[Criticisms and Analysis](#)” section, PPAs are products of their socio-technical context that raise concerns about both code and people. We must not think of PPAs out of the context of human power dynamics. The focus of PPAs is generally on the environment rather than the “root cause” of criminal events. Accordingly, crime prevention and control strategies are to manipulate environmental factors to increase the risk of crime and to reduce its perceived rewards. It is similar to when airports install metal detectors to prevent hijacking and libraries installing electronic access control inserts to make it more challenging to steal books. It is crucial to apprehend that PPAs are part of a broader goal to reduce environmental vulnerabilities encouraging crime. PPAs are tools to identify patterns and common traits among potential criminals so that support or guidance to prevent crimes can be provided beforehand. They must be technically and operationally integrated into police department operations, and their advancement depends on the efficiency of the policing practices.

PPAs are built and operated for a specific objective, subject to choices and assumptions, and must be understood as such. From this perspective, we do not need to know exactly how the PPA processes data. A PPA can be held accountable as long as we can test whether it works as designed (Kroll et al., 2017; Morley et al., 2019).⁹ An overemphasis on the transparency of the algorithms themselves is misplaced. While algorithms may be opaque in some instances, they can always be understood in terms of the goals they are designed for and the mechanism of their construction and operation. They can also be understood in terms of their inputs and outputs and the result from their application in given contexts (Kroll, 2018). What truly matters for transparency in explanation is why the algorithm is used and whether it serves its objectives and how. We should be transparent and open to community involvement when making policy and process decisions about these discussions. As a tool of predictive policing, the value of PPAs is “in their ability to provide situational awareness of crime risks and the information needed to act on those risks and preempt crime” (Perry et al., 2013, p. xxi). The relationship between the risk factors calculated by an algorithm and the result of the analysis of those factors is statistical,

⁹ Here is an example. The Crime and Victimization Risk Model (CVRM) was a statistical model used by the Chicago Police Department. It used arrest and crime incident data from within the Chicago Police Department’s record management systems to estimate an individual’s risk for becoming a party to violence. As shown on the department’s Violence Reduction Strategy web page (<https://home.chicagopolice.org/information/violence-reduction-strategy-vrs/>), the CVRM was “for the sole purpose of finding just the small group that may be at highest risk, so that the details of their crime records can be studied by experts for purposes of prioritizing the Custom Notifications program.” We do not need to know the complete details of how the algorithm works as long as we can decide whether it fulfilled its assigned purpose. 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). As reported on the department’s Violence Reduction Strategy web page (<https://home.chicagopolice.org/information/violence-reduction-strategy-vrs/>), that was a reasonably effective piece of information “to help to prioritize the Custom Notifications process,” given that “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].”

not causal. For a policing strategy to be considered effective, its results need to be tangible, such as decreased crime rates or increased arrest rates of severe offenses. Our evaluation of the results should identify confounders to avoid fallacy (as mentioned in “[Criticisms and analysis](#)” section). PPAs are to support law enforcement operations. To measure the effectiveness of PPAs, PPAs must be conceived in the broader context of law enforcement operations. It is important that police agencies understand how they work and know how to use them. PPAs will make little difference even if they are reasonably effective statistically when the police agencies do not understand the information provided. The use of PPAs should be regulated with rules to inform and structure decision-making. Moreover, the process for developing those rules should be articulated. Trade-offs of different choices will be unavoidable (Coyle and Weller, 2020; Morley et al., 2019) but expected to be acknowledged and evaluated transparently.

