Završnik A.♦

DOI: 10.1007/s12027-020-00602-0
UDC: 343.1

Criminal justice, artificial intelligence systems, and human rights

Abstract: The automation brought about by big data analytics, machine learning and artificial intelligence systems challenges us to reconsider fundamental questions of criminal justice. The article outlines the automation which has taken place in the criminal justice domain and answers the question of what is being automated and who is being replaced thereby. It then analyses encounters between artificial intelligence systems and the law, by considering case law and by analysing some of the human rights affected. The article concludes by offering some thoughts on proposed solutions for remedying the risks posed by artificial intelligence systems in the criminal justice domain.

Keywords: criminal justice; human rights; automation; algorithms; artificial intelligence; fair trial.

1 Introduction

With the advent of big data analytics, machine learning and artificial intelligence systems (henceforth 'AI systems')[1] [35], both the assessment of the risk of crime and the operation of criminal justice systems are becoming increasingly technologically sophisticated. While authors disagree whether these technologies represent a panacea for criminal justice systems - for example by reducing case backlogs - or will further exacerbate social divisions and endanger fundamental liberties, the two camps nevertheless agree that such new technologies have important consequences for criminal justice systems. The automation brought about by AI systems challenges us to take a step back and reconsider fundamental questions of criminal justice: What does the *explanation* of the grounds of a judgment mean? When is the process of adopting a judicial decision *transparent*? Who should be *accountable* for (semi-) automated decisions and how should *responsibility* be allocated within the chain of actors when the final decision is facilitated by the use of AI? What is a *fair trial*? And is the *due process of law* denied to the accused when AI systems are used at some stage of the criminal procedure?

The technical sophistication of the new AI systems used in decision-making processes in criminal justice settings often leads to a 'black box' effect [28]. The intermediate phases in the process of reaching a decision are by definition hidden from human oversight due to the technical complexity involved. For instance, multiple areas of applied machine learning show how new methods of unsupervised learning or active learning operate in a way that avoids human interven-

♦ **Završnik Aleš** - Prof. Dr. Institute of Criminology of the Faculty of Law, University of Ljubljana (Slovenia). E-mail: ales.zavrsnik@pf.uni-lj.si

[1] 1Among the several definitions of AI, the definition of the Commissioner for Human Rights is used herein:
'An AI system is a machine-based system that makes recommendations, predictions or decisions for a given set of objectives. It does so by: (i) utilising machine and/or human-based inputs to perceive real and/or virtual environments; (ii) abstracting such perceptions into models manually or automatically; and (iii) deriving outcomes from these models, whether by human or automated means, in the form of recommendations, predictions or decisions.'

tion. In the active approaches of machine learning used for natural language processing, for instance, the learning algorithm accesses a large corpus of unlabelled samples and, in a series of iterations, the algorithm selects some unlabelled samples and asks the human annotator for appropriate labels. The approach is called active as the algorithm decides what samples should be annotated by the human based on its current hypothesis. The core idea of active machine learning is to eliminate humans from the equation. Moreover, artificial neural networks (hereafter 'ANN') learn to perform tasks by considering examples, generally without being programmed with task-specific rules. As such, artificial neural networks can be extremely useful in multiple areas, such as computer vision, natural language processing, in geoscience for ocean modelling, or in cybersecurity for identifying and discriminating between legitimate activities and malicious ones. They do not demand labelled samples, e.g., in order to recognise cats in images or pedestrians in traffic, but can generate knowledge about what a cat looks like on their own. The operations in machine learning approaches are not transparent even for the researchers that built the systems and while this may not be problematic in many areas of applied machine learning, as the examples below show, AI systems must be transparent when used in judicial settings, where the explainability of decisions and the transparency of the reasoning are of significant - even civilizational - value. A decision-making process that lacks transparency and comprehensibility is not considered legitimate and non-autocratic. Due to the inherently opaque nature of these AI systems, the new tools used in criminal justice settings may thus be at variance with fundamental liberties.

