AI and Ethics
https://doi.org/10.1007/s43681-021-00117-5

ORIGINAL RESEARCH

AI audits for assessing design logics and building ethical systems: the case of predictive policing algorithms

Pamela Ugwudike[1]

Received: 30 May 2021 / Accepted: 28 October 2021

Abstract
Organisations, governments, institutions and others across several jurisdictions are using AI systems for a constellation of high-stakes decisions that pose implications for human rights and civil liberties. But a fast-growing multidisciplinary scholarship on AI bias is currently documenting problems such as the discriminatory labelling and surveillance of historically marginalised subgroups. One of the ways in which AI systems generate such downstream outcomes is through their inputs. This paper focuses on a specific input dynamic which is the theoretical foundation that informs the design, operation, and outputs of such systems. The paper uses the set of technologies known as predictive policing algorithms as a case example to illustrate how theoretical assumptions can pose adverse social consequences and should therefore be systematically evaluated during audits if the objective is to detect unknown risks, avoid AI harms, and build ethical systems. In its analysis of these issues, the paper adds a new dimension to the literature on AI ethics and audits by investigating algorithmic impact in the context of underpinning theory. In doing so, the paper provides insights that can usefully inform auditing policy and practice instituted by relevant stakeholders including the developers, vendors, and procurers of AI systems as well as independent auditors.

Keywords AI auditing · AI ethics · AI bias · Fair AI · Transparent AI · Predictive policing algorithms

1 Introduction

A fast-growing multidisciplinary literature is increasingly detailing the multifaceted biases associated with the data-driven artificial intelligence (AI)[1] systems that now inform decision making in many sectors including high stakes settings such as criminal justice systems [1–7]. AI audits have become central to efforts to mitigate unknown risks and avoid harms to create ethical systems [8] and this paper contends that a thorough evaluation of design logics or rationalities should form part of any audit, not least because the logics drive key aspects of AI design and operation and can ultimately influence outputs. To demonstrate how this occurs, this paper focuses on a key aspect of design logics which is the theoretical foundation of any AI system. In the context of AI design, I define theoretical foundations, assumptions or standpoints broadly to include either a formal theoretical framework or the creators' interpretation of the task, problem, or issue the system is designed to address, all of which inform key dimensions such as model architecture, data selection and processing, as well as the outputs.

Reinforcing the view that theoretical standpoints inform the design and outputs of systems such as predictive policing algorithms (PPAs), Kauffman and colleagues [9] rightly note in their analysis of the politics of algorithmic pattern detection for predictive policing that, 'prediction algorithms follow many styles of prediction, each driven by their own arguments and epistemic approaches to crime'. Importantly, the theoretical standpoint that drives AI processes and outputs (such as predictions) can ultimately pose social implications. This is because underpinning theoretical tenets are transmitted via developers' choices (e.g., data selection and processing decisions) into algorithmic outputs, and following deployment, into the social world. Thus, algorithmic

✉ Pamela Ugwudike
p.ugwudike@soton.ac.uk

[1] University of Southampton, Southampton, UK

[1] In this paper, the terms 'Artificial Intelligence (AI)' and 'algorithms' are used interchangeably to refer to data-driven computational models that seek to perform tasks typically associated with human beings. This reflects the common usage of both terms in recent times.

Published online: 13 December 2021

outputs, whether produced by programmed or learning models, reflect underpinning theoretical assumptions.

Whilst studies of AI bias rightly emphasise data-related problems such as the use of biased or unrepresentative data [1, 5, 6, 10], it is equally important for AI audits to investigate the capacity for theoretical assumptions to inject bias, with bias in this context referring to algorithmic outputs that consistently disadvantage subgroups along racial, gender, and other social lines.

AI bias can originate from underpinning theoretical assumptions that are typically invisible and not readily amenable to quantification and metrification but should be revealed by the creators and investigated during an audit if the aim is to create equitable models. On these bases, this paper's main purpose is to demonstrate why any audits of algorithmic systems should pay attention, not only to technical issues to do with data-related problems, but also to the links between design rationalities, in this case, theoretical assumptions, and broader social impact.

The paper's additional purpose is to use PPAs as a case example to demonstrate the importance of ensuring that audits consider the potential impact of a system's theoretical foundations. Focussing on the neo-classical, near repeat thesis, which is one of the theories that can underpin predictive policing models [11, 12], the paper shows how the theory influences the ways in which the PPAs it underpins, interpret the data fed into them, increasing the potential for adverse social outcomes such positive feedback loops that can prompt the excessive policing of historically over-policed communities. The paper also shows how this theory and its underlying assumptions about the aetiology of crime and effective crime control, can potentially foment the problem of crime displacement via the spatiotemporal dislocation of crime opportunities from predicted crime locations to nearby areas.

2 Unravelling AI design logics

All digital technologies are guided by specific ideologies, preferences, and other logics that infuse their design and tasks with meaning, giving rise to particular outputs and specific social implications. They are not created in an ideological vacuum: there is always a 'human in the loop' influenced by and influencing the social world in which digital technologies are designed and deployed, even if a digital model eventually appears to be fully automated. It is as such futile to imagine that 'the digital world is different and better than the social world' [13]. The digital is always inextricably linked to the social and one of the avenues through which the social permeates the digital is through the theory that influences how technology is designed and its operation.

