Article

Machine Learning and the Police: Asking the Right Questions

Annette Vestby[*,**] and Jonas Vestby[***]

Abstract How can we secure an accessible and open democratic debate about police use of predictive analytics when the technology itself is a specialized area of expertise? Police utilize technologies of prediction and automation where the underlying technology is often a machine learning (ML) model. The article argues that important issues concerning ML decision models can be unveiled without detailed knowledge about the learning algorithm, empowering non-ML experts and stakeholders in debates over if, and how to, include them, for example, in the form of predictive policing. Non-ML experts can, and should, review ML models. We provide a 'toolbox' of questions about three elements of a decision model that can be fruitfully scrutinized by non-ML experts: the learning data, the learning goal, and constructivism. Showing this room for fruitful criticism can empower non-ML experts and improve democratic accountability when using ML models in policing.

Introduction

Police increasingly apply advances in computer science and statistics to attempt to predict events and automate work. In this, policing is like numerous other fields; machines are, for instance, used to count votes, drive cars, predict the weather, decide loan applications, and more. Predictive analytics support risk management across the field of security governance (Hälterlein and Ostermeier, 2018). London, Los Angeles, Munich, New Orleans, Philadelphia, and Zürich are all examples of cities where police are using or have tested predictive policing software that aims to either predict where crimes are likely to take place, or who may be likely to commit a crime in the future. Machine learning (ML) is a key technology underlying many of these applications.

While ML software may rationalize otherwise laborious data-processing tasks, such as sifting through a vast cache of documents disclosed in an investigation and categorizing them (Hughes, 2017), many are concerned that using algorithmic tools to support or to automate decision-making has the inadvertent effect of reducing accountability (Barocas and Selbst, 2016; Lum and Isaac, 2016; Kroll *et al.*, 2017; Wilson, 2017). Although police accountability was a concern before the advent of predictive analytics, the use of these techniques has raised the question of whether employing ML models render humans unable to account for

[*]Annette Vestby, Doctoral researcher, Norwegian Police University College, Faculty of Law, The University of Oslo, Oslo, Norway. E-mail: anneve@phs.no

[**]Department of Criminology and Sociology of Law, University of Oslo, Oslo, Norway.

[***]Jonas Vestby, Senior researcher, Peace Research Institute Oslo, Oslo, Norway.

Policing, Volume 0, Number 0, pp. 1–15
doi:10.1093/police/paz035

decisions and how they were arrived at (Bennett Moses and Chan, 2016).

To hold police accountable for the fairness of their actions, and validity of their analyses, it is necessary to make processes of decision-making available for scrutiny. For example, transparency has been proposed as a solution to accountability issues (Pasquale, 2015; Bennett Moses and Chan, 2016; Mittelstadt *et al.*, 2016), as has the training of non-statisticians in statistics (Barocas and Selbst, 2016). Both suggestions presume that improved technical or statistical literacy is necessary to improve accountability when ML models[1] are applied in a socially consequential context such as policing. Although literacy in these fields is likely to benefit discussions among researchers, practitioners, policy-makers, and the wider public, it may not be a realistic goal. Furthermore, ML fluency alone is not enough to create morally acceptable and technically sound models (Holstein *et al.*, 2019). This article argues that technical literacy is often neither necessary nor sufficient to critically engage with the broad set of normative and technical questions raised by non-human agency in decision-making (cf. Hildebrandt, 2016a; cf. Zerilli, 2018). Such engagement is imperative to maintain and improve police accountability even in the context of new computational tools.

Besides formal accountability structures, a range of actors needs to deliberate and discuss implementation and use of ML software: internally in police organizations, between police professionals and in-house or commercial developers; stakeholders and affected populations with police and developers, and so on. Most of these cannot be expected to be experts in ML. Similarly, specialists in ML are neither experts in the broad set of issues faced in policing, nor have access to the issues visible to, for example, affected populations and end users (Marda, 2018; Holstein *et al.*, 2019). Facts are not only often uncertain in the social world in which policing operates, but values are also contested. Conversations about what good policing looks like and what its goal ought to be must allow for democratic participation (cf. Rønn, 2013). How then can we reconcile the need for cross-disciplinary and open conversation about the use of ML models in policing with the fact that the technologies themselves remain a highly specialized area of expertise? (cf. Callon *et al.*, 2009)

The ways in which technology is perceived contribute to what modes of accountability and participation it is possible to imagine (Elish and Boyd, 2017). This article demonstrates that it is not necessary to know ML algorithms to be able to engage critically with many of the important questions regarding the validity and fairness of applied ML models in policing (and it is our assumption that many, if not most, of the important aspects of police practice can be subsumed under these concepts). More inclusive mechanisms of collective decision-making (Shapiro in Sklansky, 2008), for example, in the forms of stakeholder and civil society involvement (Cath, 2018) can enhance the fairness and validity of applied ML models in policing (cf. Holstein *et al.*, 2019). This article contributes a toolbox of clear and precise questions that can be used in fora where those with and without

[1] A ‘model’ is the system of weights that will be trained using learning data and the learning algorithm. The weights are numerical and are used to calculate predictions when given new data. They can be as simple as $Y = \mathrm{b}\mathrm{X}$ (where b is the weight of input data feature X), or as complex as millions of weights connected to each other through convolutional or recurrent networks and including functions that transform the output of these systems. Commonly throughout this article, we will use ‘model’ to mean the fully trained model, that is, the model after the weights have been updated by the learning data using the learning algorithm. The fully trained model (and not the learning algorithm) is what practitioners will be using to produce new predictions that can go into decision-making. The development of the structure of the model weights (for instance, the size and number of convolutional layers) is the domain of an ML expert and can have severe implications for the ability of the model to learn from new data. In this article, therefore, we will include such considerations into the term ‘ML algorithm’, although a more common use of this term would be to include only the algorithm for how to update the weights assigned to data.

ML expertise may discuss on even terms to advance accountability in police decision-making or improve on developing or implemented ML technology.