Lawyers must be aware, moreover, of the supra-legal context and background rationale for implementing AI systems. While some reasons may be legitimate - e.g., when AI systems facilitate access to courts to individuals that might otherwise be left on the sidelines of justice -others may be disputable and require a wider social debate that can only be held outside the judicial system. Shrinking budgets, the decreased legitimacy of the judiciary, and an overload of cases may all lead to the implementation of new solutions that information technology companies are ready to offer to governments. However, proposals for outsourcing a public service to private sector providers will trigger (or should trigger) a major political discussion which must be held in more democratic fora.

Following on from this introductory contextualisation of the automation of criminal justice, this article proceeds to an outline of the automation of crime control and answers the questions of what is being automated and who is being replaced in crime control (*Part 2*). It then analyses encounters between artificial intelligence systems and the law through case law (*Part 3*) and through an analysis of some of the affected human rights (*Part 4*). Thereafter, it answers the following question: what human rights may be affected by AI systems and how? The article concludes by offering some thoughts on the proposed solutions to remedy the risks of AI systems in the criminal justice domain (*Part 5*).

2 How does automation change crime control?

2.1 The automation of policing

By means of CompStat (COMPuter STATistics, or COMParative STATistics), geospatial modelling for predicting future crime concentrations, or 'hot spots' [17], has developed into a paradigm of managerial policing employing Geographic Information Systems (GIS) to map crime. This has been advocated as a multi-layered dynamic approach to crime reduction, quality of life improvement, personnel and resource management, and not merely a computer programme. The idea is not solely to 'see crime' visually presented on a map, but rather to develop a comprehensive

managerial approach or a police management philosophy. As a 'human resource management tool', it involves 'weekly meetings where officers review recent metrics (crime reports, citations, and other data) and talk about how to improve those numbers' [7].

Compared to algorithmic prediction software, the CompStat system is calibrated less frequently. As a police officer from Santa Cruz (USA) reported: 'I'm looking at a map from last week and the whole assumption is that next week is like last week […]' [15]. CompStat relies more on humans to recognise patterns. Nevertheless, it incorporated for the first time the idea of seeing how crime evolves and focusing on 'the surface' and not the causes of crime. In this context, Siegel argues with respect to predictive analytics: 'We usually don't know about causation, and we often don't necessarily care […] the objective is more to predict than it is to understand the world […]. It just needs to work; prediction trumps explanation' [33]. In comparison to AI analytics, its limiting factor is the depth of the information and the related breadth of analysis. The amount of data is not the problem as agencies collect vast amounts of data every day; rather, the next challenge is the ability to pull operationally-relevant knowledge from the data collected.

Computational methods of 'predictive crime mapping' started to enter into crime control twelve years ago [30]. Predictive 'big data' policing instruments took another evolutionary step forward. First, advancements in AI promised to make sense of enormous amounts of data and to extract meaning from scattered data sets. Secondly, they represented a shift from being a decision support system to being a primary decisionmaker. Thirdly, they are aimed at the regulation of society at large and not only the fight against crime. (For an example of 'function-creep', see Singapore's 'total information awareness system programme'.) [18].

Police are using AI tools to penetrate deeply into the preparatory phase of crime which is yet to be committed, as well as to scrutinise already-committed crimes. With regard to ex-ante preventive measures, automation tools are supposed to excavate plotters of crimes which are yet to be committed from large amounts of data. Hence, a distinction is made between tools focusing on 'risky' individuals ('heat lists' - algorithm-generated lists identifying people most likely to commit a crime) [16] and tools focusing on risky places ('hot spot policing') [10].[2] With regard to the second, ex-post-facto uses of automation tools, there have been many success stories in the fight against human trafficking. In Europe, Interpol manages the International Child Sexual Exploitation Image Database (ICSE DB) to fight child sexual abuse. The database can facilitate the identification of victims and perpetrators through an analysis of, for instance, furniture and other mundane items in the background of abusive images - e.g., it matches carpets, curtains, furniture, and room accessories – or identifiable background noise in the video. Chatbots acting as real people are another advancement in the fight against grooming and webcam 'sex tourism'. In Europe, the Dutch children's rights organisation Terre des Hommes was the first NGO to combat webcam child 'sex tourism' by using a virtual character called 'Sweetie' [31]. The Sweetie avatar, posing as a ten-year-old Filipino girl, was used to identify offenders in chatrooms and online forums and operated by an agent of the organisation, whose goal was to gather information on individuals who contacted Sweetie and solicited webcam sex. Moreover, Terre des Hommes started engineering an AI system capable of depicting and acting as Sweetie without human intervention in order to not only identify persistent perpetrators but also to deter first-time offenders.