The extant literature on AI audits emphasises the importance of auditing practices that can take such theoretical foundations into account by systematically assessing not only the technical (e.g., predictive accuracy and explainability) but also the non-technical (e.g., underpinning design logics and principles) [14, 15]. This body of work therefore recognises the importance of proactive auditing of AI systems and their design logics as proposed by this paper. For example, some have highlighted the utility of pre-emptive impact assessments which, *inter alia*, evaluate AI design logics including the underpinning theoretical assumptions. The aim is to anticipate potentially harmful outcomes including adverse social impact [16] such as negative discrimination [17].

Indeed, the fast-growing scholarship on AI audits provides valuable insights that should inform fair AI design. But the scholarship tends to focus primarily on private sector technologies and overlook the AI systems deployed by justice systems. With its focus on applications of AI by police services, this paper expands the existing literature. It also contends in line with others such as Kazim and colleagues (2021) that, for 'holistic' AI audits, attention should not be placed solely on technical components. Non-technical issues such as the underpinning theoretical assumptions should be assessed, preferably proactively, to understand their potential impact on outputs and the broader social consequences that could arise.

This is particularly important given the insights emerging from the fast-growing multidisciplinary scholarship on AI bias which point to instances where Machine Learning algorithms and other systems have produced inequitable social outcomes that disadvantage historically disadvantaged subgroups across employment, social security, and other sectors. This scholarship highlights the discriminatory impacts of techs such as affective computing models [18], employment selection algorithms [19]; popular search engines [20], social security payment allocation algorithms [21], social media platforms [22], University admissions algorithms [23], examination result algorithms [24], and healthcare provision algorithms [25].

Criminal justice algorithms are not immune, with the extant literature demonstrating that algorithms deployed in justice systems can disadvantage racialised and low-income groups historically vulnerable to criminal justice intervention [1–7]. Key technologies in this context include risk assessment algorithms [7]; facial recognition systems [10]; and PPAs [6].

On close inspection, key arguments within the multi-faceted scholarship on AI bias tend to coalescence around three sometimes intersecting themes; broad structural and systemic critiques emphasising racial, gender, and other biases that disadvantage already marginalised subgroups, legal-philosophical critiques highlighting human rights

violations including data privacy and due process rights, and technical critiques focusing on transparency, accountability, and explainability deficits. Scholars, civil society organisations and others apply these critiques to various AI systems including those listed above.

Whilst the existing studies and debates provide very useful findings, there seems to be an inordinate focus on data-related harms, but underpinning theoretical assumptions can also operate as conduits of bias and should be scrutinised via various measures including audits, to support the design of ethical AI.

Further, focusing on data-related issues can fuel the presumption that AI systems are neutral tools that simply generate outputs based on patterns data. This perspective encourages the view that the tools should not necessarily be blamed for harmful social outcomes that arise when they are deployed. But depicting the systems as technical products that simply perform tasks for which they were designed, anthropomorphizes them, portraying them as neutral tools that are completely independent of human or social influence. Such techno-determinism must be resisted, and it should be recognised that the models are in fact sociotechnical systems that mirror or reflect the social contexts in which they are designed including the developers' chosen theoretical assumptions. Benjamin [13] defines techno-determinism as, 'The mistaken view that society is affected by but does not affect technological development'. Raji and Smart et al. [8] take this further by noting in their analysis of global ethical issues and guidelines that, 'artificial intelligence systems are not independent of their developers or of the larger sociotechnical system'. A broader understanding of AI outputs as the products of, not only the underpinning data, but also the underpinning theory, is as such required and should inform any audit intending to uncover the potential or actual source/s of algorithmic harms.

3 Remedying AI harms: the rise of ethics and audits

In response to harms associated with algorithmic models, the fast-growing field of AI ethics has emerged [26], and developers as well as procurers are increasingly called upon to address the ethical implications of their products. Indeed, the field of AI ethics has emerged internationally as the core approach to remedying algorithmic bias [8, 27], and a host of institutions and authorities have proposed useful and even high-level principles in ethical standards for several public and private sectors. The standards are nevertheless typically self-regulation mechanisms that are unenforceable and susceptible to disparate interpretation in real world contexts of algorithm design and deployment. Recognising this, Mittelstadt [27] notes that, 'the AI industry lacks proven methods to translate principles into practice' (see also [28]. This gives the creators significant latitude, and in some cases, it is possible that a professed commitment to ethical standards simply provides a veneer of design integrity, and is in essence a 'smokescreen for carrying on with business as usual' [29].

Perhaps recognising the need to ensure that ethical principles are embedded in practice, calls for AI audits have become vociferous across several jurisdictions and sectors. As Brown and colleagues [30] observe in their analysis of the field of AI audits, 'nearly every research organisation that deals with the ethics of AI has called for ethical auditing of algorithms.' Internal and external/independent audits are increasingly being implemented as accountability and transparency measures with emphases on, *inter alia*, identifying conduits of algorithmic bias, embedding ethical principles in algorithmic design, and generally pre-empting and avoiding AI harms.

As defined by the IEEE [31], an audit constitutes 'an independent evaluation of conformance of software products and processes to applicable regulations, standards, guidelines, plans, specifications, and procedures'. Originating mainly in industry and similar to measures adopted in the fields of Finance and Information Security [32], AI audits have become an important dimension of the field of AI ethics, a multidisciplinary field which as noted earlier, seeks to address the adverse impacts of such systems [26].

A key aim of AI audits is to translate ethical principles into practicable mechanisms of equitable AI and some audits have indeed been successful in uncovering algorithmic harms [10] although wide-ranging knowledge of such impact is lacking. Nevertheless, the *Gender Shades project* which conducted an audit of three commercial facial recognition systems produced by IBM, Microsoft and Face + +, found 'substantial disparities' in error rates across gender and skin colour'. For example, darker-skinned females had the highest rate of misclassification—error rates of up to 34.7% [10].

