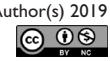
Article

European Journal of Criminology

Algorithmic justice: Algorithms and big data in criminal justice settings

European Journal of Criminology
1–20



DOI: 10.1177/1477370819876762
journals.sagepub.com/home/euc


Aleš Završnik
University of Ljubljana, Slovenia

Abstract

The article focuses on big data, algorithmic analytics and machine learning in criminal justice settings, where mathematics is offering a new language for understanding and responding to crime. It shows how these new tools are blurring contemporary regulatory boundaries, undercutting the safeguards built into regulatory regimes, and abolishing subjectivity and case-specific narratives. After presenting the context for 'algorithmic justice' and existing research, the article shows how specific uses of big data and algorithms change knowledge production regarding crime. It then examines how a specific understanding of crime and acting upon such knowledge violates established criminal procedure rules. It concludes with a discussion of the socio-political context of algorithmic justice.

Keywords

Algorithm, bias, criminal justice, machine learning, sentencing

Algorithmic governance: The context

Our world runs on big data, algorithms and artificial intelligence (AI), as social networks suggest whom to befriend, algorithms trade our stocks, and even romance is no longer a statistics-free zone (Webb, 2013). In fact, automated decision-making processes already influence how decisions are made in banking (O'Hara and Mason, 2012), payment sectors (Gefferie, 2018) and the financial industry (McGee, 2016), as well as in insurance (Ambasna-Jones, 2015; Meek, 2015), education (Ekowo and Palmer, 2016; Selingo, 2017) and employment (Cohen et al., 2015; O'Neil, 2016). Applied to social platforms, they have contributed to the distortion of democratic processes, such as general

Corresponding author:
Aleš Završnik, Institute of Criminology at the Faculty of Law, University of Ljubljana, Poljanski nasip 2, Ljubljana, SI-1000, Slovenia.
Email: ales.zavrsnik@pf.uni-lj.si

elections, with 'political contagion', similar to the 'emotional contagion' of the infamous Facebook experiment (Kramer et al., 2014) involving hundreds of millions of individuals for various political ends, as revealed by the Cambridge Analytica whistle-blowers in 2018 (Lewis and Hilder, 2018).

This trend is a part of 'algorithmic governmentality' (Rouvroy and Berns, 2013) and the increased influence of mathematics on all spheres of our lives (O'Neil, 2016). It is a part of 'solutionism', whereby tech companies offer technical solutions to all social problems, including crime (Morozov, 2013). Despite the strong influence of mathematics and statistical modelling on all spheres of life, the question of 'what, then, do we talk about when we talk about "governing algorithms"?' (Barocas et al., 2013) remains largely unanswered in the criminal justice domain. How does the justice sector reflect the trend of the 'algorithmization' of society and what are the risks and perils of this? The importance of this issue has triggered an emerging new field of enquiry, epitomized by critical algorithm studies. They analyse algorithmic biases, filter bubbles and other aspects of how society is affected by algorithms. Discrimination in social service programmes (Eubanks, 2018) and discrimination in search engines, where they have been called 'algorithms of oppression' (Noble, 2018), are but some examples of the concerns that algorithms, big data and machine learning trigger in the social realm. Predictive policing and algorithmic justice are part of the larger shift towards 'algorithmic governance'.

Big data, coupled with algorithms and machine learning, has become a central theme of intelligence, security, defence, anti-terrorist and crime policy efforts, as computers help the military find its targets and intelligence agencies justify carrying out massive pre-emptive surveillance of public telecommunications networks. Several actors in the 'crime and security domain' are using the new tools:[1] (1) intelligence agencies (see, for example, the judgment of the European Court of Human Rights in *Zakharov v. Russia* in 2015, No. 47143/06, or the revelations of Edward Snowden in 2013); (2) law enforcement agencies, which are increasingly using crime prediction software such as PredPol (Santa Cruz, California), HunchLab (Philadelphia), Precobs (Zürich, Munich) and Maprevelation (France) (see Egbert, 2018; Ferguson, 2017; Wilson, 2018); and (3) criminal courts and probation commissions (see Harcourt, 2015a; Kehl and Kessler, 2017).

The use of big data and algorithms for intelligence agencies' 'dragnet' investigations raised considerable concern after Snowden's revelations regarding the 'National Security Agency's access to the content and traffic data of Internet users. Predictive policing has attracted an equal level of concern among scholars, who have addressed 'the rise of predictive policing' (Ferguson, 2017) and the 'algorithmic patrol' (Wilson, 2018) as the new predominant method of policing, which thus impacts other methods of policing. Country-specific studies of predictive policing exist in Germany (Egbert, 2018), France (Polloni, 2015), Switzerland (Aebi, 2015) and the UK (Stanier, 2016). A common concern is the predictive policing allure of objectivity, and the creative role police still have in creating inputs for automated calculations of future crime: 'Their choices, priorities, and even omissions become the inputs algorithms use to forecast crime' (Joh, 2017a). Scholars have shown how public concerns are superseded by the market-oriented motivations and aspirations of companies that produce the new tools (Joh, 2017b). Human rights advocates have raised numerous concerns regarding predictive policing (see Robinson and Koepke, 2016). For instance, a coalition of 17 civil rights organizations has listed several

risks, such as a lack of transparency, ignoring community needs, failing to monitor the racial impact of predictive policing and the use of predictive policing primarily to intensify enforcement rather than to meet human needs (ACLU, 2016).

In Anglo-American legal systems, tools for assessing criminality have been used in criminal justice settings for a few decades. But now these tools are being enhanced with machine learning and AI (Harcourt, 2015b), and other European countries are also looking at these systems for several competing reasons, such as to shrink budgets, decrease legitimacy and address the overload of cases. The trend to 'algorithmize' everything (Steiner, 2012) has thus raised the interest of policy-makers. The European Union Fundamental Rights Agency warns that discrimination in data-supported decision-making is 'a fundamental area particularly affected by technological development' (European Union Agency for Fundamental Rights, 2018). The Council of Europe's European Commission for the Efficiency of Justice adopted a specific 'European Charter on the Use of AI in Judicial Systems' in 2018 to mitigate the above-mentioned risks specifically in justice sector.

In general, courts use such systems to assess the likelihood of the recidivism or flight of those awaiting trial or offenders in bail and parole procedures. For instance, the well-known Arnold Foundation algorithm, which is being rolled out in 21 jurisdictions in the USA (Dewan, 2015), uses 1.5 million criminal cases to predict defendants' behaviour in the pre-trial phase. Similarly, Florida uses machine learning algorithms to set bail amounts (Eckhouse, 2017). These systems are also used to ascertain the criminogenic needs of offenders, which could be changed through treatment, and to monitor interventions in sentencing procedures (Kehl and Kessler, 2017). Some scholars are even discussing the possibility of using AI to address the solitary confinement crisis in the USA by employing smart assistants, similar to Amazon's Alexa, as a form of 'confinement companion' for prisoners. Although at least some of the proposed uses seem outrageous and directly dangerous, such as inferring criminality from face images, the successes of other cases in the criminal justice system seem harder to dispute or debunk. For instance, in a study of 1.36 million pre-trial detention cases, scholars showed that a computer could predict whether a suspect would flee or re-offend better than a human judge (Kleinberg et al., 2017).