[2] For successful use of the latter, see Kadar, Maculan, Feuerriegel [20].

Some other research on preventing crime with the help of computer vision and pattern recognition with supervised machine learning seems outright dangerous [38]. Research on automated inferring of criminality from facial images based on still facial images of 1,856 real persons (half convicted) yielded the result that there are merely three features for predicting criminality: lip curvature, eye inner corner distance, and nose-mouth angle. The implicit assumptions of the researchers were that, first, the appearance of a person's face is a function of innate properties, i.e., the understanding that people have an immutable core. Secondly, that 'criminality' is an innate property of certain (groups of) people, which can be identified merely by analysing their faces. And thirdly, in the event of the first two assumptions being correct, that the criminal justice system is actually able to reliably determine such 'criminality', which implies that courts are (or perhaps should become) laboratories for the precise measurement of people's faces. The software promising to infer criminality from facial images [38] in fact illuminated some of the deep-rooted misconceptions about what crime is, and how it is defined, prosecuted, and adjudicated. The once ridiculed phrenology from the nineteenth century hence entered the twenty-first century in new clothes as 'algorithmic phrenology', which can legitimise deep-rooted implicit biases about crime [1].

Two researchers, Wu and Zhang, later admitted that they 'agree that the pungent word criminality should be put in quotation marks; a caveat about the possible biases in the input data should be issued. Taking a court conviction at its face value, i.e., as the 'ground truth' for machine learning, was indeed a serious oversight on our part' [3]. Nevertheless, their research revealed how, in the near future, further steps along the line of a corporal focus on crime control can reasonably be expected: from the analysis of walking patterns, posture, and facial recognition for identification purposes, to analysis of facial expressions and handwriting patterns for emotion recognition and insight into psychological states.

2.2 Automation in criminal courts

Courts use AI systems to assess the likelihood of recidivism and the likelihood of flight of those awaiting trial, or of offenders in bail and parole procedures. The most analysed and discussed examples come from the USA, which is also where most such software is currently being used [6]. The Arnold Foundation algorithm, which is being rolled out in 21 jurisdictions in the USA [6], uses 1.5 million criminal cases to predict defendants' behaviour in the pre-trial phase. Florida uses machine learning algorithms to set bail amounts [8].

A study of 1.36 million pre-trial detention cases showed that a computer could predict whether a suspect would flee or re-offend even better than a human judge [23].

However, while this data seems persuasive, it is important to consider that the decisions may in fact be less just. There will always be additional facts in a particular case that may be unique and go beyond the forty or so parameters considered by the algorithm in this study which might crucially determine the outcome of the deliberation process. There is thus the inevitable need for *ad infinitum* improvements. Moreover, the problem of selective labelling needs to be considered: we see results only regarding sub-groups that are analysed, only regarding people who have been released. The data that we see is data garnered based on our decisions as regards who to send to pre-trial detention. The researchers themselves pointed out that judges may have a broader set of preferences than the variables that the algorithm focuses on [23]. Finally, there is the question of what we want to achieve with AI systems, what we would like to 'optimise': decreasing crime is an important goal, but not the only goal in criminal justice. The fairness of the procedure is equally significant.

Several European countries are using automated decision-making systems for justice administration, especially for the allocation of cases to judges, e.g., in Georgia, Poland, Serbia, and Slovakia, and to other public officials, such as enforcement officers in Serbia [19]. However, while these cases are examples of indirect automated decision-making systems, they may still significantly affect the right to a fair trial. The study *'alGOVrithms - State of Play'* showed that none of the four countries using automated decision-making systems for case-allocation allows access to the algorithm and/or the source code [19]. Independent monitoring and auditing of automated decision-making systems is not possible, as the systems lack basic transparency. The main concern touches on how random these systems actually are, and whether they allow tinkering and can therefore be 'fooled'. What is even more worrying is that automated decision-making systems used for court management purposes are not transparent even for the judges themselves [19].