Another crucial but easily overlooked element of the demand for the explainability of PPAs and other machine learning algorithms is for understanding these associated assumptions, choices, and adequacy determinations (Kroll, 2018). The importance of the point is twofold. On the one hand, it is the public’s right to demand information relating to the algorithmic systems’ technical processes and the related human decisions, including, for example, information about how the police use the system to make operational decisions and policies regarding the retention, sharing, and use of the collected data (Hung and Yen, 2020). On the other hand, it will undermine public acceptance of PPAs when the above right to know is unfulfilled (Morley et al., 2019). The deployment of PPAs without community trust and involvement is neither feasible nor sustainable (Pearsall, 2010). Communication is the key to earning public trust.¹⁰

Policy Schema of Social Safety Net

In the previous sections, we examined some challenges of PPAs, including *distrust*, *efficiency*, *racism*, and *social equity*. In this section, we propose a policy schema to handle them. This schema is inspired by the similarity between public health and predictive policing. Based on the assumptions that physical and social environments may encourage predictable acts of criminal wrongdoing and that interfering with that environment would deter would-be crimes, predictive policing is to identify the where, when, and who of crime for police intervention. Accordingly, we propose a policy schema of the social safety net for predictive policing. The schema contains (1) predicting immediate risks and taking action; (2) detecting socially vulnerable

¹⁰ Take CVRM as an example again. In a 2019 review, the RAND Corporation found that the Chicago Police Department initially was not fully transparent about what was done by CVRM and that “left a great deal of room for concerns to grow and spread unchecked” (Hollywood et al., 2019, p. 38). CVRM is a party-to-violence prediction system by design. In practice, it becomes a victim prediction system because the clearance rates for shootings in Chicago were constantly low. The Chicago Police Department’s lack of communicative transparency made it mistakenly conceived as a criminal prediction system. As a result, individuals noted by the system face unnecessary stigmatization, and the system lost social acceptability.

individuals and offering help; and (3) being reviewed by and communicating with the public.

The first part is to predict immediate criminal risks and take action. The success of crime prevention does not lie in the predictions made by the algorithm but in the actions taken after the predictive outputs (Couchman, 2019; Saunders et al., 2016). As part of the advice for crime fighters, a reference guide conducted by the RAND Corporation noted that “[g]enerating predictions is just half of the predictive policing business process; taking actions to interdict crimes is the other half” (Perry et al., 2013, p. xxii). Another important point is that every specific intervention will vary by objective and situation. In addition, different PPAs produce different outputs and thus suggest different follow-up actions (Jenkins & Purves, 2020). For example, the interventions for individuals at risk of driving violence will be different from the interventions for those at risk of being victimized or the interventions for those at risk of domestic violence.¹¹ Whereas the information from police records management systems can provide some insights to identify needed interventions for the targeted group, it would require information from other government departments as well to better understand what service and other interventions an individual might need and thus require partnerships across government departments (Hollywood et al., 2019).

Second, it is crucial to identify unequal social structure breeding crimes and provide help to people in need. We want to bring attention to the potential of algorithms in the broad social safety net. As criminal involvement often links up with people with social-economic disadvantages, it is crucial in crime-fighting operations to offer help, such as job training, education, job placement, and health services, to improve stakeholders’ social welfare through a social safety net (Hung & Yen, 2020)¹² By better integrating PPAs into a broader governance framework of the social safety net, the outputs of PPAs may reflect social inequality. They can also be used to design specific intervention programs that preempt crime and improve the lives of vulnerable families and individuals in the community. The police department is only a segment within the social network.¹³ 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, additions, social work, and mental health, for example) and law enforcement agencies “in order to monitor and flag individuals or communities that are considered to be marginalized and at risk” (Roberson et al., 2020, p. 52). The aim is to “formulate

¹¹ The subgroup identified by CVRM, for example, was individuals at risk of being victimized. For the program to succeed, the followed-up interventions must be guided towards reducing the likelihood for the subgroup to be victims of violence in the future. Without sufficient information to identify needed interventions for the targeted group, the program’s chance to succeed was slim.

¹² According to the New Orleans Police Department (2011–2014), when the high-risk subgroups in the community are provided with the resources to improve, for example, their job prospects, there is indeed a significant reduction in homicide and gang-involved murders (Ferguson, 2017). The statics shows a significant difference in whether resources are implemented to increase the targeted individuals’ opportunities and chances to escape crime.