Audits such as these are clearly needed, and it is also possible that well-implemented audits may provide quality assurance and enhance public trust via enhanced transparency and accountability. But this requires a recognition that AI systems are multidimensional systems, and that algorithm auditing should extend beyond technical matters such as data provenance, quality, and processing issues. Audits should also investigate and question typically less visible dimensions, such as the theoretical assumptions about human behaviour and the social world, that can infuse tech design with potential social implications.

4 The importance of evaluating underpinning theory/ies: a case example

In the remaining sections, I use the example of PPAs to demonstrate how the theoretical foundations of an AI system can inject bias into the system, highlighting why audits should consider this problem. PPAs are 'data-mining tools that [seek to] predict and pre-empt criminal activity' [33]. The techs process historical data in an effort to forecast crime across time and space, and are portrayed by their vendors as scientific, evidence-based systems that are capable of big data analytics geared towards identifying the spatiotemporal features of future crime with a high degree of accuracy. As the National Academies of Science, Engineering, Medicine (NASEM) [34] observe in their review of how proactive policing impacts on crime and communities, 'predictive policing is a relatively new strategy, and policing practices associated with it are vague and poorly defined'. But the commonly cited examples include PredPol (now rebranded as Geolitica) [35, 36] and HunchLab [37]. These and other PPAs have been used by some police services for location-based policing to target predicted areas for more intensive policing than others, a practice that is said to facilitate the efficient allocation of policing resources for cost-effective crime prevention [38].

Studies, however, suggest that PPAs can foment discrimination by generating predictions that encourage sustained police dispatch to areas heavily populated by racial minorities [6, 39]. PPAs can also pose the risk of crime displacement whereby crime moves from predicted crime areas subject to heavier police dispatch, to other less policed locations. Here, I demonstrate that these adverse social outcomes can be traced in part to the theories that underpin the PPA, focusing primarily on the impact of near repeat theory. This theory originated in the field of epidemiology [40] and has long featured in environmental criminology [41–44]. Asking questions about this and other underpinning theoretical assumptions during audits should reveal additional factors that influence the outputs and social outcomes of PPAs.

5 Near repeat theory and PPAs

As applied in environmental criminology, near repeat theory is based on the premise that a crime event will increase the risk of a nearby crime shortly after the initial event, if deterrent measures, in this case police presence, are not instituted. Predictive policing via pre-emptive policing dispatch to the crime risk areas is therefore recommended. Thus, the emphasis of predictive policing is on pre-emption not solely crime prevention. In an analysis of predictive policing technologies, Andrejevic [33] usefully distinguishes between the two, remarking that predictive policing is, 'not about prevention in the sense of transforming the conditions that contribute to theft or fighting; it is about being in the right place to stop an imminent act before it takes place'.

Near repeat theory is theoretically aligned with neoclassical crime theories of which rational choice and routine activities theories are key examples [45]. These theories hold that crime patterns[2] are not spatiotemporally random and can be studied for proactive policing. From this perspective, crime is the product of rational human calculations and is likely to occur if the conditions are propitious in the sense that the presumed benefits outweigh the risks (e.g., detection), as is the case where crime opportunities exist, and the chances of detection are too negligible to deter the potential offender. Deterrence theory is as such relevant here and it posits that a potential offender will be deterred from committing crime if the certainty, severity and celerity of punishment are high [46].

With routine activities theory which is another neo-classical theory of crime that is aligned with the near repeat theory underpinning predictive policing, location-based or situational factors or conditions can be instituted to deter the motivated offender: removing or blocking access to a suitable target and putting in place, a suitable guardian such as a foot patrol officer [45]. Therefore, from the perspective of neoclassical criminology, crime is the product or either careful planning or opportunism or both and can be studied to identify patterns that can inform targeted policing.

This is not to say that all predictive policing algorithms rely on an explicit crime theory, but regardless of whether or not a PPA has a stated or explicit theoretical basis, AI design is driven by the developer's belief concerning how the world is or should be, and their subjective conception of the problem or task the system is designed to address. These logics converge to infuse the design and operation of the system with meaning and should be evaluated during audits.

It also appears that PPAs typically rely on police recorded crimes as a data source amongst other sources[3] for information about crime events from which predictions about

[2] These theories typically focus on street crimes at the expense of less visible but nevertheless highly serious crimes such as domestic violence and some corporate crimes.

[3] Data-driven policing models do not rely solely on administrative data. Indeed, it has been argued that criminal justice algorithms draw on datasets culled from an array of sources [3] regardless of whether or not such data are linked to criminality [50]. These include 'big data' comprising linked biometric, health, demographic, geographical, and socioeconomic data, some of which can originate from human interaction with public services and private sector organisations. Corporations can also make available to the state, social media and smartphone data for surveillance purposes. In response to this, concerns have been raised about data ownership and infringements of privacy [2].

potential near repeats are then generated. This indicates that alongside the commitment to near repeat theory, another underpinning belief is that the well-documented systemic biases that can permeate police data [47, 48] are either non-existent or can be ignored. But studies suggest that such value-driven assumptions that police recorded crime data are useful for pre-emptive identification of near repeat patterns of crime and risks, and that police presence is required in predicted crime risk locations to prevent near repeat victimisation, can foment discriminatory policing [6, 39]. These and other studies of AI bias further demonstrate why theoretical foundations and assumptions should be carefully evaluated during audits.