There are several other ongoing developments touching upon courtroom decisionmaking. In Estonia, the Ministry of Justice is financing a team to design a robot judge which could adjudicate small claims disputes of less than €7,000 [26]. In concept, the two parties will upload documents and other relevant information, and the AI will issue a decision against which an appeal with a human judge may be lodged.

2.3 Automation in prisons

New tools are used in various ways in the post-conviction stage. In prisons, AI is increasingly being used for the automation of security as well as for the rehabilitative aspect of prisonisation. A prison that houses some of China's most high-profile criminals is reportedly installing an AI network that will be able to recognise and track every prisoner around the clock and alert guards if anything seems out of place [39].

These systems are also used to ascertain the criminogenic needs of offenders that can be changed through treatment, and to monitor interventions in sentencing procedures [22]. In Finnish prisons the training of inmates encompasses AI training algorithms [25]. The inmates help to classify and answer simple questions in user studies, e.g., reviewing pieces of content collected from social media and from around the internet. The work is supposed to benefit Vainu, the company organising the prison work, while also providing prisoners with new job-related skills that could help them successfully re-enter society after they serve their sentences. Similarly, in England and Wales, the government has announced new funding for prisoners to be trained in coding as part of a £1.2m package to help under-represented groups get into such work [24]. 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 companions' for prisoners. While the 'companions' may alleviate some of the psychological stress for some prisoners, the focus on the 'surface' of the problem of solitary confinement conceals the debate about the aggravating harm of such confinement [32][3], and actually contributes to the legitimisation of solitary confinement penal policy. The shift from the real problem seems outrageous on its own.

[3] See the award-winning book on the harm caused by 'super-max' prisons, including the harm resulting from solitary confinement in Shalev [32].

3 Encounters between AI systems and the law

3.1 Due process of law and AI systems in the USA

In the American context, which is where most actual employment of AI systems in criminal justice has so far occurred, the decision on a risk assessment algorithm in the judgment in *Loomis v. Wisconsin* (2016), entitled Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), was a sobering one[4]. The COMPAS algorithm identified Loomis as an individual who presented a high risk to society due to a high risk of re-offending and the first instance court decided to refuse his request to be released on parole. In the appeal, the Supreme Court of Wisconsin decided that the recommendation from the COMPAS algorithm was not the sole grounds for refusing his request to be released on parole and hence the decision of the court did *not* violate Loomis's due process right. By confirming the constitutionality of the recommendation risk assessment algorithm, the Supreme Court of Wisconsin neglected the strength of the 'automation bias'.[5] By claiming that the lower court had the possibility to depart from the proposed algorithmic risk assessment, the Court ignored the social psychology and human-computer interaction research on the biases involved in *all* algorithmic decision-making systems, which show that once a high-tech tool offers a recommendation it becomes extremely burdensome for a human decision-maker to refute such a 'recommendation' [5]. Decision-makers regularly rate automated recommendations more positively than neutral despite being aware that such recommendations may be inaccurate, incomplete, or even wrong [13].

In the judgment in *Kansas v. Walls* (2017)[6], the Court of Appeals of the State of Kansas reached the opposite finding to *Loomis* and decided that the defendant must be allowed access to the complete diagnostic Level of Service Inventory-Revised (LSIR) assessment, which the court relied on in deciding what probation conditions to impose on him. The Court of Appeals decided that by refusing the defendant access to his LSI-R assessment the district court denied him the opportunity to challenge the accuracy of the information that the court was required to rely on in determining the conditions of his probation. By referring to the judgment in *Kansas v. Easterling*[7], the Court of Appeals decided the district court's failure to give the defendant a copy of the entire LSI-R deprived him of his constitutional right to procedural due process in the sentencing phase of his criminal proceedings.

3.2 Human rights compliance of AI systems in the EU

AI systems have a significant impact on human rights 'that engage state obligations vis-à-vis human rights.' [4][8] Since the data deluge has reached all social domains and algorithmic systems increasingly permeate various aspects of contemporary life [4][9], human rights compliance can no

[4] *State v Loomis* 881 N.W.2d 749 (Wis. 2016).