The purpose of this article is two-fold: first, it examines the more fundamental changes in knowledge production in criminal justice settings occurring due to over-reliance on the new epistemological transition – on knowledge supposedly generated without a priori theory. Second, it shows why automated predictive decision-making tools are often at variance with fundamental liberties and also with the established legal doctrines and concepts of criminal procedure law. Such tools discriminate against persons in lower income strata and less empowered sectors of the population (Eubanks, 2018; Harcourt, 2015a'; Noble, 2018; O'Neil, 2016). The following sections of this article map the novelties of automated decision-making tools – defined as tools utilizing advances in big data, algorithms, machine learning and ''AI – in criminal justice settings.

Existing research on algorithmic justice

Legal scholars have analysed the risks stemming from automated decision-making and prediction in several domains. For example, in examining automation applied to credit

scoring, Citron and Pasquale (2014) identified the opacity of such automation, arbitrary assessments and disparate impacts. Computer scientists have warned against the reductionist caveats of big data and 'exascale' computers: 'computers cannot replace archetypal forms of thinking founded on Philosophy and Mathematics' (Koumoutsakos, 2017). Research on the role of automation in exacerbating discrimination in search engines (Noble, 2018) and social security systems (Eubanks, 2018) showed how the new tools are dividing societies along racial lines. However, research on the impacts of automated decision-making systems on fundamental liberties in the justice sector has not gained enough critical reflection.

Legal analysis of the use of algorithms and AI in the criminal justice sector have focused on the scant case law on the topic, especially on the landmark decision in *Loomis v. Wisconsin* (2016) in the USA, in which the Supreme Court of Wisconsin considered the legality of using the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) risk assessment software in criminal sentencing. The *Loomis* case raised two related legal issues, that is, due process and equal protection (Kehl and Kessler, 2017). In the same year, ProPublica's report on 'Machine bias' (Angwin et al., 2016) triggered fierce debates on how algorithms embed existing biases and perpetuate discrimination. By analysing the diverging assessments of the outcomes of the COMPAS algorithm – by ProPublica, on the one hand, and the algorithm producer Northpointe, on the other – scholars have identified a more fundamental difference in the perception of what it means for an algorithm to be 'successful' (Flores et al., 2016). Probation algorithm advocates and critics are both correct, given their respective terms, claims the mathematician Chouldechova (2016), since they pursue competing notions of fairness.

Along with these concerns, scholars dubbed the trend to use automated decision-making tools in the justice sector 'automated justice' (Marks et al., 2015), which causes discrimination and infringes the due process of law guaranteed by constitutions and criminal procedure codes. Similarly, Hannah-Moffat (2019) warns that big data technologies can violate constitutional protections, produce a false sense of security and be exploited for commercial gain. 'There can and will be unintended effects, and more research will be needed to explore unresolved questions about privacy, data ownership, implementation capacities, fairness, ethics, applications of algorithms' (Hannah-Moffat, 2019: 466).

According to Robinson (2018), three structural challenges can arise whenever a law or public policy contemplates adopting predictive analytics in criminal justice: (1) what matters versus what the data measure; (2) current goals versus historical patterns; and (3) public authority versus private expertise.

Based on research on the automation of traffic fines, O'Malley (2010) showed how a new form of 'simulated justice' has emerged. By complying with automated fining regimes – whereby traffic regulation violations are automatically detected and processed and fines enforced – we buy our freedom from the disciplinary apparatuses but we succumb to its 'simulated' counterpart at the same time, claims O'Malley. The judicial-disciplinary apparatus still exists as a sector of justice, but a large portion of justice is – concerning the 'volumes of governance' – 'simulated': 'the majority of justice is delivered by machines as 'a matter of one machine "talking" to another' (O'Malley, 2010).

Scholars are also critical of the *purposes* of the use of machine learning in the criminal justice domain. Barabas et al. (2017) claim that the use of actuarial risk assessments is not a neutral way to counteract the implicit bias and to increase the fairness of decisions in the criminal justice system. The core ethical concern is not the fact that the statistical techniques underlying actuarial risk assessments might reproduce existing patterns of discrimination and the historical biases that the data reflect, but rather the *purpose* of risk assessments (Barabas et al., 2017).

One strand of critique of automated decision-making tools emphasizes the limited role of risk management in the multiple purposes of sentencing. Namely, minimizing the risk of recidivism is only one of the several goals in determining a sentence, along with equally important considerations such as ensuring accountability for past criminal behaviour. The role of automated risk assessment in the criminal justice system should not be central (Kehl and Kessler, 2017). It is not even clear whether longer sentences for riskier prisoners might decrease the risk of recidivism. On the contrary, prison leads to a slight increase in recidivism owing to the conditions in prisons, such as the harsh emotional climate and psychologically destructive nature of prisonization (Gendreau et al., 1999).

Although cataloguing risks and alluding to abstract notions of 'fairness' and 'transparency' by advocating an 'ethics-compliant approach' is an essential first step in addressing the new challenges of big data and algorithms, there are more fundamental shifts in criminal justice systems that cannot be reduced to mere infringements of individual rights. What do big data, algorithms and machine learning promise to deliver in criminal justice and how do they challenge the existing knowledge about crime and ideas as to the appropriate response to criminality?

New language and new knowledge production

Today, law enforcement agencies no longer operate merely in the paradigm of the ex post facto punishment system but also employ ex ante preventive measures (Kerr and Earle, 2013). The new preventive measures stand in sharp contrast to various legal concepts that have evolved over the decades to limit the executive power to legal procedures. New concepts and tools are being invented to understand crime (knowledge production) and to act upon this knowledge (crime control policy). These include, for instance, 'meaning extraction', 'sentiment analysis', 'opinion mining' and 'computational treatment of subjectivity', all of which are rearranging and blurring the boundaries in the security and crime control domain.

For instance, the post-9/11 German anti-terrorist legislation invented the notion of a 'sleeping terrorist', based on the presumed personal psychological states and other circumstances of known terrorists. Such preventive 'algorithmic' identification of a target of state surveillance criminalized remotely relevant antecedents of crime and even just mental states. The notion presents a form of automated intelligence work whereby the preventive screening of Muslims was conducted because of the 'general threat situation' in the post-9/11 period in order to identify unrecognized Muslim fundamentalists as 'terrorist sleepers'. The idea of preventive screening as a dragnet type of investigation was that the known characteristics of terrorists are enough to infer the criminality of future criminals. The legislation was challenged before the Federal

Constitutional Court, which limited the possibility of such screening (4 April 2006) by claiming that the general threat situation that existed after the terrorist attacks on the World Trade Center on 9/11 was *not* sufficient to justify that the authorities start such investigations (Decision 1BvR 518/02). The legislation blurred the start and finish of the criminal procedure process. The start of a criminal procedure is the focal point when a person becomes a suspect, when a suspect is granted rights, for example Miranda rights, that are aimed at remedying the inherent imbalances of power between the suspect and the state. The start of the criminal procedure became indefinite and indistinct (see Marks, et al., 2015) and it was no longer clear when a person 'transformed' into a suspect with all the attendant rights.