Background

Predictive policing can be considered as a particular technique under the wider umbrella of intelligence-led policing (ILP) (Fyfe *et al.*, 2018). ILP emerged as a practical, managerial programme for basing decisions about police services on objective data analysis (Ratcliffe, 2016). Systematic collection and analysis of intelligence are intended to improve both the effectiveness of interventions against crime, providing more accurate targeting, and the cost efficiency (Innes and Sheptycki, 2004; Tilley, 2008). In predictive policing, as in ILP, analysis and decisions are centralized and rationalized; predictive policing '[emphasizes] the objective, scientific selection of strategies and tactics, and puts a premium on centralized, rationalized, bureaucratic decision-making.' (Sklansky, 2011, p. 4)

Police accountability

Keeping police organizations and officials answerable and responsible is a key component of democratic policing, and has long been a concern of police researchers and practitioners (cf. Goldstein, 1960; Reiner, 2013). Control over individual and organizational police conduct has been sought in part through accountability systems by which police may be answerable to the public, a bureaucracy, or the law (Dowdle, 2017). In terms of the position of police forces within the democratic system, accountability can mean political control over the police, or cooperation between the police and government, whereby the police are expected to provide explanations for decision-making (Chan, 1999, pp. 252–253).[2]

The application of predictive or automation software to support decision-making may fundamentally challenge the ability of officers and organizations to account for decision-making processes, as well as obfuscate responsibility in 'multi-agent structures' composed of humans and computational tools (Bennett Moses and Chan, 2016, p. 12). The opacity of 'algorithms'—applied predictive models or automated decision-making systems—remains at the core of the concerns about their use (Diakopoulos, 2015; Burrell, 2016; Mittelstadt *et al.*, 2016; Wilson, 2017). There is a worry that algorithms 'are opaque in the sense that if one is a recipient of the output of the algorithm (the classification decision), rarely does one have any concrete sense of how or why a particular classification has been arrived at from inputs' (Burrell, 2016, p. 1).

When one or more elements of the decision-making process are not comprehensible, either of the aforementioned conceptions of accountability is challenged. A statistical model, typically embedded in commercial, off-the-shelf software, works as a 'black box', where inputs (e.g. geospatial data on crime or demographics) are processed into output (e.g. a forecast or classification) through a calculation that remains invisible to the end-user. While arguably not essentially inscrutable (Kroll, 2018), the process is practically inscrutable to non-experts (cf. Latour, 1999), and can make the basis and rationale for decisions unclear. How can there be effective political control over decision-making if a key component in the shaping of the decision-making is essentially unknowable? How can the police give full account of their decisions if they rested, in part, on an analysis that they themselves are unable to explain?

[2] A distinction is made in the literature between a traditional, legalist concept of accountability and a more recent form focused on value-for-money and effectiveness. 'The new accountability' has shifted the accountability emphasis from a legalist or public-interest standard to one 'committed to fiscal restraint, efficiency, performance and the cutting back of the public sector' (Chan, 1999, p. 254). Or, '[a]ccountability has become accountancy' (Reiner, 2013, p. 1).

Transparency has been held up as part of the ideal solution to the challenges posed by ML to accountable decision-making (Pasquale, 2015; Bennett Moses and Chan, 2016; Hildebrandt, 2016b). To achieve transparency, information must be both accessible and comprehensible (Mittelstadt *et al.*, 2016, p. 6). This is difficult, however, when it comes to semi-autonomous learning machines. Some have thus argued that accountability may be possible without full transparency (e.g. the disclosure of source code) by designing accountability into the software (Kroll *et al.*, 2017). In addition to technical scrutiny and oversight, the application of algorithmic decision-making or algorithm-supported technology requires societal oversight, including public debate (Marda, 2018; Zweig *et al.*, 2018). Building on insights in this vein, we provide in this article an operationalization of these principles in the form of illuminating questions that lower the bar for entry into debates about the use of ML models in policing. Doing this, we effectively point to and demarcate a space where statistical or data science literacy is not a prerequisite to participation.[3]

ML and policing

A widely cited understanding of learning in the context of artificial intelligence (AI) is that learning has occurred if 'an agent[4] improves its performance on future tasks after making observations about the world' (Russell and Norvig, 2010, p. 693). This understanding requires agreement as to what it means to become better at a task. Two judges of an agent's performance might disagree over how much the agent has learned. It follows that agreement on a judgement of how well any agent is learning depends on a prior agreement on how to evaluate performance. Some matters are easier to reach agreement about than others. There are, for instance, performance criteria that aptly capture what it means for vehicles to merge onto a highway (Knight, 2017) and we might not expect too much disagreement on this point. It is harder to reach agreement about more complex social issues such as how to balance performance measures of law enforcement against minor offences given that there are possible costs to citizen trust in police (cf. Lum and Nagin, 2017). Just as agreement that someone is learning is more likely if agreement has already been achieved as to what it means to become better, so agreement that ML is useful is more likely if everyone already agrees on the learning goal of the machine learner.

While machines have been able to learn from data for quite some time, in the last decades, machines have become able to learn and excel at cognitive tasks, such as labelling objects in pictures and identifying words from sound. One technological application of this has been automated number plate recognition (APNR). Mounted on police vehicles, APNR has facilitated police monitoring of offenders (Stanier, 2016). These developments in ML capability came through a combination of new learning algorithms (some developed from the 1950s and onwards), more computational power, and the development of code to use the machine computational power effectively to solve the learning problems (Schmidhuber, 2015).

[3] Of course, such discussions need to be part of a wider, multifaceted accountability system, which it falls outside the scope of this article to address. The toolbox of questions offered in this article cannot, for example, reduce the opacity created by commercial secrecy (Burrell, 2016), which would require to make accountability actionable (Wright *et al.*, 2015), for example, by legislating a right to receive an explanation of machine decisions when requested (Norwegian Board of Technology, 2018), as well as having independent ML experts and non-ML-experts evaluate the outcomes of decisions made by machine models (Bennett Moses and Chan, 2016).

[4] We will use 'agent' in this text to mean something or someone that is capable of making decisions (if not actually acting them out in the world), such as a human individual, an organization, or a machine-learned model.