[5] In appealing to the United States Supreme Court, the Court denied the writ of certiorari, thus declining to hear the case, on 26 June 2017. *Loomis v Wisconsin*, 881 N.W.2d 749 (Wis. 2016), cert. denied, 137 S.Ct. 2290 (2017).

[6] *State of Kansas v. John Keith Walls*, 116,027, The Court of Appeals of the State of Kansas (2017).

[7] In *State v. Easterling* the Kansas Supreme Court held that a convicted defendant has a constitutional right to due process at sentencing, which requires 'the sentencing court to assure itself that the information upon which it relies to fix sentence is reliable and accurate, and [further requires] the sentencing court to ensure that the defendant have an effective opportunity to rebut the allegations likely to affect the sentence is fully applicable under these circumstances.' *State of Kansas v. David E. Easterling*, 289 Kan. 470, 481, 213 P.3d 418 (2009).

[8] *Committee of Experts on Human Rights Dimensions of Automated Data Processing and Different Forms of Artificial Intelligence* (*MSI-AUT*) [4], paras. 10 and 11 of the Preamble.

[9] See *Committee of Experts on Human Rights Dimensions of Automated Data Processing and Different Forms of Artificial Intelligence* (*MSI-AUT*) [4].

longer be seen as the exclusive domain of privacy and personal data protection [2][10] and non-discrimination and equality law [11]. Automated systems have been introduced to replace humans in the banking, insurance, education, and employment sectors, as well as in armed conflicts. They have influenced general elections and core democratic processes. Personal data protection regime is thus not sufficient to address all of the challenges as regards ensuring the compliance of AI systems with human rights. The human rights implications are then necessarily manifold, as the Committee of Experts on Internet Intermediaries (MSI-NET) at the

Council of Europe [34] rightly acknowledges. The human rights that may be impacted through the use of automated processing techniques and algorithms are: (1) the right to a fair trial and due process, (2) privacy and data protection, (3) freedom of expression, (4) freedom of assembly and association, (5) the right to an effective remedy, (6) the prohibition of discrimination, (7) social rights and access to public services, and (8) the right to free elections. Moreover, as fundamental freedoms are interdependent and interrelated, *all human rights* are potentially impacted by the use of algorithmic technologies, e.g., in education, social welfare, democracy, and judicial systems. The developments with the AI used in social systems and domains may even 'disrupt the very concept of human rights as protective shields against state interference.' [34, p. 32].

3.2.1 Equality and discrimination

Over-policing, as the most visible example of discrimination stemming from predictive policing programmes, occurs when the police patrol areas with more crime, which in turn amplifies the need to police areas already policed. It is a prime example of the 'vicious circle' effect of the use of machine learning in the crime control domain [12]. However, under-policing is even more critical, as the police do not scrutinize some areas as much as others, which leads to a disproportionate feeling and experience of justice. Some types of crime are then more likely to be prosecuted than others and the central principle of legality - the requirement that all crimes be prosecuted *ex officio*, as opposed to the principle of opportunity, by which prosecutors decide on prosecution at their own discretion, is thus not respected [40]. The use of predictive software to ascertain the treatment of perpetrators of white-collar crimes may neglect the fact that the enculturation of such offenders did not fail in any meaningful way [21]. On the contrary, such offenders are typically distinguished and respected citizens, e.g., CEOs, physicians, judges, or university professors. Critical criminologists have shown how the definition of crime itself—and even more so the prosecution of crime - is inherently political: law enforcement agencies are forced to make 'political' decisions about which crime to prosecute and investigate due to limited resources and personnel. They prioritise activities either explicitly or implicitly. Inequality in predictive policing then changes the perception of what counts as 'serious crime' in the first place. Hedge fund operations with sub-prime mortgages packaged in 'derivatives' are then reduced to unfortunate 'bad luck', despite impoverishing large parts of the population as their savings or equity evaporate. Predictive policing software has not been able to capture this important shift.

3.2.2 Personal data protection

With regard to the implications of the use of AI systems for personal data protection, the set of barriers to the adverse impacts of AI systems includes rights, such as the explicit consent of data subjects to the processing of their personal data, the data minimization principle, the principle of purpose limitation, and the set of rights relating to when automated decision-making is allowed. The

[10] See also: *Veale, Edwards* [37].