6 The impact of theory on outcomes: PPAs and positive feedback loops

In this section, I draw on relevant studies to illustrate how near repeat assumptions can influence a PPA's interpretations of input data and foment discriminatory outcomes. One such PPA is an internationally used commercial algorithm that was developed by researchers in the US. The published version of the algorithm indicates that it is a seismology algorithm that utilises Epidemic Type Aftershock Sequence (ETAS) crime forecasting and Self-Exciting Point Processes (SEPPs) [35, 49]. In other words, it is a modified earthquake forecasting model that analyses time-series data to predict crime risk locations over time. The model, therefore, proceeds on the basis that an event can temporarily escalate the likelihood of a proximate and similar future event just as earthquakes generate aftershocks. This is clearly consistent with the near repeat thesis and whilst developer-led studies emphasise the PPA's efficacy in predicting and reducing crime [35], independent studies by Lum and Isaac [6] and Ensign et al. [39] indicate that the near repeat assumption evident in its SEPPs design and operation, can ultimately breed discriminatory policing.

The studies suggest that this adverse social outcome is likely when the SEPPs model relies on police crime data which can be imbued with racially biased policing whereby some minorities, particularly black people, are criminalised at a higher rate than others e.g., via 'over-searching' and 'over-patrolling' [48]. The affected subgroup would therefore be overrepresented in the data used by the model for near repeat analysis, fuelling predictions of spatiotemporal crime risks in the areas in which they reside, and encouraging police dispatch accordingly. Worse still, in their study, Lum and Isaac [6] also found evidence of a vicious cycle whereby police dispatch to predicted areas activated excessive policing, generating more distorted crime data which when fed back into the PPA, produced further biased predictions. In their analysis of the same PPA using the policing data from Lum and Isaac's [6] study, Ensign and colleagues [39] recorded similar findings.

These studies show that PPAs inspired by a near repeat/SEPPs design logic, can though their predictions encourage the concentration of policing in the same areas. Policing those areas yields datasets, which when fed back into the system, triggers a positive feedback loop. The studies therefore suggest that the PPAs can have self-reinforcing properties fuelled by their underpinning near repeat/SEPPs design. Systematics in the models generate feedback loops that repeatedly target the same areas even if in reality the patterns in data reflect an artificial inflation of crime rates created by repeated police dispatch to affected areas. This outcome occurs because, in line with near repeat assumptions, the model correlates police recorded crime (including racially biased records) with risks of near repeat victimisation, and generates predictions accordingly.

Unfortunately, the studies cited above suggest that some PPAs seem unable to unravel such nuances in datasets and mitigate the impact of systemic problems such as racially biased policing. Where an algorithm is driven by the near repeat thesis and the creators' assumption that police records equate to criminality and can be used to identify near repeat crime risks, the algorithm's data analysis and outputs will reflect these underpinning theoretical assumptions of how the social world is and should be, and would in fact, mask forms of discrimination embedded in the data. Therefore, unless steps are taken during design [51], the algorithms will automatically assume that criminalisation rates (e.g., arrest rates) are proxies for crime rates, ignoring other factors such as biased policing.

Some developers of PPAs recognise this problem and as has been noted, several claim that, to address the issue of potential over-policing in affected areas, their tools rely on crimes reported to the police, not the crimes the police observe following dispatch to predicted crime areas [52, 53]. But this approach can still introduce bias if other factors other than actual victimisation can explain higher crime reports in some areas, e.g., crime reporting propensities or rates that vary across racial, gender, socioeconomic, and other lines. Bias is also likely in areas that are historically prone to over policing, unless it can be shown that heightened police patrol in such areas does not influence crime reporting rates. These should all be evaluated during audits.

Meanwhile, this analysis of the links between near repeat assumptions and discriminatory outcomes provides a useful real-life example of how theoretical choice influences social outcomes. The example also shows that independent studies in this area seem more nuanced and able to shed light on the connections between *input* (e.g., theory + data processing instructions), model run (data processing) AND *output* (e.g., prediction).

7 The impact of theory on outcomes: PPAs and crime displacement effects

Another adverse outcome that can potentially stem from the deployment of PPAs based on the near repeat thesis is crime displacement whereby the PPAs trigger the dislocation of crime from predicted areas/police dispatch locations to other areas. This again exemplifies why underpinning theories and their potential impact should be investigated during any AI audit.

Crime displacement is the relocation or displacement of crime from one point in time and/or a location to another. It can occur as, (1) a temporal response that involves postponing the intended crime to a later time, (2) a spatial or geographical response that relocates the intended criminality to a different area, (3) a tactical response in which the intended crime is executed using a different method, (4) a target-related response whereby the intended crime is executed on a different target, (5) a crime type response in which a criminal act different from the intended crime is committed, and 6) an offender response that occurs when different offenders replace the potential offenders who would have been deterred by blocked situational opportunities such as police patrol [54–56]. There is also a conceptual distinction between benign and malign crime displacement. The former is said to occur when the volume of crime displaced is significantly lower and the types of crime are less harmful than the prevented crimes. On the other hand, malign displacement is the outcome when the crimes displaced raise crime rates in predicted low crime locations perhaps even to the levels in predicted high crime areas where policing resources are concentrated [56, 57].