In addition to being able to learn cognitive tasks, another equally important ML development has been the invention of learning algorithms that can approximate complex functions and select important features without overfitting[5] (Hastie et al., 2009) the model to the training sample. These algorithmic improvements have made it possible for the machine to learn from datasets with thousands of labelled features so that it can pick out features (variables) and a functional form that is likely to perform well when predicting new samples. The implication is that the variables used in ML models are not necessarily chosen by human field experts, but rather by the ML algorithm itself, and that decisions are made less based on theories developed by humans, and more from a 'what works' perspective in terms of ML predictive power. Not surprisingly, these new abilities have made machine-learned models increasingly useful to agencies in decision-making and practice. ML models have been used, for example, by the UK Serious Fraud Office to identify legally privileged material among millions of disclosed documents in an investigation (Hughes, 2017), and by the Norwegian Labour Inspection Authority to predict high-risk workplaces to be inspected by the agency (Øyvann, 2017).

When discussing whether to use ML in police decision-making, it is important to compare ML, not to ideal decision-making, but to human decision-making (Bennett Moses and Chan, 2016, p. 7). Machines reach decisions in suboptimal environments based on inconclusive, inscrutable, and misguided evidence (Mittelstadt *et al.*, 2016). Whenever decision-making leads to unfair outcomes, processes may be hard to trace and it is 'rarely straightforward to identify who should be held responsible for the harm caused' (Mittelstadt *et al.*, 2016, p. 5). This is, however, a fundamental problem of decision-making *per se*, and not unique to decisions made or supported by machines (Zerilli *et al.*, 2018).

Humans excel at learning from cognitive data. Through hearing sounds, watching faces, and observing our surroundings, we distinguish syllables, words, sentences, and meaning. We can connect the dots between a smile, a sarcastic tone, the literal meaning of a sentence, and what the speaker intended to say. We can read books and news and talk to people, and from these activities, draw conclusions such as 'Democratic governance cannot allow police unfettered authority to achieve security; rather, police must do so in a manner that not only is within legal bounds but also is acceptable to citizens.' (Lum and Nagin, 2017, p. 361). Computers still do not make as comprehensive use of cognitive data as humans do. And whereas humans are always collecting (if not learning from) the whole sensory range of their experiences, only specific data (e.g. images/sound/video of a certain kind) are commonly collected for the purpose of training computers.

An important difference between machine and human learning is that ML is based on known algorithms. The Merriam-Webster definition of algorithm is 'a procedure for solving a mathematical problem . . . in a finite number of steps that frequently involves repetition of an operation' (Algorithm, n.d.). Humans, of course, also have procedures to solve problems in a finite number of steps and that frequently involve repetition of an operation. However, even the person using them may not always know or understand these procedures.

Since we both know the algorithms machines use (we write them down in programming languages), and can control the data by which they have learned (we can reset their biases at any time, feed particular training data to the model, or stop the learning process at any time), the learning and

[5] A model is overfit when it adjusts too much to quirks of the learning sample, and thus ends up performing worse on new data.

subsequent decisions are, in principle, more transparent in the case of machines than they are for humans (Zerilli *et al.*, 2018).[6] After all, we have not written the code for human learning, and we have little control over the input data that humans have used in their training. There is thus some irony in that one of the main critiques of the use of ML in decision-making is that machine decisions are opaque.

One possible explanation for this discrepancy is that it can be relatively straightforward to ask humans how they came to their decisions. It would be reasonable to expect a police chief to explain the facts, interpretations, and priorities behind her/his decision-making.[7] It can be much more difficult to produce similar explanations for why a machine model ended up with its biases;[8] in many cases, it can even be difficult to describe these biases in straightforward language. The learning opacity of machines may, in principle, be lower than it is for humans, but in practice, it is higher. As humans, we are better equipped to inquire of other humans how they reached their conclusions than we are to interrogate a machine model.

This opacity, although understandable, is worrisome because it could entail 'de-responsibilisation' of human actors in mixed networks of human and machine actors (Mittelstadt *et al.*, 2016, p. 12). While discriminatory policing practices have also arisen from purely human practices, '[...], filtering this decision-making process through sophisticated software that few people understand lends unwarranted legitimacy to biased policing strategies' (Lum and Isaac, 2016, p. 19). In other words, the machine's output may appear 'de-subjectified' (Završnik, 2017) and thus be interpreted by end-users as more objective than it actually is (cf. Elish and Boyd, 2017; Waardenburg *et al.*, 2018).

However, we believe that the opacity problem should not be exaggerated, and that it is necessary to distinguish between various sources of complexity, impenetrability, and even obfuscation (cf. Burrell, 2016). We disagree that ML algorithms are 'inherently opaque' (Hildebrandt, 2016b, p. 57), and furthermore, we argue that common variations on 'the fallacy of inscrutability' (Kroll, 2018) belie the potential for empowerment of non-specialists in debates over the use of ML technologies.

In fact, many relevant normative and factual judgements that comprise decisions by humans often do not depend on knowing or understanding the exact interplay of data and algorithm behind the decision (c.f. 'System 1' in Kahneman, 2011). Moreover, we are perfectly able to understand human behaviour without consideration of the inner workings of the neural network that is our brain (Dennett, 1995; Zerilli *et al.*, 2018). A useful starting point to begin to understand an agent's actions is to consider the previous experiences of the agent, what the agent wishes to accomplish, and what consequences the agent anticipates from its actions. We suggest that it can be helpful to structure a discussion between ML experts and non-ML experts around three elements that mirror this type of inspection: (1) the type of data we use to learn; (2) the learning goal we set; and (3) how later actions affect subsequent training data. These are elements that those who are not ML experts can understand and usefully discuss with ML experts,

[6] The blog posts of Andrej Karpathy (the director of AI at Tesla) contain excellent illustrations and examples of these points (http://karpathy.github.io/). The posts guide the reader through central algorithms and provide their source code. Karpathy has even written a JavaScript implementation of convolutional neural networks (https://cs.stanford.edu/people/karpathy/convnetjs/), so the reader can follow the training ('learning') process in real time on a web browser.

[7] For computers, we have not focused much on building software that provides post hoc explanations for a given machine decision in a way that any reasonable person would be able to comprehend. While the result of such an explanation in principle would be more transparent, the communication tools needed are not (yet) there (DARPA, 2016).