General Data Protection Directive (henceforth GDPR)[11] offers some points of reference here. In cases of automated processing, the data controller must implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, for instance by ensuring him or her the right to obtain human intervention on the part of the controller, to express his or her point of view, and to contest the decision (Art. 22, para. 3 of the GDPR). The GDPR includes the right of the data subject to receive 'meaningful information about the logic involved' in automated processing. (See Arts. 13, 14, and 15.)

Automated decisions, which produce adverse legal effects concerning the data subject or significantly affect him or her, are prohibited pursuant to Article 11 of the Law Enforcement Directive[12] (henceforth 'Law Enforcement Directive'), unless they are authorised by Union or Member State law, which also has to ensure appropriate safeguards for the rights and freedoms of the data subject. In line with the provisions of the Law Enforcement Directive, judicial decisions made entirely by an algorithmic tool can never be legal.

3.2.3 The right to a fair trial

The use of algorithms in criminal justice systems raises serious concerns with regard to Article 6 (concerning the right to a fair trial) of the European Convention on Human Rights[13] and Article 47 of the Charter[14], and the principle of the equality of arms and adversarial proceedings as established by the European Court of Human Rights [34, p. 11]. The fair trial standards contained in Article 6 of the ECHR guarantee the accused the right to participate effectively in the trial and include the presumption of innocence, the right to be informed promptly of the cause and nature of the accusation, the right to a fair hearing and the right to defend oneself in person.

The right to effective participation may be violated in a variety of different situations, ranging from poor acoustics in the courtroom[15] to preventing the accused from being present at the trial or from examining a witness testifying against him or her.[16] The latter is also one of the minimal guarantees of a fair trial contained in Art. 6, para. 3 and normally requires that all evidence against the accused be produced in his or her presence at a public hearing, which gives the defendant an effective opportunity to challenge the evidence against him or her.[17] The right to confrontation does not apply merely to witnesses, as the term is usually understood under national law, since it has an autonomous meaning in the Convention system that goes beyond its ordinary meaning and also includes experts, expert witnesses, and victims. In any case in which the deposition serves to a material degree as the basis for the conviction of the defendant, it constitutes evidence for the prosecution to which the Convention guarantees apply.[18] The right enshrined in Article 6(3)(d) can even be

[11] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L119/89.

[12] Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA [2016] OJ L119/89.

[13] Convention for the Protection of Human Rights and Fundamental Freedoms (European Convention on Human Rights) (hereinafter: ECHR).

[14] Charter of Fundamental Rights of the European Union [2012] OJ C 326 (hereinafter: Charter).

[15] *Stanford v the United Kingdom*, App no 16757/90 (ECtHR 11 April 1994).

[16] *ermi v Italy*, App no 18114/02 (ECtHR 18 October 2006).

[17] *Al-Khawaja and Tahery v the United Kingdom*, App no 26766/05, 2228/06 (ECtHR 15 December 2011); *Asani v the former Yugoslav Republic of Macedonia*, App no 27962/10 (ECtHR 1 February 2018).

[18] *Luca v Italy*, App no 33354/96 (ECtHR 27 February 2001), §41.

applied to documentary evidence[19] and computer files[20] relevant to the criminal accusations against the defendant. Therefore, in order to ensure effective participation in a trial, the defendant must also be able to challenge the algorithmic score that is the basis of his or her conviction.

However, the right to confrontation is not absolute and may be restricted if certain conditions are met. The traditional approach of the European Court of Human Rights was that the right to a fair trial was violated if a conviction was based either solely or to a decisive degree on an uncontested statement (the 'sole or decisive rule').[21] However, in *Al Khawaja and Tahery* the Court partially departed from its previous jurisprudence, stating that the admission of untested evidence will not automatically result in a breach of Article 6 (1): when assessing the overall fairness of a trial, the European Court of Human Rights has to consider whether it was necessary to admit such evidence and whether there were sufficient counterbalancing factors, including strong procedural safeguards.[22]