Similarly, the notion of a 'person of interest' blurs the existing concepts of a 'suspect'. In Slovenia, the police created a Twitter analysis tool during the so-called 'Occupy Movement' to track 'persons of interest' and for *sentiment analysis*. The details of the operation remain a secret because the police argued that the activity falls under the publicly inaccessible regime of 'police tactics and methods'. With such reorientation, the police were chasing the imagined indefinable future. They were operating in a scenario that concerned not only what the person might have committed, but what the person might become (see Lyon, 2014) – the focus changed from past actions to potential future personal states and circumstances.

The 'smart city' paradigm is generating similar inventions. In the Eindhoven Living Lab, consumers are tracked for the purpose of monitoring and learning about their spending patterns. The insights are then used to nudge people into spending more (Galič, 2018). The Living Lab works on the idea of adjusting the settings of the immediate sensory environment, such as changing the music or the street lighting. Although security is not the primary objective of the Living Lab, it is a by-product of totalizing consumer surveillance. The Lab treats consumers like Pavlov's dogs, because the Lab checks for 'escalated behaviour' to arrive at big-data-generated determinations of when to take action. The notion of escalated behaviour includes yelling, verbal or otherwise aggressive conduct or signs that a person has lost self-control (de Kort, 2014). Such behaviour is targeted in order for it to be defused (Galič, 2018). The goals of the intervention are then to induce subtle changes, such as 'a decrease in the excitation level'.' This case shows how the threshold for intervention in the name of security drops: interventions are made as soon as the behaviour is considered 'escalated'.

To conclude, the described novel concepts, that is, 'terrorist sleeper', and so on, are producing new knowledge about crime and providing new thresholds and justifications to act upon this knowledge. The new mathematical language serves security purposes well, writes Amoore (2014). However, the new concepts, such as 'meaning extraction', 'sentiment analysis' and 'opinion mining', are blurring the boundaries in the security and crime control domain. The concepts of a 'suspect', 'accused' or 'convicted' person serve as regulators of state powers. Similarly, the existing standards of proof serve as thresholds for ever more interference in a person's liberties. But these new concepts and mathematical language no longer sufficiently confine agencies or prevent abuses of power. The new mathematical language is helping to tear down the hitherto respected walls of criminal procedure rules.

Risks and pitfalls of algorithmic justice

Domain transfer – what gets lost in translation?

What are the specific risks stemming from the use of big data, algorithms and machine learning in the criminal justice domain?

The statistical modelling used in criminal justice originates from other domains, such as earthquake prediction (for example Predpol), where ground-truth data are more reliable and the acceptability of false predictions is different. Optimization goals may differ across domains. For instance, online marketers want to capture users' attention as much as possible and would rather send more – albeit irrelevant – products. Such 'false-positive' rankings in marketing are merely nuisances. In predictive policing, optimization on the side of 'false positives' has significantly more profound consequences. Treating innocent individuals as criminals severely interferes with fundamental liberties.

Concerning data, first of all, criminality – by default – is never fully reported. The dark figure of crime is a 'black box' that can never be properly encompassed by algorithms. The future is then calculated from already selected facts about facts. Predictions can be more accurate in cases where 'reality' does not change dramatically and where the data collected reflect 'reality' as closely as possible. Second, crime is a normative phenomenon, that is, it depends on human values, which change over time and place. Algorithmic calculations can thus never be accurately calibrated given the initial and changing set of facts or 'reality'.

Criminal procedure codes are a result of already realized balancing acts as to what should be optimized. The existing legal doctrines in constitutional and criminal law already embody decisions regarding the strength of the competing values, such as efficiency and fairness. For instance, in a democratic, liberal response to crime, it is better to let 10 criminals walk free than to sentence one innocent person – this is the litmus test of authoritarian and democratic political systems, and tech-savvy experts are now negotiating these balancing acts.

Such balancing acts are fairness trade-offs: what to optimize for in which domain is different. Discussion of the fairness of the COMPAS probation algorithm shows how the discussion unfolded on different levels. Whereas ProPublica claimed that COMPAS was biased towards black defendants by assigning them higher risk scores, the developer of the algorithm, Northpointe (now Equivant), argued that in fact the algorithm directly reflected past data, according to which blacks more often committed a crime upon being released (Spielkamp, 2017). ProPublica compared the false positives of both groups: blacks were rated a higher risk but re-offended less frequently, whereas white defendants were rated a lower risk and re-offended more frequently. In other words, a disproportionate number of black defendants were 'false positives': they were classified by COMPAS as high risk but subsequently were not charged with another crime. These differences in scores – the higher false-positive rates for blacks – amounted to unjust discrimination. On the other hand, Northpointe (now Equivant) argued that COMPAS was equally good at predicting whether a white or black defendant classified as high risk would re-offend (an example of a concept called 'predictive parity'). The two camps seem to perceive fairness differently. There are a few fairness criteria that can be used and these 'cannot all be simultaneously satisfied when recidivism prevalence differs across groups'

(Chouldechova, 2016). These then are the fairness trade-offs: 'Predictive parity, equal false-positive error rates, and equal false-negative error rates are all ways of being "fair", but are statistically impossible to reconcile if there are differences across two groups – such as the rates at which white and black people are being rearrested' (Courtland, 2018).

There are competing notions of fairness and predictive accuracy, which means a simple transposition of methods from one domain (for example, earthquake prediction to sentencing) can lead to distorted views of what algorithms calculate. The current transfer of statistical modelling in the criminal justice domain neglects the balance between competing values or, at best, re-negotiates competing values without adequate social consensus.

A similar refocusing was carried out by Barabas et al. (2017), who argued 'that a core ethical debate surrounding the use of regression in risk assessments is not simply one of bias or accuracy. Rather, it's one of purpose.' They are critical of the use of machine learning for predicting crime; instead of risk scores, they recommend 'risk mitigation' to understand the social, structural and psychological drivers of crime.

Building databases

Databases and algorithms are human artefacts. 'Models are opinions embedded in mathematics [and] reflect goals and ideology' (O'Neil, 2016: 21). In creating a statistical model, choices need to be made as to what is essential (and what is not) because the model cannot include all the nuances of the human condition. There are trustworthy models, but these are based on the high-quality data that are fed into the model continuously as conditions change. Constant feedback and testing millions of samples prevent the self-perpetuation of the model.