It is worth noting that similar to near repeat theory, crime displacement theory is underpinned by the rational offender model of offending behaviour and the presumption that potential offenders are motivated by the presence of situational crime opportunities. Therefore, removing opportunities for crime, by for instance, increasing police presence and the chances of apprehension should deter the rational offender and improve crime prevention rates [45]. However, the perceptual character of deterrence means that other factors such as impulsiveness and lack of concern about possible apprehension can fuel offending behaviour regardless of the crime prevention strategies in place [46]. This means that again, the crimes that are relevant here are likely to be premediated street crimes. Whilst some scholars studying the environmental or ecological correlates of crime focus on sociological explanations to do with structural inequalities in their explanation of why crime rates vary across time and space [58] others emphasise environmental factors such as the absence of situational crime prevention strategies, for example hotspot policing [59]. It is, however, worth considering whether crime displacement, whatever its cause, is a potential risk of using PPAs that situate police patrols within a particular location as a crime prevention measure. This risk should be considered during audits of PPAs rooted in near repeat assumptions which as we have seen, can encourage such tactical strategies: specifically, the PPAs can prompt the sustained concentration of policing in the same areas.

Studies exploring whether or not crime displacement occurs following such heightened police presence in an area, emerge in part from the field of experimental criminology and from meta-analytic reviews [60, 61]. Most of the studies have focussed mainly on 'hotspot' policing inspired by the analogue clinical predictions of frontline police officers or digital crime mapping systems based on geographical information systems (GIS) technology. It is, however, worth noting that the difference between 'hotspot' policing and predictive policing driven by PPAs is not always clear-cut apart from the claim that PPAs can predict crimes [34]. Meanwhile, the prevailing design of crime displacement studies seems to predominantly involve comparisons of 'hotspot' policing areas with surrounding areas to observe changes in several crime-related variables such as: frequency of calls to police or police calls to people's address, and also rates of recorded crime. Increases in these variables within the 'hot spot' policing areas and decreases or stability in neighbouring locations (compared with baseline crime rates before and after the hot spot intervention) are considered indicators of limited or no crime displacement [60, 61]. Indeed, a reduction in crime rates within the surrounding areas is considered evidence that the intervention produced diffusion benefits. Such benefits it is claimed, arise when police concentration in an area acts as a deterrent, reducing in crime rates in the area and nearby locations.

Nevertheless, other studies have found evidence of crime displacement triggered by concentrated police presence. In their study of 'hot spot' policing and crime displacement, Andresen and Malleson [62] compared changes in crime rates within low crime areas and 'hot spot' locations following an initiative to introduce a heavy presence of foot patrol officers in the target areas. They found evidence of spatial crime displacement to border areas where police presence was not as concentrated, meaning that crime rates increased in those areas. Other studies have similarly found evidence of such displacement and some have shown that its occurrence can differ across time, space, and types of crime and it can even occur within the targeted locations [63]. This indicates that it can occur within small-sized geographical areas, posing implications for PPAs which tend to target locations at the granular level (e.g., street level).

These are at best mixed results from studies and debates about the links between heightened police presence and crime displacement, revealing the need to consider the

problem during audits of PPAs, since studies have shown that PPAs underpinned by near repeat logics can trigger concentrated policing in an area. Unfortunately, much of the existing studies focus on 'hotspot' policing, and independent research on PPAs specifically is lacking. However, some of the creators and proponents contend that the models do not trigger crime displacement. They argue that instead, PPAs increase the risks of perceived detection and produces a deterrent effect that reduces aggregate offending rates [64].

But it is clear that independent studies and audits are needed in this area. Issues that should be considered include, not only the spatial dimension of crime displacement, but also its temporal potential, that is, the possibility that crime displacement can occur in the long term. Another potential problem that should be considered is whether potential offenders relocate their activities distally to areas further away from a PPA prediction area but still within less policed catchment areas or grid cells that are susceptible to crime displacement. This is particularly problematic for PPAs since as already the technologies are designed to be applied within smaller geo-spatial regions for best outputs in terms of accuracy of prediction. Such deployment makes sense operationally since a key aim is to reduce overall crime rates by concentrating scarce resources in smaller-sized locations. But it means that the PPAs are limited by their inability to capture events beyond the granular level and spatiotemporal displacement could readily occur. Reinforcing this, Shapiro [53] remarks in his analysis of predictive policing and its social impact that, 'predictive policing is simply a more granular version of hot spot policing and as such has raised new concerns about crime displacement'. Shapiro goes on to imply that 'saturation and patrol predictability' are factors that can enhance potential offenders' awareness of heightened risks of apprehension, possibly encouraging them to relocate their activities to other less policed areas.

When the limitations of crime displacement studies are considered alongside the dearth of independent research on the potential for PPAs rooted in near repeat assumptions to trigger such adverse social outcome, it becomes clear that systematic analyses of the algorithms are needed and should be addressed via audits and independent research. This is particularly pressing given that the existing studies focus mainly on 'hotspot' policing informed by analogue processes such as the clinical predictions of frontline officers and analysts, or by geo-spatial crime mapping software and geographical information systems (GIS). Importantly, this discussion about the links between near repeat theory and crime displacement risks demonstrates the importance of unravelling and challenging underpinning theoretical foundations during audits.

8 Conclusion

This paper has shown that it is important for audits to evaluate how theoretical assumptions influence algorithm design and outputs. Using the case example of PPAs, and drawing on recent studies, the paper demonstrates that it is important to consider during audits that, (1) algorithmic feedback loops leading to the labelling and over policing of historically marginalised communities, and (2) the problem of crime displacement, are potential outcomes that can arise when a PPA is rooted in near repeat theory. In terms of (1) above, labelling certain areas as chronically criminogenic impacts on the agency of the residents who can become vicariously labelled. Other outcomes include adverse impacts on property prices and applications for mortgages and insurance. Regarding crime displacement, this is a problem that, if it occurs, relocates victimisation to other locations, rendering predictive policing counterproductive and as such ineffective.