The problems posed by AI systems are very similar to those presented by anonymous witnesses or undisclosed documentary evidence as AI systems are opaque (as discussed in the introduction). At least some degree of disclosure is necessary in order to ensure a defendant has the opportunity to challenge the evidence against him or her and to counterbalance the burden of anonymity. Absent or anonymous witnesses, although not *per se* incompatible with the right to a fair trial, can only participate in a criminal procedure as a measure of last resort and under strict conditions ensuring that the defendant is not placed at a disadvantage. Such a rule should be applied to the use of AI systems used in criminal justice settings. A fair balance should be struck between the right to participate effectively in the trial, on the one hand, and the use of opaque AI systems designed to help judges arrive at more accurate assessments of the defendant's future conduct, on the other [29]. The right to cross-examine witnesses should be interpreted so as to also encompass the right to examine the data and the underlying rules of the risk-scoring methodology. In probation procedures, such a right should entail ensuring it is possible for a convicted person to question the model applied - from the data fed into the algorithm to the overall model design.

The use of algorithmic tools in criminal procedure could also violate some other aspects of the right to a fair trial, in particular the right to a randomly selected judge, the right to an independent and impartial tribunal, and the presumption of innocence.

3.2.4 Presumption of innocence

Besides affecting many dimensions of inequality, AI decision-making systems may collide with several other fundamental liberties. Similar to 'redlining', the 'sleeping terrorist' concept used in German anti-terrorist legislation infringed 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. O'Neil offers the illustrative example of the case of Robert McDaniel, a twenty-two-year-old high school student who received increased police attention due to a predictive programme's analysis of his social network and residence in a poor and dangerous neighbourhood: '… he was unlucky. He has been surrounded by crime, and many of his acquaintances have gotten caught up in it. And largely because of these circumstances—and not his own actions - he has been deemed dangerous. Now the police have their eye on him.' [27, p. 102-103].

[19] *Mirilashvili v Russia*, App no 6293/04 (ECtHR 11 December 2008), §158–159.
[20] *Georgios Papageorgiou v Greece*, App no 59506/00 (ECtHR 9 May 2003), §37.
[21] *Doorson v the Netherlands*, App no 20524/92 (ECtHR 26 March 1996).
[22] *Al-Khawaja and Tahery v the United Kingdom*, §152.

3.2.5 Effective remedy

Automated techniques and algorithms used for crime prevention purposes facilitate forms of secret surveillance and 'data-veillance' that are impossible for the affected individual to know about. The European Court of Human Rights has underlined that the absence of notification at any point undermines the effectiveness of remedies against such measures.[23] The right to an effective remedy implies the right to a reasoned and individual decision. Article 13 of the European Convention on Human Rights stipulates that everyone whose rights have been violated shall have an effective remedy before a national authority. The available remedy should be effective in practice and in law. As noted in the *Study on the Human Rights Dimensions of Automated Data Processing Techniques*:

Automated decision-making processes lend themselves to particular challenges for individuals' ability to obtain effective remedy. These include the opaqueness of the decision itself, its basis, and whether the individuals have consented to the use of their data in making this decision, or are even aware of the decision affecting them. The difficulty in assigning responsibility for the decision also complicates individuals' understanding of whom to turn to [to] address the decision. The nature of decisions being made automatic, without or with little human input, and with a primacy placed on efficiency rather than humancontextual thinking, means that there is an even larger burden on the organisations employing such systems to provide affected individuals with a way to obtain [a] remedy [34, p. 24].

3.2.6 Other rights

New notions in the pre-emptive crime paradigm, such as 'sleeping terrorist', are in collision with the principle of legality, i.e., *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 'person of interest', redirects law enforcement agency activities towards individuals not yet considered 'suspects'. The new notions being invented contravene the established standards of proof in criminal procedure.

AI systems should respect a certain set of rights pertaining to tribunals, i.e., the right to a randomly selected judge, which requires that the criteria determining which court - and which 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 (as discussed in the section on automation in criminal courts.

4 Discussion: toward solutions

How should we design human-rights-compliant AI systems that respect the rule of law standards of the 'analogue world'? The trend to 'algorithms' everything has raised the interest of policymakers. They share concern over the impact of algorithms on fundamental liberties and how to make 'algorithms accountable'. In the European context, the Council of Europe's European Commission for the Efficiency of Justice (CEPEJ) adopted the 'European Charter on the Use of AI in Judicial Systems' at the end of 2018 to mitigate the above-mentioned risks specifically in the justice sector [10]. Similar concerns can be noticed elsewhere in the world, most notably in the USA, where the New York City Council was the first to pass a law on algorithmic decision-making trans-

[23] *Roman Zakharov v Russia*, App no 47143/06 (ECtHR 4 December 2015).

parency[24]. The law sets up a task force to monitor the fairness and validity of the algorithms used by municipal agencies.