Defining the goals of the problem and deciding what training data to collect and how to label the data, among other matters, are often done with the best intentions but with little awareness of their potentially harmful downstream effects. Such awareness is even more needed in the crime control domain, where the 'training data' on human behaviour used to train algorithms are of varying quality.

On a general level, there are at least two challenges in algorithmic design and application. First, compiling a database and creating algorithms for prediction always require decisions that are made by humans. 'It is a human process that includes several stages, involving decisions by developers and managers. The statistical method is only part of the process for developing the final rules used for prediction, classification or decisions' (European Union Agency for Fundamental Rights, 2018). Second, algorithms may also take an unpredictable path in reaching their objectives despite the good intentions of their creators.

Part of the first question relates to the data and part to the algorithms. In relation to data, the question concerns how data are collected, cleaned and prepared. There is an abundance of data in one part of the criminal justice sector; for example, the police collect vast amounts of data of varying quality. These data depend on victims reporting incidents, but these reports can be more or less reliable and sometimes they do not even exist. For some types of crime, such as cybercrime, the denial of victimization is often the norm (Wall, 2007). Financial and banking crime remains underreported, even less prosecuted or concluded by a court decision. There is a base-rate problem in such types

of crime, which hinders good modelling. The other problem with white-collar crime is that it flies below the radar of predictive policing software.

Furthermore, in the case of poor data, it does not help if the data-crunching machine receives more data. If the data are of poor quality, then 'garbage in garbage out'.

The process of preparing data in criminal justice is inherently political – someone needs to generate, protect and interpret data (Gitelman, 2013).

Building algorithms

The second part of the first question relates to building algorithms. If 'models are opinions embedded in mathematics' (O'Neil, 2016: 21), legal professionals working in criminal justice systems need to be more technically literate and able to pose relevant questions to excavate values embedded in calculations. Although scholars (Pasquale, 2015) and policy-makers alike (European Union Agency for Fundamental Rights, 2018) have been vocal in calling for the transparency of algorithms, the goal of enabling detection and the possibility of rectifying discriminatory applications misses the crucial point of machine learning and the methods of neural networks. These methods are black-boxed by default. Demanding transparency without the possibility of explainability remains a shallow demand (Veale and Edwards, 2018). However, third-party audits of algorithms may provide certification and auditing schemes and increase trust in algorithms in terms of their fairness.

Moreover, building better algorithms is not the solution – thinking about problems only in a narrow mathematical framework is reductionist. A tool might be good at predicting who will fail to appear in court, for example, but it might be better to ask why people do not appear and, perhaps, to devise interventions, such as text reminders or transportation assistance, that might improve appearance rates. 'What these tools often do is help us tinker around the edges, but what we need is wholesale change', claims Vincent Southerland (Courtland, 2018). Or, as Princeton computer scientist Narayanan argues, 'It's not enough to ask if code executes correctly. We also need to ask if it makes society better or worse' (Hulette, 2017).

Runaway algorithms

The second question deals with the ways algorithms take unpredictable paths in reaching their objectives. The functioning of an algorithm is not neutral, but instead reflects choices about data, connections, inferences, interpretations and inclusion thresholds that advance a specific purpose (Dwork and Mullingan, 2013). In criminal justice settings, algorithms calculate predictions via various proxies. For instance, it is not possible to calculate re-offending rates directly, but it is possible through proxies such as age and prior convictions. Probation algorithms can rely only on measurable proxies, such as the frequency of being arrested. In doing so, algorithms must not include prohibited criteria, such as race, ethnic background or sexual preference, and actuarial instruments have evolved away from race as an explicit predictor. However, the analysis of existing probation algorithms shows how prohibited criteria are not directly part of calculations but are still encompassed through proxies (Harcourt, 2015b). But, even in cases of such

'runaway algorithms',' the person who is to bear the effects of a particular decision will often not even be aware of such discrimination. Harcourt shows how two trends in assessing future risk have a significant race-related impact: (1) a general reduction in the number of predictive factors used, and (2) an increased focus on prior criminal history, which forms part of the sentencing guidelines of jurisdictions in the USA (Harcourt, 2015b). Criminal history is among the strongest predictors of arrest and a proxy for race (Harcourt, 2015b). In other words, heavy reliance on the offender's criminal history in sentencing contributes more to racial disparities in incarceration than dependence on other robust risk factors less bound to race. For instance, the weight of prior conviction records is apparent in the case of Minnesota, which has one of the highest black/white incarceration ratios. As Frase (2009) explains, disparities are substantially greater in prison sentences imposed and prison populations than regarding arrest and conviction. The primary reason is the heavy weight that sentencing guidelines place on offenders' prior conviction records.

Blurring probability, causality and certainty

Risk assessments yield probabilities, not certainties. They measure correlations and not causations. Although they may be very good at finding correlations, they cannot judge whether or not the correlations are real or ridiculous (Rosenberg, 2016). Correlations thus have to be examined by a human; they have to be ascribed a meaning. Just as biases can slip into the design of a statistical model, they can also slip into the interpretation – the interpreter imposes political orientations, values and framings when interpreting results. Typically, data scientists have to work alongside a domain-specific scientist to ascribe meaning to the calculated results.

For instance, at what probability of recidivism should a prisoner be granted parole? Whether this threshold ought to be a 40 percent or an 80 percent risk of recidivism is an inherently 'political' decision based on the social, cultural and economic conditions of the given society. Varying social conditions are relevant; for example, crowded prisons may lead to a mild parole policy, whereas governance through fear or strong prison industry interests that influence decision-making may lead to a harsher policy. Decision-making based on the expected risk of recidivism can be more precise if certainty as to the predicted rate of offending is added – with the introduction of the Bayesian optimization method (Marchant, 2018). Models that are based not solely on the expected risk rate but also on the *certainty* of the predicted risk rate can be more informative.

However, although a human decision is more precise because the expected values can be compared according to the certainty of those values, it still needs to be made with regard to what is too much uncertainty to ground the probation decision on the expected risk rate. Quantifying uncertainty by analysing the probability distribution around the estimated risk thus enables comparisons between different risk assessments based on different models, but it does not eliminate the need to draw a line, that is, to decide how much weight should be given to the certainty of the expected risk rate. There is still a need to ascribe meaning to the probability and 'translate' it in the context of the judicial setting.

Human or machine bias

Human decision-making in criminal justice settings is then often flawed, and stereotypical arguments and prohibited criteria, such as race, sexual preference or ethnic origin, often creep into judgments. Research on biases in probation decisions, for instance, has demonstrated how judges are more likely to decide 'by default' and refuse probation towards the end of sessions, whereas immediately after having a meal they are more inclined to grant parole (Danziger et al., 2011). Can algorithms help prevent such embarrassingly disproportionate and often arbitrary courtroom decisions?