[8] Here, 'bias' just means the values of the weights in the model. These values will lead the model to produce biased results, preferably towards producing outcomes that we deem as proper given the learning task.

and that do not rely on the algorithm used for learning.[9,10]

A useful assumption for non-ML experts when discussing ML models is to assume that the learning algorithm chosen by the ML expert is optimal for achieving the established goal with the given data. While this assumption is many times wrong, it has the benefit of making much of the complexity of ML, such as knowing how recurrent neural networks function, irrelevant. We believe this assumption can lower the bar for non-experts for entry into a discussion with ML experts and facilitate a fruitful debate.

We do not imply that ML experts should be left to their own devices when it comes to designing the optimal learning algorithm for a given problem. Rather, the institutions that can ensure optimal learning algorithms, such as competitive environments and peer review, are clearly important. Our point is to delineate those aspects of the development of ML decision-making that can be the domain of all, experts and non-experts alike, and identify those aspects that require ML knowledge.

Optimal in this context is not a normative term, and there is a key distinction to be drawn between the concepts optimal and good. Computation and statistics offer the ability to test in a cost-effective way a vast number of possible models. For example, we can use ML algorithms to run a large number of tests to decide which parameters are important predictors of individual recidivism (cf. Berk and Bleich, 2013). The goal of an ML algorithm is to identify the optimal parameters for reaching the defined learning goal, disregarding such things as ethical concerns pertinent to policing unless these are explicitly operationalized and programmed (Norwegian Board of Technology, 2018, p. 12). Optimization means choosing the parameters that make the most accurate predictions given the data and learning used, so that the best performance possible is achieved within that given frame. A suboptimal algorithm will result in poor learning, whatever the machine is set to learn—whether it is good or bad, morally speaking. Bad decisions can arise even assuming an optimal learning algorithm.

In a survey of ML practitioners about how to improve fairness in their systems, the most commonly reported strategy was to collect more training data, and respondents struggled to anticipate which subpopulations and forms of unfairness they needed to consider (Holstein *et al.*, 2019). Both findings point to the benefit of lowering the bar for democratic participation in the development and auditing of machine-learned models. It is crucial to realize that ML specialists are not necessarily the experts in answering or having knowledge about issues of fairness or of how models will be perceived, used, and work in an applied context. Rather, these issues can be perceived by experts and stakeholders in domains other than ML.[11] Machine-aided decision-making, as in the case of human decision-making overall, benefits in the end when people can discuss these issues in open, democratic forums (cf. Elster, 1998; Habermas, 2000).

Asking about fairness and validity: a toolbox

As a society we have an interest in crime prevention and efficient policing, but we also have an interest in ensuring that law enforcement strategies, including deployment and surveillance

[9] Indeed, the algorithms in ML are used to learn some goal from data. Furthermore, the performance of machine-learned models is not generally measured in the beauty or structure of the algorithm, but in how well the model perform on the learning task for a particular set of data (Hastie et al., 2009).

[10] The terms ‘interpretable’ and ‘explainable’ AI are used in the wider AI field and literature. Work to increase interpretability and understanding of ML models (i.e. procedures that (unlike the suggestion put forward by us in this article) depend on the ML algorithm used) is underway in a field of research called ‘explainable AI’ (DARPA, 2016).

[11] One responder replied ‘You’ll know if there’s fairness issues if someone raises hell online’ (Holstein, *et al.*, 2019, p. 7).

> decisions, are effective, fair, and just. This requires understanding, testing, and governance
> (Bennett Moses and Chan, 2016, p. 14).

Broadly speaking, decisions can be criticized with respect to two different issues: the validity[12] of the decision and the fairness of the decision. To consider the validity of the model, we ask: did the decision lead to the intended result? To evaluate validity, a reviewer would need to consider whether the learning model reflects actual performance based on the agreed-upon performance metric, or whether the performance metric itself measures what we intended to measure.[13] Since learning goals can be quite abstract and contested (e.g. the goal of reducing crime), the scope of validity issues is likely to overlap with domains outside those of programmers and statisticians. However, even quite narrow issues, such as selection bias in the training data, may be easier for non-ML experts to expose who may know, for instance, how data are collected. As an example of the latter, Sheptycki (2004) found that information was more likely to be recorded by police officers if it was considered by them as useful to successfully prosecute a crime.

Reviewing the fairness of a decision resulting from either a human or a machine model, involves asking whether the intended result, and the means to achieve it, were good? Evaluating fairness is a normative endeavour. It entails, in this context, to consider if the learning goal, the process that improves learning, and the means for achieving learning success, are determined in a democratically legitimate way. Ensuring the possibility of an open and democratic debate is both a requirement as well as part of the solution to the fairness issue.

What follows is a toolbox of questions that non-experts can ask creators of machine-learned models with the expectation of receiving understandable answers. Replies in the form of 'however, we have accounted for this in our model' require modelling decisions that could be stated explicitly, and these decisions need to be known caveats for everyone using the model. We have divided the toolbox into sections with questions about data, about learning, and about constructivism. We discuss both validity and fairness issues in each section. The goal of the toolbox is to empower non-ML experts in debates with ML experts.

Asking questions about the data

Some crimes are more likely than others to be recorded by the police, and only recorded crimes become crime data. Thus, crime statistics have passed through a process of selection. The first stage in the process is legislative; this is when certain acts are criminalized. A further selection occurs because some crimes are not reported or discovered by the public and police; in addition, reporting practices may vary with crime type and district. Some are unlikely to be discovered let alone reported if not for systems for inspection or mandated reporting. Economic crime is an example of the latter category; an example is tax avoidance, where reporting depends on audits and inspections by designated agencies (Korsell, 2015), and the finance industry can mobilize secrecy to resist financial crime surveillance (Pasquale, 2015, 2017).

Practices, methods, and emphases of the police, and the other agencies, businesses, and citizens that

[12] In the applied ML context that we are mostly thinking of in this article, where learned models are used to make judgments in new cases, we are concerned about the external validity of the model. Of course, external validity depends on internal validity, and many of the issues we discuss in this section would affect internal validity as well as external.