¹³ For the situation in the UK, see Babuta and Oswald (2020) and Crawford and Evans (2012).

a plan of intervention that mobilizes multiple sectors, collaborating to provide services and support” to the targeted individuals or communities.^{14,15} The intervention programs aim to reduce crime by eliminating the possibility of someone responding that they did not know they were committing an offense or had no alternative but to commit a crime. For these reasons, the programs employing predictive policing aim to alter certain aspects of the physical and social environment to discourage predictable acts of criminal incidents.¹⁶ When properly used, PPAs may reflect social inequality, which is partly due to the unequal distribution of social resources, such as opportunities and wealth. By better integrating PPAs into a broader governance framework of the social safety net, PPAs help distribute social services to some of the most vulnerable and most needed subgroups in society.

Third, the employment of PPAs should be noted and reviewed by the public because a large part of public fear and distrust of PPAs is due to poor communication between the police and the community. AI is a powerful tool, and its usage should be regularly checked and revised to avoid abuse. Continuous measurements are vital to identifying areas for improvement, modifying interventions, and distributing resources. The use of PPAs should face public audits and be overseen through democratic procedures, including parliament and civic society. If something goes wrong, the legal system should be able to hold someone accountable and avoid repeating the same mistake. 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, policymakers, citizens, and developers and designers in the endeavor.

In summary, the progress of human civilization has benefited from the invention of various tools, but humans have also tended to destroy each other with these tools historically. AI could be a potent tool. Designing a safety net for this tool or establishing usage guidelines helps reduce abuse. However, more importantly, it is

¹⁴ As Roberson et al. (2020, p. 55) put it, these programs “offer a venue for service providers from various sectors (police, education, addictions, social work, mental health, etc.) to regularly convene and discuss clients who meet a defined threshold of risk. The intent of these discussions is to formulate a plan of intervention that mobilizes multiple sectors, collaborating to provide services and support to the individual or families. To mitigate risk before harm occurs, [they] aim to connect clients to services within 24 to 48 h of a case being presented to the group.”

¹⁵ The programs “[hold] that violent crime can be dramatically reduced when law enforcement, community members, and social services providers join together to directly engage with street groups and gangs to clearly communicate: (1) a law enforcement message that any future violence will be met with clear, predictable, and certain consequences; (2) a moral message against violence by community representatives; and (3) an offer of help for those who want it” (von Ulmenstein and Sultan, 2011, p. 7). For details, see Kennedy and Friedrich (2014).

¹⁶ For example, as part of the efforts of the Office of Community Oriented Policing Services of the U.S. Department of Justice, law enforcement is encouraged to practice community policing by working with the communities they serve. It is noted that the role of law enforcement in the group violence intervention program is to identify the high-risk groups of exposure to violence, either as victims or as perpetrators, and to notify the targeted individuals who “are subjects of special law enforcement attention” (Kennedy & Friedrich, 2014, p. 26). The notification usually includes a custom legal assessment explaining the target’s legal exposure and information on social service resources available for the target and his/her families. Also see Braga et al. (2018) for a review of recent research on the effectiveness of this approach. They note that the existing theoretical literature and empirical evidence suggest that this approach generates noteworthy crime reductions.

necessary to increase the incentives or rewards of human reciprocal altruism and fair cooperation by changing social institutions. This ensures that humans can benefit from the tool while reducing abuse.

Conclusions and Further Questions

This article aims to contribute to the debate by analyzing key criticisms and challenges to the employment of PPAs and offering a policy schema to handle them. We argue that, with appropriate management and human oversight, predictive policing algorithms can help achieve social goods.

To summarize, the “[Criticisms and Analysis](#)” section clarifies multiple factors involved in the complexity of the debate by classifying and examining the main objections to the employment of PPAs. It argues that some issues (e.g., distrust, bias prediction, transparency, and social inequality) are genuine challenges, while others are more rhetorical. The “[Explainability, Accountability, and Communication](#)” section discusses further concerns regarding accountability. It is noted that we should conceive the PPAs in the broader context of law enforcement operations. More attention should be given to explain why the algorithm is being used and what mechanisms exist to hold the creators and the operators accountable. The “[Policy Schema of Social Safety Net](#)” section proposes a schema of the social safety net for predictive policing to address the challenges. This article concludes with the view that banning PPAs helps little to solve the problems. Similar to other means adopted in the progress of human civilization (e.g., knives and fire), PPAs are instruments. Whether it is favorable or dangerous depends on how people use them (Kleinberg et al., 2018). We believe that better integrating PPAs into a broader governance framework of the social safety net will lead to a positive impact and help reduce racism and achieve equity.