The first answer can be that no bias is desirable and a computer can offer such an unbiased choice architecture. But algorithms are 'fed with' data that is not 'clean' of social, cultural and economic circumstances. Even formalized synthetic concepts, such as averages, standard deviations, probability, identical categories or 'equivalences', correlation, regression, sampling, are 'the result of a historical gestation punctuated by hesitations, retranslations, and conflicting interpretations' (Desrosières, 2002: 2). De-biasing would seem to be needed.

However, cleaning data of such historical and cultural baggage and dispositions may not be either possible or even desirable. In analysing language, Caliskan et al. (2017) claim that, first, natural language necessarily contains human biases. The training of machines on language corpora entails that AI will inevitably absorb the existing biases in a given society (Caliskan et al., 2017). Second, they claim that de-biasing may be possible, but this would lead to 'fairness through blindness' because prejudice can creep back in through proxies (Caliskan et al., 2017). Machine biases are then an inevitable consequence also in algorithmic decision-making systems, and such biases have been documented in criminal justice settings (see Angwin et al., 2016). Moreover, big data often suffer from other biases, such as confirmation bias, outcome bias, blind-spot bias, the availability heuristic, clustering illusions and bandwagon effects.

The second dilemma concerns whether de-biasing is even desirable. If machine decision-making is still the preferred option, then engineers building statistical models should carry out a de-biasing procedure. However, constitutions and criminal procedure codes have all been adopted through a democratic legislative process that distilled the prevailing societal interests, values, and so on of the given society. Although it is not difficult to be critical of this process of 'distillation', which is itself extremely partial or even 'captured' by the 'rich and powerful', it is still relatively open to scrutiny in comparison with a process of de-biasing conducted behind closed doors by computer scientists in a laboratory. The point is that de-biasing entails that inherently political decisions are to be made, for example as to what is merely gendered and what is sexist language that needs to be 'cleaned', or what is hate speech targeting minorities and which differential treatment should be deemed to be discriminatory. In a machine-based utopia, such decisions would thereby be relegated to the experts of the computer science elite. In this sense, de-biasing is not even desirable.

Do we as a society prefer human bias over machine bias? Do we want to invest in human decision-makers or in computerized substitutes?

Judges already carry out risk assessments on a daily basis, for example when deciding on the probability of recidivism. This process always subsumes human experience,

culture and even biases. They may work against or in favour of a given suspect. Human empathy and other personal qualities that form part of flourishing human societies are in fact types of bias that overreach statistically measurable 'equality'. Such decision-making can be more tailored to the needs and expectations of the parties to a particular judicial case (Plesničar and Šugman Stubbs, 2018). Should not such personal traits and skills or emotional intelligence be actively supported and nurtured in judicial settings? Bernard Harcourt (2006) explains how contemporary American politics has continually privileged policing and punishment while marginalizing the welfare state and its support for the arts and the commons. In the light of such developments, increasing the quality of human over computerized judgement should be the desired goal.

Judicial evolution and decision-making loops

Proponents of self-learning models claim that what is needed is the patience to make things 'really' work: 'smarter' algorithms have to be written, and algorithms should monitor existing algorithms ad infinitum (Reynolds, 2017). The programmes will be provided with feedback in terms of rewards and punishments, and they will automatically progress and navigate the space of the given problem – which is known as 'reinforcement learning'. The logics become caught up in escalating and circular decision-making loops ad infinitum.

However, this insatiable demand reveals itself as the ideology of big data. The argument exposes an apologetic gesture: what is needed is patience and the motivation to make things work. The argument claims that technology is not yet sufficiently perfected but that self-learning machines may remedy the current deficiencies. The inaccuracies resulting from algorithmic calculations are then also deficiencies in their own right. This argument neglects the fact that social life is always more picturesque and users of the systems will have to incorporate new parameters of criminal cases on a regular basis.

Judicial precedents by default started as outliers. Newly occurring circumstances receive more weight and completely change the reasoning that prevailed in former precedents. However, if the logic becomes caught up in escalating and circular decision-making loops, then the new circumstances of a case will never change the existing reasoning. Legal evolution may thus be severely hindered.

Subjectivity and the case-specific narrative

More than a decade ago Franko Aas (2005) identified how biometric technology used in criminal justice settings had been changing the focus from 'narrative' to supposedly more 'objective' and unbiased 'computer databases'. Franko was critical of the transition and characterized it as going 'from narrative to database'. Instead of the narrative, she furthermore claimed, the body has come to be regarded as a source of unprecedented accuracy because 'the body does not lie' (Franko Aas, 2006).

With AI tools, supposedly objective value-free 'hard-core' science' is even further replacing subjectivity and the case-specific narrative. Today, criminal justice systems are on the brink of taking yet another step in the same direction – from the database towards automated algorithmic-based decision-making. Whereas the database was still a static

pool of knowledge that needed some form of intervention by the decision-makers, for example so as to at least include it in the context of a specific legal procedure, automated decision-making tools adopt a decision independently. The narrative, typically of a risky individual, is not to be trusted and only the body itself can provide a reliable confession. For example, a DNA sample can reveal the home country of a refugee and not their explanation of 'their origin. Today, moreover, even decision-makers are not to be believed. Harcourt (2015b) succinctly comments that, for judicial practitioners, the use of automated recommendation tools is acceptable despite the curtailment of the discretion of the practitioners, whose judgement is often viewed negatively and associated with errors. Relying on automated systems lends an aura of objectivity to the final decision and diffuses the burden of responsibility for their choices (Harcourt, 2015a).

In the transition towards the complete de-subjectivation in the decision-making process, a sort of erasure of subjectivity is at work (see Marks et al., 2015). The participants in legal proceedings are perceived as the problem and technology as the solution in the post-human quest: 'The very essence of what it means to be human is treated less as a feature than a bug' (Rushkoff, 2018). It is 'a quest to transcend all that is human: the body, interdependence, compassion, vulnerability, and complexity.' The transhumanist vision too easily reduces all of reality to data, concluding that 'humans are nothing but information-processing objects' (Spiekermann et al., 2017). If perceived in such a manner, machines can easily replace humans, which can then be tuned to serve the powerful and uphold the status quo in the distribution of social power.

Psychological research into human–computer interfaces also demonstrates that automated devices can fundamentally change how people approach their work, which in turn leads to new kinds of errors (Parasuraman and Riley, 1997). A belief in the superior judgement of automated aids leads to errors of commission – when people follow an automated directive despite contradictory information from other more reliable sources of information because they fail to either check or discount that information (Skitka et al., 2000). It is a new type of automation bias, where the automated decisions are rated more positively than neutral (Dzindolet et al., 2003). Even in the existing stage of non-binding semi-automated decision-making systems, the process of arriving at a decision changes. The perception of accountability for the final decision changes too. The decision-makers will be inclined to tweak their own estimates of risk to match the model's (see Eubanks, 2018).

Human rights implications

Scholars and policy-makers have dealt extensively with the impacts of predictive modelling methods on the principle of non-discrimination and equality (European Union Agency for Fundamental Rights, 2018) and the implications for personal data protection (for example, Brkan, 2017). Although these are pertinent, the impacts of predictive modelling methods in criminal justice settings go well beyond these two concerns. Let us first examine the impact on the principle of equality.