[13] One aspect of internal validity is whether the model reflects causal mechanisms. In most settings, a machine-learned model would be answering a much more pragmatic question (such as, are we becoming better at doing a specific task, as defined by the learning goal and the data used to train the agent and test its performance?). There is no guarantee that the model would learn actual causal mechanisms. There exist some arguments for the connection between learning and causality, such as the probably approximately correct theorem (Valiant, 1984). However, in many ML applications, we are more concerned with what works than with why it works. ML can be used to probe hypotheses about causal effects, however (Rubin, 1974).

report to the police thus shape the composition of the data. Thus, it is not straightforward to establish the relationship between these known crimes and the ‘actual’ extent and distribution of crimes (i.e. the dark figure; see e.g. Reiner, 2016, p. 108). However, crimes that control agencies focus on and that are not generally reported by anyone else are particularly vulnerable to over-representation in the data in relation to their actual distribution in the universe of crime.

Problematic as well as desirable policing practices inscribe themselves on police-generated data. A study by the Human Rights Data Analysis Group provides an illustrative example (Lum and Isaac, 2016). The study modelled predictive policing forecasts using the published algorithm for PredPol (Mohler *et al.*, 2015) and police data on drug policing in Oakland, CA, and then compared the forecasts with patterns of drug use estimated from national survey data on drug use and health. It found that using the PredPol algorithm, ‘black people would be targeted by predictive policing at roughly twice the rate of whites’, despite estimates showing roughly equal levels of drug use (Lum and Isaac, 2016, p. 18). Low-income people and non-Whites other than Blacks would also be disproportionately targeted, that is, over-policed.

This example shows how input data used to train machines and humans alike can lead to invalid models and unfair practice. In this case, the invalid model or belief is that that targeting Black residential areas is a reasonable way to conduct drug policing, despite the fact that patterns of drug use suggest that Black residential areas should not have higher incidences of drug use. The result is unfair police practice, whereby Black citizens and neighbourhoods are policed more than Whites despite the lack of an objective basis in racial patterns of drug offence.

Those without expertise in ML can ask the following about data:

- what input data are used? What set has been used to train the model? What set is used to test performance? When and where were the data collected?
- are there named variables? If so, what are they and which contribute most to the decisions? How are these named variables operationalized and measured?
- does input data capture features (directly or indirectly) that should not be relevant to the decision? For example, are any input features correlated with gender in such a way that model decisions are different if you are male or female?
- is the data representative of the field that the model decisions affect? For example, has the model been tested in the setting where it is applied? What are the most obvious differences between the training setting and the current setting? Do we need to make any adjustments for particular groups or decisions? and
- how are the data collected? For example, were they collected with the intention of being used for these kinds of decisions? Do we know of any selection biases (either by design or due to practical issues) with regard to the data collection? Who collects the data?

Asking questions about learning

All learning has a goal. In ML models, goals can be more or less explicit. Regardless of whether the learning is supervised, unsupervised, or reinforced,[14] it is possible and meaningful to ask what the overarching learning goal is and what specific

[14] In supervised learning, the correct response for any given input is provided so that the learning algorithm can attempt to reduce the error given this solution. Unsupervised learning uses rules, like similarity, to cluster observations. Here, the learning goal might be to cluster what we deem as relevant observations together. Lastly, in reinforcement learning, rewards and punishments for specific actions are provided to induce specific behaviour in the actor using the model.

rule or measurement is being used as the reference for determining if a model is learning.

As discussed earlier, it is easier to reach agreement on whether an agent is learning when agreement has already been established regarding the larger issue of how to evaluate performance. However, this is often not the case with social issues. Most social issues are complex; values are often in dispute and the facts may be uncertain. This complicates police decision-making. For example, given that resources are finite, should the police maximize their response to minor offences, or focus efforts at crime prevention? (cf. Lum and Nagin, 2017). Can police analysts objectively adjudicate this by measuring the harm (a value concept) caused or prevented (Rønn, 2013), and define where resources might 'do "the most good"' in a way that all agree with? (Sklansky, 2008, p. 122)

ML models optimize against particular learning goals that must be operationalized and measured. Since some types of outcomes are easier to measure than others, there is an inherent bias in ML models for choosing the learning goals that are easiest to measure.[15] Outcomes that have already been measured, such as the location of arrests, thus become more attractive than unmeasured outcomes, such as citizen response to police tactics (Lum and Nagin, 2017). When inherent bias is transferred from the machine models into actual decision-making, the consequences can be wide-ranging as the HRDAG study shows (Lum and Isaac, 2016).

When a learning goal, or what constitutes good performance of that goal, is disputed, and when learning goals are operationalized differently than what we ideally would want, predictions from ML models must be applied with caution, if at all. An extreme example can be found in Wu and Zhang (2016) who claim that their ML model can automatically identify criminals from facial characteristics only, and 'empirically establish the validity of automated face-induced inference on criminality, despite the historical controversy surrounding this line of enquiry' (Wu and Zhang, 2016, p. 1). Here, the model does not separate criminals from non-criminals, but rather photos of convicts and suspects from a set of ID photos taken from the Internet. The authors themselves agree with critics who argue that a difference in socio-economic status in the two sets could possibly explain why the model manages to separate the sets (Wu and Zhang, 2016, p.3). If we, for the sake of argument, bypass the looming question of the purpose and value of automatic recognition of 'criminals', that is, the learning goal, it should be obvious that it would be highly problematic to use a model purporting to identify criminals that may in fact simply identify poverty.

Two clear concerns when thinking about employing an ML model in decision-making processes are (1) whether the operationalized goal optimized against in the ML model is delivering good performance also when measured against a more general and overarching learning goal and (2) whether the operationalized goal produces unwanted side effects. Humans commonly disagree on how best to solve social issues, and institutions such as political parties, academia, and the media, may facilitate discussion that is needed to reach agreement. Within these discussions, narrow arguments about the performance of ML models should be regarded as arguments about efficiency, not efficacy.