One may wonder whether the proposed solution can be applied to detect vulnerable groups and help improve social resilience—a society’s capacities to cope with or respond to external natural and social disturbances—in a society. Our preliminary view is positive. For example, recent studies have found that natural hazards do not cause harm to social members equally. Economically disadvantaged groups are more likely to be impacted when facing natural disasters (Lin, 2015; Lin et al., 2017). Thus, by analyzing potentially vulnerable groups and offering more resources for preventive countermeasures in advance, we can reduce these groups’ vulnerability. Furthermore, systematic discrimination is a common unjust structure in society. The structure may be implicit and not easy to identify. Through AI, it may be possible to find the patterns of these revelations and systemic biases and then find ways to improve them. Of course, those issues are far beyond the scope of this article; they nevertheless constitute a valuable topic for future research.

Funding Funding was provided by Ministry of Science and Technology, Taiwan (Grant No. 107-2410-H-001-101-MY3).

References

- Amnesty International. (2020). *We sense trouble: Automated discrimination and mass surveillance in predictive policing in the Netherlands*. Retrieved November 11, 2020, from <https://www.amnesty.org/download/Documents/EUR3529712020ENGLISH.PDF>
- Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *Machine bias*. ProPublica. Retrieved November 11, 2020, from <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- Aougab, T., Ardila, F., Athreya, J., Goins, E., Hoffman, 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.
- Asher-Schapiro, A. (2020, June 25). *California City bans predictive policing in U.S. First*. Reuters. Retrieved November 11, 2020, from <https://www.reuters.com/article/us-usa-police-tech-trfn/california-city-bans-predictive-policing-in-u-s-first-idUSKBN23V2XC>
- 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 November 20, 2020, from https://rusi.org/sites/default/files/rusi_pub_165_2020_01_algorithmic_policing_babuta_final_web_copy.pdf
- Berthold, M. (2020, November 14). *You can't eliminate bias from machine learning, but you can pick your bias*. VentureBeat. Retrieved November 20, 2020, from <https://venturebeat.com/2020/11/14/you-cant-eliminate-bias-from-machine-learning-but-you-can-pick-your-bias/>
- Braga, A. A., Weisburd, D., & Turchan, B. (2018). Focused deterrence strategies and crime control: An updated systematic review and meta-analysis of the empirical evidence. *Criminology & Public Policy*, 17(1), 205–250.
- Castelvecchi, D. (2020). *Mathematicians urge colleagues to boycott police work in wake of killings*. Retrieved November 11, 2020, from <https://www.nature.com/articles/d41586-020-01874-9>
- Couchman, H. (2019). *Policing by machine*. Liberty, LIB11. Retrieved November 20, 2020, from <https://www.libertyhumanrights.org.uk/wp-content/uploads/2020/02/LIB-11-Predictive-Policing-Report-WEB.pdf>
- Coyle, D., & Weller, A. (2020). 'Explaining' machine learning reveals policy challenges. *Science*, 368(6498), 1433–1434.
- 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.). Oxford University Press.
- Deto, R. (2020, August 25). *Pittsburgh city council introduces police facial recognition, predictive policing ban*. Pittsburgh City Paper. Retrieved November 11, 2020, from <https://www.pghcitypaper.com/pittsburgh/pittsburgh-city-council-introduces-police-facial-recognition-predictive-policing-ban/Content?oid=17879052>
- Egbert, S., & Krasmann, S. (2020). Predictive policing: Not yet, but soon preemptive? *Policing and Society*, 8(3), 1–15.
- Engqvist, T. (2020). The bias paradox: Are standpoint epistemologies self-contradictory? *Episteme*. <https://doi.org/10.1017/epi.2020.21>
- Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., & Venkatasubramanian, S. (2018). Runaway feedback loops in predictive policing. *Proceedings of Machine Learning Research*, 81, 1–12.
- Fajnzylber, P., Lederman, D., & Loayza, N. (2002). Inequality and violent crime. *The Journal of Law and Economics*, 45(1), 1–39.
- Ferguson, A. G. (2017). *The rise of big data policing: Surveillance, race, and the future of law enforcement*. New York University Press.
- Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge University Press.
- Goel, S., Shroff, R., Skeem, J. L., & Slobogin, C. (2018). *The accuracy, equity, and jurisprudence of criminal risk assessment*. Retrieved November 11, 2020. <https://doi.org/10.2139/ssrn.3306723>
- Hao, K. (2019, December 27). In 2020, let's stop AI ethics-washing and actually do something. *MIT Technology Review*. Retrieved November 11, 2020, from <https://www.technologyreview.com/2019/12/27/57/ai-ethics-washing-time-to-act/>
- Hardyns, W., & Rummens, A. (2018). Predictive policing as a new tool for law enforcement? Recent developments and challenges. *European Journal on Criminal Policy and Research*, 24, 201–218.