Audits that seek to uncover the black box of AI design rationalities can help avoid these problems. Using the example of PPAs, some questions that auditors of such technologies could ask include: What is the developer's aetiological position on crime causation? What crime control approaches does it encourage? Which subgroups are most vulnerable to criminalisation when crime is defined this way and why? What type/s of policing or crime control approaches are encouraged by such definitions? What are the implications for diverse subgroups including those that are historically vulnerable to biased criminalisation? What are the goals for which the algorithm will be optimised (e.g., technical efficiency without consideration of social impact?). These questions could be tailored to suit the specifics of other AI systems.

Commercial algorithms of which the commonly used PPAs represent an example, are trade secrets protected by proprietary laws. Therefore, the creators are not required to release their code. Besides, the inner workings of black box systems particularly ML models can become opaque, making it difficult to anticipate, understand and mitigate risks. But an audit framework should investigate underpinning design logics as a key step towards demystifying the techs. This should contribute to efforts to ensure that the logics are open, transparent, and amenable to robust critique, to mitigate or prevent harmful social impacts.

Inviting the creators or developers to reflect on and answer questions such as those listed above (during the design process) should help ensure that underpinning assumptions are clearly stated and documented. It has been noted that 'With artificial intelligence systems it can be difficult to trace model output back to requirements because these may not be explicitly documented, and issues may

only become apparent once systems are released' [8]. This can prompt a reactive approach to algorithm auditing, but proactive measures are needed for harm avoidance and part of this should be an evaluation of underpinning theoretical assumptions.

We have seen how adopting a theory that defines crime events as criminogenic in the sense that they trigger additional crimes in the same areas leads to an emphasis on the use of predictive models that end up profiling and labelling the same areas with all the negative social connotations and implications of such labelling. The models also pose the risk of malign spatiotemporal crime displacement. Therefore, audits should consider whether PPAs anticipate these problems and have in place, remedial strategies.

As AI design in several Western countries e.g., the US and the UK is currently dominated by a specific demographic (mostly white males of high social status) [13] democratisation via design justice [65] involving the participation of affected groups in tech design could help rebalance current power dynamics that exclude already marginal groups from tech design processes such as selecting a guiding theory. Design justice can therefore help and diversify AI theoretical foundations. Broussard [66] argues that 'when development teams are small, like-minded, and not diverse', values that breed inequitable outcomes can become normalised. Perhaps this explains why the previously cited scholarship on AI bias suggests that AI benefits and risks are unevenly distributed, with racialised communities and socially marginal groups bearing much of the risks.

To conclude, whilst several existing studies of AI bias focus on data-related harms, this paper broadens the debate by considering theory-related issues and their potential downstream effects. Critical areas of future research include independent empirical studies of the impact of theory of AI operation, outputs, deployment and outcomes. The studies are very much needed to broaden current understandings and inform AI audits.

Declarations

Conflict of interest The author did not receive any funding or support from any organisation for the submitted work, there are no conflicts of interest, and the author wrote the entire paper. There are no datasets, codes, or ethical approval requirements for the paper and no participants were involved in the work.

References

1. Angwin J. Jeff Larson, J.: Bias in criminal risk scores Is mathematically inevitable, researchers say. https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say (2016). Accessed June 2018
2. Hannah-Moffat, K.: Algorithmic risk governance: big data analytics, race and information activism in criminal justice debates. Theor. Criminol. **23**(4), 453–470 (2018)
3. Green, B. Chen, Y.: Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In Proceedings of the Conference on Fairness, Accountability, and Transparency. ACM, 90–99 (2019)
4. Hao, K.: AI is sending people to jail—and getting it wrong. https://www.technologyreview.com/2019/01/21/137783/algorithms-criminal-justice-ai/ (2019). Accessed September 2019
5. Hao, K. Stray, J.: Can you make AI fairer than a judge? Play our courtroom algorithm game', MIT Technology Review. https://www.technologyreview.com/s/613508/ai-fairer-than-judge-criminal-risk-assessment-algorithm (2019). Accessed September 2019
6. Lum, K., Isaac, W.: To predict and serve? Significance **13**, 14–19 (2016)
7. Ugwudike, P.: Digital prediction technologies in the justice system: The implications of a 'race-neutral' agenda. Theoretical Criminology. Online First 2020. https://journals.sagepub.com/doi/abs/10.1177/1362480619896006 (2020)
8. Raji, I. D. Smart, A. et al.: Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. In Conference on Fairness, Accountability, and Transparency (FAT* '20), January 27–30, 2020, Barcelona, Spain. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3351095.3372873 (2020). Accessed November 2020
9. Kaufmann, M., Egbert, S., Leese, M.: Predictive policing and the politics of patterns. Br. J. Criminol. **59**, 674–692 (2019)
10. Buolamwini, J., Timnit, G.: Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. FAT (2018). Proceedings of Machine Learning Research. 81, 1–15 http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf (focuses on systems created by IBM, Microsoft Face++) (2018). Accessed July 2020
11. Haberman, C. P., Ratcliffe, J. H.: The predictive policing challenges of near repeat armed street robberies. Policing: A Journal of Policy and Practice 6(2), 151–166 (2012)
12. Perry, W.L.: Predictive policing: The role of crime forecasting in law enforcement operations. Rand Corporation, Santa Monica, CA (2013)
13. Benjamin, R.: Race After Technology: Abolitionist Tools for the New Jim Code. Polity Press, Cambridge (2019)
14. Kazim, E., Koshiyama, A.S., Hilliard, A., Polle, R.: Systematizing audit in algorithmic recruitment. J. Intelligence **9**(3), 46 (2021). https://doi.org/10.3390/jintelligence9030046
15. Koshiyama, A., Kazim, E., Treleaven, P., et al.: Towards algorithm auditing: A survey on managing legal, ethical and technological risks of AI, ML and associated algorithms. SSRN Electron J. (2021).
16. Kazim, E., Koshiyama, A.: The interrelation between data and AI ethics in the context of impact assessments. AI Ethics **1**, 219–225 (2021). https://doi.org/10.1007/s43681-020-00029-w