A further concern is that the ML model optimizes against many, but not all, aspects of the overarching learning goal(s). In developing a machine model and measuring the data that goes into learning, some aspects can be lost. By openly discussing the purpose of the agent, and what the overarching learning goals should be, it is possible to identify the elements that the ML model is not optimizing against and take appropriate action. When the

[15] A similar dynamic is discussed in relation to management by output measurement in the public sector. Smith terms a potential consequence 'tunnel vision', which 'can be defined as an emphasis on phenomena that are quantified in the performance measurement scheme at the cost of unquantified aspects of performance.' (1995, p. 284)

ML model only optimizes against some of the established goals, we should be wary about letting the ML model decide actions directly.

Those without expertise in ML can ask:

- what is the overarching learning goal? For example, what would we, as a society, like to accomplish by making these decisions?
- what specific rule(s) or measurement(s) are used as the reference for whether a model is learning? For example, what is the dependent variable(s)? What kind of similarity rule is being used? What kinds of actions are rewarded or punished? How is the rule operationalized and measured?
- is there agreement on the learning goal?
- is the specific learning goal a complete description of what the agent is supposed to achieve? and
- will optimizing action or decision-making against this learning goal take effort away from, or actively work against, other goals?

Asking questions about constructivism

Our models, be they machine or mental, affect the world when we use them to make decisions. In policing, making some sort of wanted impact is of course the point. Predictive analyses are meant to guide action, 'to identify likely targets for police intervention' (Perry *et al.*, 2013, p. xiii). Decisions, actions, analyses, policies, and local and historical contexts contribute to present day policing concepts and practices. Unlike in the field of physics, say, our policing decisions affect social systems. We use the term constructivism to denote this insight.

The constructivist fact about the social world raises three main concerns for ML in decision-making. First, data can become outdated or otherwise fail to generalize; as a result, they will no longer provide good guidance for decision-making. Secondly, past decisions can reinforce unwanted or erroneous patterns used in the training of models. Thirdly, a narrow focus on predictive performance within the bounds of the learning goal can make more difficult arguing for decisions intended to break or change social patterns.

The first concern mainly regards validity. ML algorithms can be used to make models that are optimized for a variety of settings. However, making models that fit particular settings can be difficult and time-consuming. In practice, therefore, it is reasonable to assume that the use of data that is unfit as a basis for generalization is widespread. Thus, input from outside the ML domain of expertise is important, in particular to make clear what data the model is optimizing against, and to demand that machine-learned models be shown to perform well in the particular settings in which they are implemented.

ML models identify patterns in data. When police implement ML models with mistaken causal assumptions, such as the ones exposed in the HRDAG study, they will reinforce the erroneous correlational patterns that underlie the model. These can then be picked up by later generation ML models and used to improve performance on their set learning task. If we continue to learn using models based on the same incorrect assumptions, and to rely on data that are reinforcing the correlational patterns, then we reproduce the same error (Zhang *et al.*, 2018).

Predictability is desirable because it commonly promises great cost-efficiency. The emphasis on resource efficiency is a selling point for predictive policing; it moves 'law enforcement from focusing on what happened to focusing on what will happen and how to effectively deploy resources in front of crime, thereby changing outcomes.' (Beck and McCue, 2009, p. 1). However, a sole emphasis on predictability can lead to choosing the learning goals that are easiest to predict, or to relying on correlational patterns that may have dubious causal merit to predict more accurately.

The validity issues discussed in the previous paragraphs have strong implications for fairness and the democratic quality of policing. Policing distributes benefits to and burdens on citizens, and impacts the distribution of security among individuals and communities (Brodeur, 2010, pp. 135–136).

The democratic quality of policing is among its important moral dimensions. At a minimum, police action must be legal. But a commitment to democracy places demands on the police above this minimal threshold. For instance, the anti-inegalitarian view of democracy in the work of Ian Shapiro, entails ‘ongoing opposition to patterns of unjustifiable hierarchy (Sklansky, 2008, p. 109). Maximizing the democratic quality of policing means ‘making it as effective as possible in combating unjustified patterns of private domination and unthreatening as possible as a tool of official domination.’ (Sklansky, 2008, p. 109). To focus law enforcement disproportionately on disadvantaged groups embeds domination, not least through the reinforcing effect of the data stream going back into the police organization. A consequence of constructivism is, therefore, that we cannot ignore causality or ethics and rely solely on predictive performance in decision-making.

Interestingly, the possible pitfalls related to pattern reproduction also point to where ML models can improve on human learning and practice. Algorithmic tools can detect discrimination (Mittelstadt *et al.*, 2016, p. 15), but in contrast to individuals and organizations, they can be used to actively withhold from analysis dubious relationships between, for example, ethnicity and crime or ZIP code. While there is reason to be sceptical of purely technical solutions to protect, for example, a complex social concept such as ‘fairness’ (Lipton and Steinhardt, 2018), however, the work done to identify discriminatory practices and mitigate unfairness in and through algorithmic tools also represents opportunities to improve on human decision-making (Zerilli, 2018; Zhang, *et al.*, 2018; Holstein *et al.*, 2019).

Those without expertise in ML can ask:

- can the machine decision, if acted upon, affect later training data?
- does the machine model represent a causal relationship, or is it a pragmatic solution?
- does the model rely on correlations that likely only improve performance due to historical practices? Are these historical practices morally contested? and
- would we like to break or change certain observable patterns in society? If so, what potential consequences would this change involve for the machine model?

Conclusion

As police departments seek to prevent both harm and spend resources frugally, they are increasingly adopting proactive policies and techniques. However, the use of predictive tools requires careful consideration, and we have argued that ML expertise is not necessary to participate in debates over many important facts and normative issues. Questions about the purpose of technology or police are both moral and political ones (cf. Turkle, 2004). Our goal is to empower non-technical experts and stakeholders and encourage their participation in debates over applied ML in policing, as well as in processes of ML model development. Several arguments have been made that such participation is not only technically and morally necessary (Cath, 2018; Holstein *et al.*, 2019; cf. Rønn, 2013) but also feasible (cf. Zerilli, 2018). This article contributes a toolbox of questions that in effect operationalizes such calls and provides context that illustrates the utility and purpose of asking them in the police and related crime control domains. Asking about the data, the learning goal, and how model decisions affect later data are three concrete lines of inquiry that non-experts can understand, and should discuss.