- Haslanger, S. (2015). Distinguished lecture: Social structure, narrative and explanation. *Canadian Journal of Philosophy*, 45(1), 1–15.
- Haslanger, S. (2017). Racism, ideology, and social movements. *Res Philosophica*, 94(1), 1–22.
- Heaven, W. D. (2020, July 17). Predictive policing algorithms are racist. They need to be dismantled. *MIT Technology Review*. Retrieved November 11, 2020, from <https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/>
- Hollywood, J. S., Mckay, K. N., Woods, D., & Agniel, D. (2019). *Real-time crime centers in Chicago*. Retrieved November 11, 2020, from https://www.rand.org/content/dam/rand/pubs/research_reports/RR3200/RR3242/RAND_RR3242.pdf
- Hosein, A. O. (2018). Racial profiling and a reasonable sense of inferior political status. *Journal of Political Philosophy*, 26(3), e1–e20.
- Hung, T.-W., & Yen, C.-P. (2020). On the person-based predictive policing of AI. *Ethics and Information Technology*. <https://doi.org/10.1007/s10676-020-09539-x>
- Hung, T. W. (2020). A Preliminary Study of Normative Issues of AI Prediction. *EurAmerica*, 50(2), 229–252. [https://doi.org/10.7015/JEAS.202006_50\(2\).0004](https://doi.org/10.7015/JEAS.202006_50(2).0004)
- Hitachi Inc. (2019). Hitachi provides an AI environment in research on Kanagawa prefecture police's crime and traffic accident prediction techniques. <https://reurl.cc/IL6d2E>. Accessed 16 Jan 2020.
- Illinois Institute of Technology. (2019). Crime and victimization risk model (CVRM) fact sheet. Retrieved November 11, 2020, from <https://home.chicagopolice.org/wp-content/uploads/2019/01/FACT-SHEET-Crime-and-Victimization-Risk-Model-1.pdf>
- Jenkins, R., & Purves, D. (2020). *AI ethics and predictive policing: A roadmap for research*. Retrieved November 11, 2020, from <http://aipolicing.org/year-1-report.pdf>
- Kennedy, D. M., & Friedrich, M. (2014). *Custom notifications: Individualized communication in the group violence intervention*. Office of Community Oriented Policing Services. Retrieved November 20, 2020, from https://nnscommunities.org/wp-content/uploads/2017/10/GVI_Custom_Notifications_Guide.pdf
- Kerr, I., & Earle, J. (2013). Prediction, preemption, presumption. *Stanford Law Review Online*, 65(September), 65–72.
- Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2018). Discrimination in the age of algorithms. *Journal of Legal Analysis*, 10, 113–174.
- Kroll, J. A. (2018). The fallacy of inscrutability. *Philosophical Transactions of the Royal Society A*, 376, 20180084.
- Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Harlan, Yu. (2017). Accountable algorithms. *University of Pennsylvania Law Review*, 165, 633–705.
- Kusner, M. J., & Loftus, J. R. (2020). The long road to fairer algorithms. *Nature*, 578(7793), 34–36.
- Lin, K.-H., Lee, H.-C., & Lin, T.-H. (2017). How does resilience matter? An empirical verification of the relationships between resilience and vulnerability. *Natural Hazards*, 88(2), 1229–1250.
- Lin, T.-H. (2015). Governing natural disasters: State capacity, democracy, and human vulnerability. *Social Forces*, 93(3), 1267–1300.
- Lin, Y. T., Hung, T. W., & Huang, L. T. L. (2020). Engineering equity: How AI can help reduce the harm of implicit bias. *Philosophy & Technology*. <https://doi.org/10.1007/s13347-020-00406-7>
- 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.
- Ohyama, T., & Amemiya, M. (2018). Applying crime prediction techniques to Japan: A comparison between risk terrain modeling and other methods. *European Journal on Criminal Policy and Research*, 24(4), 469–487.
- Pearsall, B. (2010). Predictive policing: The future of law enforcement? *NIJ Journal*, 266, 16–19.
- Perry, W. L., McInnis, B., Price, C. C., Smith, S. C., & Hollywood, J. S. (2013). *The role of crime forecasting in law enforcement operations*. RAND Corporation. Retrieved November 11, 2020, from https://www.rand.org/content/dam/rand/pubs/research_reports/RR200/RR233/RAND_RR233.pdf
- 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*, 94, 192–233.
- Roberson, K., Khoo, C., & Song, Y. (2020). *To surveil and predict: A human rights analysis of algorithmic policing in Canada*. Retrieved November 11, 2020, from <https://citizenlab.ca/wp-content/uploads/2020/08/To-Surveil-and-Predict.pdf>