17. Raji, D. I., Smart, A, White, R. N. et al.: Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. In FAT* '20, 27–30 (2020)
18. Rhue, L.: Racial Influence on Automated Perceptions of Emotions (November 9, 2018). Available at SSRN: https://ssrn.com/abstract=3281765 or http://dx.doi.org/https://doi.org/10.2139/ssrn.3281765 (2019). Accessed August 2020
19. Ajunwa, I., Friedler, S.A., Scheidegger, C., Venkatasubramanian, S.: Hiring by algorithm: predicting and preventing disparate impact, Yale Law School Information Society Project Conference Unlocking the Black Box: The Promise and Limits of Algorithmic Accountability in the Professions http://sorelle.friedler.net/papers/SSRN-id2746078.pdf (2016). Accessed June 2018
20. Noble, S.: Algorithms of Oppression. New York University Press, New York (2018)
21. Eubanks, V.: Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. St Martins Press, New York (2018)
22. Ugwudike, P. and Fleming, J.: Artificial Intelligence, digital capital, and epistemic domination on Twitter: A study of families affected by imprisonment. Punishment and Society. Online First. https://doi.org/10.1177/14624745211014391 (2021)
23. Adams, R.: UCAS orders inquiry into 'racial profiling' of UK students. https://www.theguardian.com/education/2018/apr/24/ucas-orders-inquiry-into-racial-profiling-of-uk-students (2018). Accessed September 2018
24. Porter, J.: UK ditches exam results generated by biased algorithm after student protests. https://www.theverge.com/2020/8/17/21372045/uk-a-level-results-algorithm-biased-coronavirus-covid-19-pandemic-university-applications (2020). Accessed October 2020
25. Price, M.: Hospital 'risk scores' prioritize white patients. https://www.sciencemag.org/news/2019/10/hospital-risk-scores-prioritize-white-patients (2019). Accessed September 2018.
26. Kazim, E. and Koshiyama, A.: A High-Level Overview of AI Ethics. Available at SSRN: https://ssrn.com/abstract=3609292 or http://dx.doi.org/https://doi.org/10.2139/ssrn.3609292 (2020). Accessed October 2020
27. Mittelstadt, B.: AI Ethics: Too Principled to Fail? SSRN, https://doi.org/10.2139/ssrn.3391293 (2019). Accessed November 2019
28. Whittlestone, J., Nyrup, R., Alexandrova, A. Cave, S.: The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions. In Proceedings of the AAAI/ACM Conference on AI Ethics and Society, Honolulu, HI, USA. https://dl.acm.org/doi/https://doi.org/10.1145/3306618.3314289 (2019). Accessed May 2019
29. Sloane, M.: Inequality Is the Name of the Game: Thoughts on the Emerging Field of Technology, Ethics and Social Justice. In Proceedings of the Weizenbaum Conference 2019 "Challenges of Digital Inequality – Digital Education, Digital Work, Digital Life" (pp. 1–9). Berlin https://doi.org/10.34669/wi.cp/2.9 (2019). Accessed December 2019
30. Brown, S., Davidovic, J., Hasan, A.: The algorithm audit: scoring the algorithms that score us. Big Data Soc. (2021). https://doi.org/10.1177/2053951720983865
31. IEEE: IEEE Standard for Software Reviews and Audits. IEEE Std 1028–2008 1–53 (2008) https://doi.org/10.1109/IEEESTD.2008.4601584
32. Raji, I. D., Buolamwini, J.: Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In AAAI/ACM Conf. on AI Ethics and Society (2019). https://doi.org/10.1145/3306618.3314244
33. Andrejevic, M.: To pre-empt a thief. Int. J. Commun. **11**, 879–896 (2017)
34. NASEM: Proactive Policing: Effects on Crime and Communities. Washington, DC: The National Academies Press. https://doi.org/10.17226/24928. (2018). Accessed August 2018.
35. Mohler, G., O., et al.: Randomized controlled field trials of predictive policing. J. Am. Stat. Assoc 110 (512) 1399–1411 (2015)
36. PredPol: Geolitica: A New Name, A New Focus https://blog.predpol.com/geolitica-a-new-name-a-new-focus (2021a). Accessed May 2021
37. Heffner, J.: Algorithms & Explanation: A Humble Framing https://www.law.nyu.edu/sites/default/files/upload_documents/Jeremy%20Heffner%20Slides_0.pdf (2019). Accessed March 2020
38. PredPol: The Cost of PredPol & How to Justify Your Purchase https://blog.predpol.com/the-cost-of-predpol-how-to-justify-your-purchase (2021b). Accessed May 2021
39. Ensign, D., Friedler, S.A., Neville, S., Scheidegger, C. Venkatasubramanian, S.: Runaway feedback loops in predictive policing, paper presented at the Machine Learning Research. Conference on Fairness, Accountability, and Transparency, 81:1–12. https://www.fatml.org/schedule/2017/presentation/runaway-feedback-loops-predictive-policing (2018) Accessed September 2018.
40. Knox, E.G.: Epidemiology of childhood leukaemia in Northumberland and Durham. Br. J. Prev. Soc. Med. **18**, 17–24 (1964)
41. Haberman, C.P., Hatten, D., Crater, J.G., Piza, E.L.: The sensitivity of repeat and near repeat analysis to geocoding algorithms. J. Crim. Just. (2020). https://doi.org/10.1016/j.jcrimjus.2020.101721
42. Meijer, A., Wessels, M.: Predictive policing: review of benefits and drawbacks. Int. J. Public Adm. **42**(12), 1031–1039 (2019)
43. Johnson, S. D., Bernasco, W., Bowers, K. J., Elffers, H., Ratcliffe, J. H., Rengert, G. Townsley, M.: Space–time patterns of risk: a cross national assessment of residential burglary victimization.' J. Quant. Criminol. 23: 201, 219 (2007)
44. Ratcliffe, J.H.: The hotspot matrix: a framework for the spatio-temporal targeting of crime reduction. Police Pract. Res. **5**(1), 5–23 (2004)
45. Clarke, R. V. Felson, M.: Routine activity and rational choice: Volume 5 (Advances in Criminological Theory). Transaction Publishers (2004)
46. Nagin, D.: Deterrence in the twenty-first century. Crime Justice **42**(1), 199–263 (2013)
47. Shiner, M., Carre, Z., Delsol, R. Eastwood, N.: The Colour of Injustice: 'Race', drugs and law enforcement in England and Wales. https://www.lse.ac.uk/united-states/Assets/Documents/The-Colour-of-Injustice.pdf (2018). Accessed February 2019.
48. Vomfell, L., Stewart, N.: Officer bias, over-patrolling and ethnic disparities in stop and search. Nat. Hum. Behav. **5**, 566–575 (2021)
49. Mohler, G.O., Short, M.B., Brantingham, P.J., Schoenberg, F.P., Tita, G.E.: Self-exciting point process modelling of crime. J. Amer. Statist. Assoc. **106**(493), 100–108 (2011)
50. Kitchin, R.: Big data, new epistemologies and paradigm shifts. Big Data Soc. (2014). https://doi.org/10.1177/2053951714528481
51. Kamiran, F., Calders, T., Pechenizkiy, M.: Techniques for discrimination-free predictive models. In: Custers, B., Calders, T., Schermer, B., Zarsky, T. (eds.) Discrimination and privacy in the information society: data mining and profiling in large databases, pp. 223–240. Springer, Heidelberg (2013)
52. Ferguson, G.: Policing predictive policing, Washington University law. Review **94**(5), 1109–1189 (2017)
53. Shapiro, A.: Predictive policing for reform? Indeterminacy and intervention in big data policing. Surveill. Soc. **17**(3/4), 456–472 (2019)
54. Johnson, S.D., Guerette, R.T., Bowers, K.: Crime displacement: what we know, what we don't know, and what it means for crime reduction. J. Exp. Criminol. **10**(4), 549–571 (2014)
55. Reppetto, T.A.: Crime prevention and the displacement phenomenon. Crime Delinq. **22**, 166–177 (1976)
56. Barr, R., Pease, K.: Crime placement, displacement, and deflection. Crime Justice **12**, 277–318 (1990)