Over-policing as an outcome of the use of predictive policing software – when the police patrol areas with more crime, which in turn amplifies the need to police these areas already policed – has been shown to be a prime example of the 'vicious circle'

effect in the crime control domain (Ferguson, 2017). However, under-policing is even more critical. First, the police do not scrutinize some areas as much as others, which leads to a disproportionate feeling and experience of justice. Second, some types of crime are more likely to be prosecuted than others. The central principle of legality – the requirement that all crimes be prosecuted ex officio, as opposed to the principle of opportunity – is thus not respected. The crimes more likely committed by the middle or upper classes then fly under the radar of criminal justice systems. Such crimes often become 're-labelled' as 'political scandals' or merely occupational 'risk'. Greater criminality in the banking system as a whole has been ignored (Monaghan and O'Flynn, 2013) and is increasingly regarded as part of the 'normal' part of financialized societies. This is not to claim that there are no options for the progressive use of big data and AI in the 'fight' against 'crimes of the powerful'. For instance, the Slovenian Ministry of Finance's financial administration is using machine learning to detect tax-evasion schemes and tax fraud (Spielkamp, 2019), which could be directed towards the most powerful multinational enterprises and their manipulations with transfer pricing. However, predictive policing software has not been used to such ends to the same extent as for policing 'the poor'.

Algorithmic decision-making systems may collide with several other fundamental liberties. Similar to 'redlining', the 'sleeping terrorist' concept (as discussed above) infringes upon the presumption of innocence. The mere probability of a match between the attributes of known terrorists and a 'sleeping' one directs the watchful eye of the state to the individual. Similarly, there is a collision with the principle of legality, that is *lex certa*, which requires the legislature to define a criminal offence in a substantially specific manner.

Standards of proof are thresholds for state interventions into individual rights. However, the new language of mathematics, which helps define new categories such as a 'person of interest', re-directs law enforcement agency activities towards individuals not yet considered a 'suspect'. The new notions being invented contravene the established standards of proof in criminal procedure.

In the specific context of a criminal trial, several rights pertain to the principle of the equality of arms in judicial proceedings and the right to a fair trial. Derivative procedural rights, such as the right to cross-examine witnesses, should be interpreted so as to also encompass the right to examine the underlying rules of the risk-scoring methodology. In probation procedures, this right should entail ensuring that a convicted person has the possibility to question the modelling applied – from the data fed into the algorithm to the overall model design (see the *Loomis* case).

Systems of (semi-)automated decision-making should respect a certain set of rights pertaining to tribunals. That is, the principle of the natural court requires that the criteria determining which court (or a specific judge thereof) is competent to hear the case be clearly established in advance (the rule governing the allocation of cases to a particular judge within the competent court, thus preventing forum shopping), and the right to an independent and impartial tribunal.

Conclusion

Automated analysis of judicial decisions using AI is supposed to boost efficiency in the climate of the neoliberal turn and political pressure 'to achieve more with less'. In such

a context, automated decision-making in criminal justice proves itself to be an 'algorithmic avatar' of neoliberalism (Benbouzid, 2016; Desrosières, 2002).

There is a prevailing sentiment that the AI tools will vaporize biases and heuristics inherent in human judgement and reasoning, which will in turn increase the legitimacy of criminal justice agencies and confine infliction of punishment to 'pure' scientific method and 'reason'. We can trace the idea to confine criminal justice to 'pure' scientific method to Cesare Beccaria, who claimed in his seminal booklet *On Crimes and Punishments* ([1764] 1872) that criminal justice must be confined to reason. In the context of the post-enlightenment, this meant changing the rules of criminal procedure, such as prohibition of torture, and the rules of substantive criminal law, such as detaching crime from sin (Adler et al., 1998: 51). Beccaria laid down several crime policy principles pertaining to the criminal justice system, such as the principle of separation of powers, according to which judges may *only apply* the law (by following Montesquieu). His ideas helped humanize criminal procedure and legitimized the then newly emerging classes and modern state, when the feudal values of the Ancien Régime started to melt. Similarly, today we can observe how justifications for introducing AI into criminal justice again allude to 'reason' and 'objectivity' – as if judges are not reliable enough to make unbiased decisions. The introduction of AI into criminal justice settings also reinforces the social power of the new emerging digital elite, which capitalizes on the ideology of big data and algorithmic impartiality.

However, delivering justice with new tools – big data, algorithms and machine learning – does not lead to a de-biased digital bonanza. As the article shows, de-biasing of language and criminal justice data may not even be possible. Moreover, since de-biasing would override the democratic procedures designed to prevent abuses of the elite, it is not even desired.

If the decision-maker is biased, then the logical move is to eliminate the human from the decision-making loop. However, in reality, eliminating the human factor proved not to be the key to solving the problem. Automated systems further exacerbate the problems they were employed to improve (see Eubanks 2018: 80–1). This article shows how the trend towards de-subjectivation brought by technology may lead to unintended consequences, such as hindering legal evolution, reinforcing the 'eternal past', eliminating case-specific narratives and erasing subjectivity. Moreover, automation in the criminal justice system directly interferes with several constitutional liberties and rights of defence in the criminal procedure.

However, one should' not entirely neglect the value of automation in criminal justice settings. Automation can increase knowledge about the criminal justice system itself. If the structure of a specific judicial procedure leads to discriminatory and arbitrary outcomes, then the introduction of automated tools may improve justice. But the question of what to improve remains open: human decision-makers or computerized substitutes? We should learn from the current experiments showing how automation can be counterproductive. What advocates of the automated decision-making systems neglect is the importance of the ability to bend the rules and reinterpret them according to social circumstances (Eubanks, 2018).

If criminal justice procedures are being computerized, we are surely likely to ask ourselves 'what is being left out?' 'Judicial opinions are themselves works of literature

with rich hypertextual potential', claims Birkhold (2018). Also, we want empathetic judges (Plesničar and Šugman Stubbs, 2018) and not executory 'cold-blooded' machines. Criminal procedure has several conflicting goals. If the end goal of the criminal procedure is to find the historical truth, then Miranda rights and exclusionary rules are an absolute obstacle in the truth-discovering process. If its goal is to reconcile the parties and solve a dispute between an accused and the state, then these rules have an important place in criminal procedure. Their goal is to facilitate meaning and other values from the process, such as respect for the rule of law.

Finally, we must acknowledge the fact that, in criminal justice systems, some procedures should not be subjected to automatization (similarly in the context of policing, by Oswald et al., 2018). There is simply too high an impact upon society and upon the human rights of individuals for them to be influenced by a reduced human agency relegated to machines.