Acknowledgments

The authors are grateful to all who have provided valuable feedback on this article, in particular the two anonymous reviewers, Helene O.I. Gundhus and the New Trends in Modern Policing project group, Lynn P. Nygaard, and the Young Nordic Police Research Network.

Funding

This work is funded by the European Research Council, Grant No. 648291, the Norwegian Research Council Grant No. 238170, and the Science Studies Colloquium Series at the University of Oslo.

References

Algorithm (n.d.). *Merriam-Webster Online*. https://www.merriam-webster.com/dictionary/algorithm (accessed 16 October 2018).

Barocas, S. and Selbst, A. D. (2016). 'Big Data's Disparate Impact'. *California Law Review* **104**: 671–732.

Beck, C. and McCue, C. (2009). 'Predictive Policing: What Can we Learn from Wal-Mart and Amazon about Fighting Crime in a Recession?'. *Police Chief* **76** (11): 18–24.

Bennett Moses, L. and Chan, J. (2016). 'Algorithmic Prediction in Policing: Assumptions, Evaluation, and Accountability'. *Policing and Society* **28** (7): 1–17.

Berk, R. A. and Bleich, J. (2013). 'Statistical Procedures for Forecasting Criminal Behavior: A Comparative Assessment'. *Criminology & Public Policy* **12**(3): 513–544.

Brodeur, J. P. (2010). *The Policing Web*. New York: Oxford University Press.

Burrell, J. (2016). 'How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms'. *Big Data & Society* **3**(1): 1–12.

Callon, M., Lascoumes, P. and Barthe, Y. (2009). *Acting in an Uncertain World: An Essay on Technical Democracy*. Cambridge, MA: MIT Press.

Cath, C. (2018). 'Governing Artificial Intelligence: Ethical, Legal and Technical Opportunities and Challenges'. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **376**(2133): 20180080.

Chan, J. B. (1999). 'Governing Police Practice: Limits of the New Accountability'. *The British Journal of Sociology* **50**(2): 251–270.

DARPA (2016). *Broad Agency Announcement. Explainable Artificial Intelligence (XAI)*. https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf (accessed 31 March 2019).

Dennett, D. (1995). *Darwin's Dangerous Idea: Evolution and the Meanings of Life*. New York: Simon & Schuster.

Diakopoulos, N. (2015). 'Algorithmic Accountability: Journalistic Investigation of Computational Power Structures'. *Digital Journalism* **3**(3): 398–415.

Dowdle, M. W. (2017). 'Public Accountability: Conceptual, Historical and Epistemic Mappings'. In Drahos, P. (ed), *Regulatory Theory: Foundations and Applications*. Canberra: ANU Press.

Elish, M. C. and Boyd, D. (2017). 'Situating Methods in the Magic of Big Data and AI'. *Communication Monographs* **85**(1): 57–80.

Elster, J. (1998). *Deliberative Democracy*. Cambridge: Cambridge University Press.

Fyfe, N., Gundhus, H. O. I. and Rønn, K. V. (eds) (2018). *Moral Issues in Intelligence-Led Policing*. London; New York: Routledge.

Goldstein, J. (1960). 'Police Discretion Not to Invoke the Criminal Process: Low-Visibility Decisions in the Administration of Justice'. *The Yale Law Journal* **69**(4): 543–594.

Habermas, J. (2000). *The Inclusion of the Other: Studies in Political Theory*. Cambridge, MA: MIT Press.

Hälterlein, J. and Ostermeier, L. (2018). 'Special Issue: Predictive Security Technologies'. *European Journal for Security Research* **3**(2): 91–94.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*. New York: Springer.

Hildebrandt, M. (2016a). 'New Animism in Policing: Re-Animating the Rule of Law?'. In Bradford, B., Loader, I., Jauregui, B. and Steinberg, J. (eds), *The Sage Handbook of Global Policing*. Los Angeles, CA: Sage, pp. 406–428.

Hildebrandt, M. (2016b). 'The New Imbroglio. Living with Machine Algorithms'. In Janssens, L. (ed), *The Art of Ethics in the Information Society*. Amsterdam: Amsterdam University Press, pp. 55–60.

Holstein, K., Vaughan, J. W., Daumé H. III, Dudík, M. and Wallach, H. (2019). 'Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?' ArXiv: 1812.05239 [Cs]. https://doi.org/10.1145/3290605.3300830 (accessed 20 March 2019).

Hughes, D. (2017). 'Robot Investigators 'Could Be Used to Examine Documents in Criminal Cases'. *The Independent* (14 December 2017). https://www.independent.co.uk (accessed 20 November 2018).

Innes, M. and Sheptycki, J. W. (2004). 'From Detection to Disruption: Intelligence and the Changing Logic of Police Crime Control in the United Kingdom'. *International Criminal Justice Review* **14**(1): 1–24.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Knight, W. (2017). 'Reinforcement Learning: 10 Breakthrough Technologies 2017;. *MIT Technology Review* (April). https://www.technologyreview.com/s/603501/10-breakthrough-technologies-2017-reinforcement-learning/ (accessed 23 November 2018).

Korsell, L. (2015). 'On the Difficulty of Measuring Economic Crime'. In van Erp, J., Huisman, W. and Vande Walle G. (eds), *The Routledge Handbook of White-Collar and Corporate Crime in Europe*. Abingdon, Oxon; New York: Routledge, pp. 111–127.

Kroll, J. A. (2018). 'The Fallacy of Inscrutability'. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **376**(2133): 20180084.

Kroll, J. A., Huey, J., Barocas, S. *et al.* (2017). 'Accountable Algorithms'. *University of Pennsylvania Law Review* **165**: 633–705.

Latour, B. (1999). *Pandora's Hope: Essays on the Reality of Science Studies*. Cambridge, MA; London, England: Harvard University Press.

Lum, C. and Nagin, D. S. (2017). 'Reinventing American Policing'. *Crime and Justice* **46**(1): 339–394.