- Saunders, J., Hunt, P., & Hollywood, J. S. (2016). Predictions put into practice: A quasi-experimental evaluation of Chicago's predictive policing pilot. *Journal of Experimental Criminology*, 12(3), 347–371.
- Schuilenburg, M. (2021). *Hysteria: Crime, media, and politics*. Routledge.
- Selbst, A. D. (2017). Disparate impact in big data policing. *Georgia Law Review*, 51(1), 109–195.
- Shapiro, A. (2017). Reform predictive policing. *Nature News*, 541(7638), 458.
- Sheehey, B. (2019). Algorithmic paranoia: The temporal governmentality of predictive policing. *Ethics and Information Technology*, 21, 49–58.
- Skeem, J., & Lowenkamp, C. (2016). Risk, race, and recidivism: Predictive bias and disparate impact. *Criminology*, 54, 680–712.
- Skeem, J., & Lowenkamp, C. (2020). Using algorithms to address trade-offs inherent in predicting recidivism. *Behavioral Sciences & the Law*, 38, 259–278.
- Sloane, S., Baillargeon, R., & Premack, D. (2012). Do Infants have a sense of fairness? *Psychological Science*, 23(2), 196–204.
- Susser, D. (2021). Predictive policing and the ethics of preemption. In B. Jones & E. Mendieta (Ed.), *The ethics of policing: An interdisciplinary perspective*. New York University Press.
- 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.
- 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 November 20, 2020, from https://nnscommunities.org/wp-content/uploads/2017/10/LE_Case_Studies.pdf
- Wisconsin v. Loomis. (2016). Retrieved November 11, 2020, from <https://www.wicourts.gov/sc/opinion/DisplayDocument.pdf?content=pdf&seqNo=171690>
- Zarsky, T. Z. (2013). Transparent predictions. *University of Illinois Law Review*, 2013(4), 1503–1570.
- Zheng, R. (2018). Bias, structure, and injustice: A reply to Haslanger. *Feminist Philosophy Quarterly*. <https://doi.org/10.5206/fpq/2018.1.4>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.