Acknowledgements

I am grateful to the reviewers for their comments. Special thanks are also owed to the participants at the colloquium 'Automated Justice: Algorithms, Big Data and Criminal Justice Systems', held at the Collegium Helveticum in Zürich, 20 April 2018, for their helpful comments and suggestions.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research leading to this article has received funding from the EURIAS Fellowship Programme (the European Commission Marie Skłodowska-Curie Actions – COFUND Programme – FP7) and the Slovenian Research Agency, research project 'Automated Justice: Social, Ethical and Legal Implications', no. J5-9347.

ORCID iD

Aleš Završnik https://orcid.org/0000-0002-4531-2740

Note

1. For the purposes of this article, the notions 'new tools' and 'automated decision-making tools' are used to encapsulate the terms 'big data', 'algorithms', 'machine learning' and 'artificial intelligence'. Each of these notions is highly contested and deserves its own discussion, but a detailed analysis of these tools is not central to identifying the social and legal implications of their use in criminal justice systems.

References

ACLU (American Civil Liberties Union) (2016) Statement of Concern About Predictive Policing by ACLU and 16 Civil Rights Privacy, Racial Justice, and Technology Organizations, 31 August. URL (accessed 4 September 2019): https://www.aclu.org/other/statement-concern-about-predictive-policing-aclu-and-16-civil-rights-privacy-racial-justice.

Adler F, Mueller GOW and Laufer WS (1998) *Criminology*, 3rd edn. Boston, MA: McGraw-Hill.

Aebi C (2015) Evaluation du système de prédiction de cambriolages résidentiels PRECOBS. MA thesis, École des Sciences Criminelles, Université de Lausanne, Switzerland.

Ambasna-Jones M (2015) The smart home and a data underclass. *The Guardian*, 3 August.

Amoore L (2014) Security and the incalculable. *Security Dialogue* 45(5): 423–439.
Angwin J, Larson J, Mattu S and Kirchner L (2016) Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*, 23 May. URL (accessed 4 September 2019): https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
Barabas C, Dinakar K, Ito J, Virza M and Zittrain J (2017) Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. Cornell University. arXiv:1712.08238 [cs, stat].
Barocas S, Hood S and Ziewitz M (2013) Governing algorithms: A provocation piece, 29 March. URL (accessed 4 September 2019): https://ssrn.com/abstract=2245322.
Beccaria CB ([1764] 1872) *An Essay on Crimes and Punishments*. With a Commentary by M. de Voltaire. A New Edition Corrected. Albany: W.C. Little & Co. URL (accessed 4 September 2019): https://oll.libertyfund.org/titles/2193.
Benbouzid B (2016) Who benefits from the crime? *Books&Ideas*, 31 October.
Birkhold MH (2018) Why do so many judges cite Jane Austen in legal decisions? *Electric Literature*, 24 April.
Brkan M (2017) Do algorithms rule the world? Algorithmic decision-making in the framework of the GDPR and beyond. SSRN Scholarly Paper, 1 August.
Caliskan A, Bryson JJ and Narayanan A (2017) Semantics derived automatically from language corpora contain human-like biases. *Science* 356(6334): 183–186.
Chouldechova A (2016) Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Cornell University. arXiv:1610.07524 [cs, stat].
Citron DK and Pasquale FA (2014) The scored society: Due process for automated predictions. *Washington Law Review* 89. University of Maryland Legal Studies Research Paper No. 2014–8: 1–34.
Cohen JE, Hoofnagle CJ, McGeveran W, Ohm P, Reidenberg JR, Richards NM, Thaw D and Willis LE (2015) Information privacy law scholars' brief in Spokeo, Inc. v. Robins, 4 September 2015. URL (accessed 4 September 2019): https://ssrn.com/abstract=2656482.
Courtland R (2018) Bias detectives: The researchers striving to make algorithms fair. *Nature* 558(7710): 357–360.
Danziger S, Levav J and Avnaim-Pesso L (2011) Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences* 108(17): 6889–6892.
De Kort Y (2014) Spotlight on aggression. Intelligent Lighting Institute, Technische Universiteit Eindhoven 1: 10–11.
Desrosières A (2002) *The Politics of Large Numbers: A History of Statistical Reasoning*. Cambridge, MA: Harvard University Press.
Dewan S (2015) Judges replacing conjecture with formula for bail. *New York Times*, 26 June.
Dwork C and Mullingan DK (2013) It's not privacy, and it's not fair. *Stanford Law Review* 66(35): 35–40.
Dzindolet MT, Peterson SA, Pomranky RA, Pierce LG and Beck HP (2003) The role of trust in automation reliance. *International Journal of Human-Computer Studies* 58(6): 697–718.
Eckhouse L (2017) Opinion | Big data may be reinforcing racial bias in the criminal justice system. *Washington Post*, 2 October.
Egbert S (2018) About discursive storylines and techno-fixes: The political framing of the implementation of predictive policing in Germany. *European Journal for Security Research* 3(2): 95–114.
Ekowo M and Palmer I (2016) *The Promise and Peril of Predictive Analytics in Higher Education*. New America, Policy Paper, 24 October.
Eubanks V (2018) *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St Martin's Press.

European Union Agency for Fundamental Rights (2018) #BigData: Discrimination in data-supported decision making. FRA Focus, 29 May.

Ferguson AG (2017) *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement*. New York: NYU Press.

Flores AW, Bechtel K and Lowenkamp CT (2016) False positives, false negatives, and false analyses: A rejoinder to 'Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks'. *Federal Probation Journal* 80(2): 9.

Franko Aas K (2005) *Sentencing in the Age of Information: From Faust to Macintosh*. London: Routledge-Cavendish.

Franko Aas K (2006) 'The body does not lie': Identity, risk and trust in technoculture. *Crime, Media, Culture* 2(2): 143–158.

Frase RS (2009) What explains persistent racial disproportionality in Minnesota's prison and jail populations? *Crime and Justice* 38(1): 201–280.

Galič M (2018) Živeči laboratoriji in veliko podatkovje v praksi: *Stratumseind 2.0* – diskusija živečega laboratorija na Nizozemskem. In: Završnik A (ed.) *Pravo v dobi velikega podatkovja*. Ljubljana: Institute of Criminology at the Faculty of Law.

Gefferie D (2018) The algorithmization of payments. Towards Data Science, 27 February. URL (accessed 4 September 2019): https://towardsdatascience.com/the-algorithmization-of-payments-how-algorithms-are-going-to-change-the-payments-industry-5dd3f266d4c3.

Gendreau P, Goggin C and Cullen FT (1999) The effects of prison sentences on recidivism. User Report: 1999-3. Department of the Solicitor General Canada, Ottawa, Ontario.

Gitelman L (ed.) (2013) *'Raw Data' Is an Oxymoron*. Cambridge, MA: MIT Press.

Hannah-Moffat K (2019) Algorithmic risk governance: Big data analytics, race and information activism in criminal justice debates. *Theoretical Criminology*. 23(4): 453–470.

Harcourt BE (2006) *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age*. Reprint edition. Chicago: University of Chicago Press.