Lum, K. and Isaac, W. (2016). 'To Predict and Serve?'. *Significance* **13**(5): 14–19.

Lipton, Z. C. and Steinhardt, J. (2018). 'Troubling Trends in Machine Learning Scholarship'. ArXiv Preprint ArXiv: 1807.03341. https://arxiv.org/pdf/1807.03341.pdf (accessed 21 March 2019).

Marda, V. (2018). 'Artificial Intelligence Policy in India: A Framework for Engaging the Limits of Data-Driven Decision-Making'. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **376**(2133): 1–19.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S. and Floridi, L. (2016). 'The Ethics of Algorithms: Mapping the Debate'. *Big Data & Society* **3**(2): 1–21.

Mohler, G. O., Short, M. B., Malinowski, S. *et al.* (2015). 'Randomized Controlled Field Trials of Predictive Policing'. *Journal of the American Statistical Association* **110**(512): 1399–1411.

Norwegian Board of Technology (2018). *Artificial Intelligence: Opportunities, Challenges and a Plan for Norway*. Oslo: Norwegian Board of Technology.

Øyvann, S. (2017). 'AI finner bedrifter som skal ha tilsyn [AI identifies businesses for inspection]'. *Computerworld* (8 July 2017). http://www.cw.no/artikkel/offentlig-it/ai-finner-bedrifter-som-skal-ha-tilsyn (accessed 15 January 2019).

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge: Harvard University Press.

Pasquale, F. (2017). 'Paradoxes of Privacy in an Era of Asymmetrical Social Control'. In Završnik, A. (ed), *Big Data, Crime and Social Control*. Abingdon, Oxon; New York: Routledge.

Perry, W. L., McInnis, B., Price, C. C., Smith, S. C. and Hollywood, J. S. (2013). *Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations*. Santa Monica, CA: RAND.

Ratcliffe, J. H. (2016). *Intelligence-Led Policing*. Abingdon, Oxon; New York: Routledge.

Reiner, R. (2013). 'Who Governs? Democracy, Plutocracy, Science and Prophecy in Policing'. *Criminology & Criminal Justice* **13**(2): 161–180.

Reiner, R. (2016). *Crime: The Mystery of the Common-Sense Concept*. Cambridge; Malden, MA: Polity Press.

Rønn, K. V. (2013). 'Democratizing Strategic Intelligence: On the Feasibility of an Objective, Decision-Making Framework When Assessing Threats and Harms of Organized Crime'. *Policing* **7**(1): 53–62.

Rubin, D. B. (1974). 'Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies'. *Journal of Educational Psychology* **66**(5): 688–701.

Russell, S. and Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*. Boston, MA: Pearson.

Schmidhuber, J. (2015). 'Deep Learning in Neural Networks: An Overview'. *Neural Networks* **61**: 85–117.

Sheptycki, J. (2004). 'Organizational Pathologies in Police Intelligence Systems: Some Contributions to the Lexicon of Intelligence-Led Policing'. *European Journal of Criminology* **1**(3): 307–332.

Sklansky, D. A. (2008). *Democracy and the Police*. Stanford, CA: Stanford University Press.

Sklansky, D. A. (2011). 'The Persistent Pull of Police Professionalism'. *New Perspectives in Policing* 1–19.

Smith, P. (1995). 'On the Unintended Consequences of Publishing Performance Data in the Public Sector'. *International Journal of Public Administration* **18**(2–3): 277–310.

Stanier, I. (2016). 'Enhancing Intelligence-Led Policing: Law Enforcement's Big Data Revolution'. In Bunnik, A., Cawley, A., Mulqueen, M. and Zwitter, A. (eds), *Big Data Challenges*. London: Palgrave Macmillan UK, pp. 97–113.

Tilley, N. (2008). 'Modern Approaches to Policing: Community, Problem-Oriented and Intelligence-Led'. In Newburn T. (ed), *Handbook of Policing*. Cullompton, Devon: Willan Publishing, pp. 373–403.

Turkle, S. (2004). 'How Computers Change the Way We Think'. *The Chronicle of Higher Education* **50**(21): B26.

Valiant, L. G. (1984). 'A Theory of the Learnable'. *Communications of the ACM* **27**(11): 1134–1142.

Waardenburg, L., Sergeeva, A. and Huysman, M. (2018). 'Hotspots and Blind Spots'. In Schultze, U., Aanestad, M., Mähring, M., Østerlund, C. and Riemer K. (eds.), *Living with Monsters? Social Implications of Algorithmic Phenomena, Hybrid Agency, and the Performativity of Technology*, Cham: Springer International Publishing, pp. 96–109.

Wilson, D. (2017). 'Algorithmic Patrol. The Futures of Predictive Policing'. In Završnik, A. (ed.), *Big Data, Crime and Social Control*. Abingdon, Oxon; New York: Routledge, pp. 146–155.

Wright, D., Rodrigues, R., Raab, C. *et al.* (2015). 'Questioning Surveillance'. *Computer Law & Security Review* **31**(2): 280–292.

Wu, X. and Zhang, X. (2016). 'Responses to Critiques on Machine Learning of Criminality Perceptions' (Addendum of arXiv: 1611.04135). *ArXiv: 1611.04135 [Cs]*. http://arxiv.org/abs/1611.04135 (accessed 8 January 2019).

Završnik, A. (2017). 'Big Data. What Is It and Why Does It Matter for Crime and Social Control?' In Završnik, A. (ed), *Big Data, Crime and Social Control*. Abingdon, Oxon; New York: Routledge, pp. 24–50.

Zerilli, J., Knott, A., Maclaurin, J. and Gavaghan, C. (2018). 'Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard'? *Philosophy & Technology*, 1–23. https://doi.org/10.1007/s13347-018-0330-6.

Zhang, B. H.3, Lemoine, B. and Mitchell, M. (2018). Mitigating Unwanted Biases with Adversarial Learning. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, *ACM*. pp. 335–340.

Zweig, K. A., Wenzelburger, G. and Krafft, T. D. (2018). 'On Chances and Risks of Security Related Algorithmic Decision Making Systems'. *European Journal for Security Research* **3**(2): 181–203.