57. Guerette, R.T., Bowers, K.J.: Assessing the extent of crime displacement and diffusion of benefits: a review of situational crime prevention evaluations. Criminology **47**, 1331–1368 (2009)
58. Bottoms, A.: Developing Socio-Spatial Criminology. In: Maguire, M., Morgan, R., Reiner, R. (eds.) Oxford Handbook of Criminology, 5th edn. Oxford University Press, Oxford (2012)
59. Weisburd, D., Telep, C.: Hot spots policing: what we know and what we need to know. J. Exp. Criminol. **30**(2), 200–220 (2016)
60. Braga, A.A., Papachristos, A.V., Hureau, D.M.: The effects of hot spots policing on crime: an updated systematic review and meta-analysis. Justice Q **31**(4), 633–663 (2014)
61. Telep, C.W., Weisburd, D., Gill, C.E., Vitter, Z., Teichman, D.: Displacement of crime and diffusion of crime control benefits in large-scale geographic areas: a systematic review. J. Exp. Criminol. **10**, 515–548 (2014)
62. Andresen, M.A., Malleson, N.: Police foot patrol and crime displacement: a local analysis. J. Contemp. Crim. Justice **30**(2), 186–199 (2014)
63. Waples, S., Gill, M., Fisher, P.: Does CCTV displace crime? Criminol. Crim. Just. **9**(2), 207–224 (2009)
64. PredPol: The Myth of Crime Displacement https://www.predpol.com/crime-displacement-predpol/ (2021c). Accessed May 2021
65. Costanza-Chock, S.: Design Justice: towards an intersectional feminist framework for design theory and practice. In Proceedings of the Design Research Society 2018. SSRN: https://ssrn.com/abstract=3189696 (2018)
66. Broussard, M.: Artificial Unintelligence: How Computers Misunderstand the World. MIT Press, Cambridge (2018)