Harcourt BE (2015a) *Exposed: Desire and Disobedience in the Digital Age*. Cambridge, MA: Harvard University Press.

Harcourt BE (2015b) Risk as a proxy for race. *Federal Sentencing Reporter* 27(4): 237–243.

Hulette D (2017) Patrolling the intersection of computers and people. Princeton University, Department of Computer Science, News, 2 October. URL (accessed 4 September 2019): https://www.cs.princeton.edu/news/patrolling-intersection-computers-and-people.

Joh EE (2017a) Feeding the machine: Policing, crime data, & algorithms. *William & Mary Bill of Rights Journal* 26(2): 287–302.

Joh EE (2017b) The undue influence of surveillance technology companies on policing. *New York University Law Review* 92: 101–130.

Kehl DL and Kessler SA (2017) Algorithms in the criminal justice system: Assessing the use of risk assessments in sentencing. URL (accessed 4 September 2019): http://nrs.harvard.edu/urn-3:HUL.InstRepos:33746041.

Kerr I and Earle J (2013) Prediction, preemption, presumption: How big data threatens big picture privacy. *Stanford Law Review Online* 66(65): 65–72.

Kleinberg J, Lakkaraju H, Leskovec J, Ludwig J and Mullainathan S (2017) *Human Decisions and Machine Predictions*. Working Paper 23180, National Bureau of Economic Research, Cambridge, MA.

Koumoutsakos P (2017) Computing . Data .. Science . . . Society: On connecting the dots. Talk at the Collegium Helveticum, Zurich, 9 March. URL (accessed 4 September 2019): https://www.cse-lab.ethz.ch/computing-data-science-society-on-connecting-the-dots/.

Kramer ADI, Guillory JE and Hancock JT (2014) Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences* 111(24): 8788–8790.

Lewis P and Hilder P (2018) Leaked: Cambridge Analytica's blueprint for Trump victory. *The Guardian*, 23 March.

Lyon D (2014) Surveillance, Snowden, and big data: Capacities, consequences, critique. *Big Data & Society* 1(2):1–13.

McGee S (2016) Rise of the billionaire robots: How algorithms have redefined hedge funds. *The Guardian*, 15 May.

Marchant R (2018) Bayesian techniques for modelling and decision-making in criminology and social sciences. Presentation at the conference 'Automated Justice: Algorithms, Big Data and Criminal Justice Systems', Collegium Helveticum, Zürich, 20 April. URL (accessed 4 September 2019): https://collegium.ethz.ch/wp-content/uploads/2018/01/180420_automated_justice.pdf.

Marks A, Bowling B and Keenan C (2015) Automatic justice? Technology, crime and social control. SSRN Scholarly Paper, 19 October. In: Brownsword R, Scotford E and Yeung K (eds) *The Oxford Handbook of the Law and Regulation of Technology*. Oxford University Press, forthcoming. Queen Mary School of Law Legal Studies Research Paper No. 211/2015; TLI Think! Paper 01/2015. URL (accessed 4 September 2019): https://ssrn.com/abstract=2676154.

Meek A (2015) Data could be the real draw of the internet of things – but for whom? *The Guardian*, 14 September.

Monaghan LF and O'Flynn M (2013) The Madoffization of society: A corrosive process in an age of fictitious capital. *Critical Sociology* 39(6): 869–887.

Morozov E (2013) *To Save Everything, Click Here: Technology, Solutionism, and the Urge to Fix Problems that Don't Exist*. London: Allen Lane.

Noble SU (2018) *Algorithms of Oppression: How Search Engines Reinforce Racism*, 1st edn. New York: NYU Press.

O'Hara D and Mason LR (2012) How bots are taking over the world. *The Guardian*, 30 March.

O'Malley P (2010) Simulated justice: Risk, money and telemetric policing. *British Journal of Criminology* 50(5): 795–807.

O'Neil C (2016) *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown.

Oswald M, Grace J, Urwin S, et al. (2018) Algorithmic risk assessment policing models: Lessons from the Durham HART model and 'experimental' proportionality. *Information & Communications Technology Law* 27(2): 223–250.

Parasuraman R and Riley V (1997) Humans and automation: Use, misuse, disuse, abuse. *Human Factors* 39(2): 230–253.

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

Plesničar MM and Šugman Stubbs K (2018) Subjectivity, algorithms and the courtroom. In: Završnik A (ed.) *Big Data, Crime and Social Control*. London: Routledge, Taylor & Francis Group, 154–176.

Polloni C (2015) Police prédictive: la tentation de ' dire quel sera le crime de demain'. *Rue89*, 27 May.

Reynolds M (2017) AI learns to write its own code by stealing from other programs. *New Scientist*, 25 February.

Robinson D and Koepke L (2016) Stuck in a pattern. Early evidence on 'predictive policing' and civil rights. *Upturn*.

Robinson DG (2018) The challenges of prediction: Lessons from criminal justice. *14 I/S: A Journal of Law and Policy for the Information Society* 151. URL (accessed 4 September 2019): https://ssrn.com/abstract=3054115.

Rosenberg J (2016) Only humans, not computers, can learn or predict. *TechCrunch*, 5 May.

Rouvroy A and Berns T (2013) Algorithmic governmentality and prospects of emancipation. *Réseaux* No. 177(1): 163–196.

Rushkoff D (2018) Future human: Survival of the richest. *OneZero*, Medium, 5 July.

Selingo J (2017) How colleges use big data to target the students they want. *The Atlantic*, November 4.
Skitka LJ, Mosier K and Burdick MD (2000) Accountability and automation bias. *International Journal of Human-Computer Studies* 52(4): 701–717.
Spiekermann S, Hampson P, Ess C, et al. (2017) The ghost of transhumanism & the sentience of existence. URL (accessed 4 September 2019): http://privacysurgeon.org/blog/wp-content/uploads/2017/07/Human-manifesto_26_short-1.pdf.
Spielkamp M (2017) Inspecting algorithms for bias. *MIT Technology Review*, 6 December.
Spielkamp M (ed.) (2019) *Automating Society*. Berlin: AW AlgorithmWatch.
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: Society, Security, Innovation and Ethics*. London: Palgrave Macmillan, 97–113.
Steiner C (2012) *Automate This*. New York: Penguin.
Veale M and Edwards L (2018) Clarity, surprises, and further questions in the Article 29 Working Party draft guidance on automated decision-making and profiling. *Computer Law and Security Review* 34(2): 398–404.
Wall DS 2007 *Cybercrime. The Transformation of Crime in the Information Age*. Cambridge, UK; Malden, MA: Polity Press.
Webb A (2013) Why data is the secret to successful dating. Visualised. *The Guardian*, 28 January.
Wilson D (2018) Algorithmic patrol: The futures of predictive policing. In: Završnik A (ed.) *Big Data, Crime and Social Control*. London, New York: Routledge, Taylor & Francis Group, 108–128.