The use of AI in criminal justice and policing potentially affects several criminal procedure rights: the presumption of innocence; the right to a fair trial (including the equality of arms in judicial proceedings, the right to cross-examine witnesses); the right to an independent and impartial tribunal (including the right to a randomly selected judge); the principle of non-discrimination and equality; and the principle of legality (i.e., *lex certa*), and blurs the existing standards of proof.

AI is becoming even more complex with the concept of deep learning with artificial neural networks. Further technological developments might improve this (e.g., the ongoing research on 'explainable AI' may remedy the opacity of current AI approaches), but for the time being *transparency* is not much more than an illusion.

There is a sentiment that A.I. tools will vaporise the biases and mental shortcuts (heuristics) inherent to human judgment and reasoning. This is a powerful reason why AI technologies have too quickly been given too much power to tackle and solve essentially social (and not technological) problems. Social scientists, including lawyers, must engage more intensively with computer and data scientists in order to build a human-rights-compliant approach.

Listing the relevant fundamental rights and analysing case studies may be of great benefit as regards the human rights compliance of the novel systems that may be used in the future. However, we may still find *any* list inadequate. When a process of deciding by automated means involves the use of automated reasoning to aid or replace a decision-making process that would otherwise be performed by humans, *any human right* may be affected depending on the social domain in which the systems are employed.

Listing possible actors in the chain of building and employing AI systems may also lead to all-encompassing lists of state and private sector actors. The deepening of the digital ecosystem has led to a situation where responsibilities are becoming increasingly spread to a number of dependant actors. We can map responsibility in several ways: from the obligations of states to the obligations of the private sector; from data preparation to writing algorithm code (how data is cleaned and prepared, which data is taken in and used, and which data is left out of the calculus, etc.); from algorithmic design and development to implementation processes, etc. With the deepening of the digital ecosystem it becomes much more burdensome to determine who is responsible for certain data intake and algorithmic output. The acts committed might not even reach the existing thresholds of accountability. It may even be unjust to hold an actor accountable for the consequences of activities that are generally of great benefit to a society. An actor may be generating a risk our societies are willing to accept as 'socially permissible risk' [14].

One way forward is to learn from experiments from domains other than that of justice. In her succinct analysis of automated welfare systems in the USA, Eubanks [9] shows how removing human discretion from public assistance eligibility assessment seemed like a compelling solution to ending discrimination against African-Americans in the welfare system. If human decision-makers are biased, then moving towards eliminating humans from the decision-making loop seems logical. However, despite the fact that such a move towards automation and the elimination of the human from the decision-making process may intuitively feel like the right move to make, the experiences

[24] A local law in relation to automated decision systems used by agencies, No 2018/049. Available at: http://legistar.council.nyc.gov/LegislationDetail.aspx?ID=3137815&GUID=437A6A6D-62E1-47E2-9C42-461253F9C6D0.

that Eubanks uncovered show that this may well be counterproductive. What advocates of automated decision-making systems neglect is the importance of the ability to bend the rules and reinterpret them according to social circumstances [9]. Removing human discretion thus is a double-edged sword: it can reduce human bias, but it can also exacerbate past injustices or produce new ones.

Similarly, in Turkle's analysis of the social acceptability of computerised decisionmaking systems, she claims that when a system is perceived as discriminatory and one that creates racially disparate outcomes in sentencing, disadvantaged African Americans would choose a computerised judge rather than a human judge [36]. After all, human judges tend to be white middle-aged men. The 'tough on crime' laws that established mandatory minimum sentences form any categories of crime and removed part of judges' discretion did make US criminal justice fairer - but all defendants were hit hard and prisons soon became overcrowded. Ironically, writes Eubanks [9, p. 81], the adoption of 'tough on crime' laws were a result of organising by both conservative 'law-and-order' types and by some progressive civil rights activists who saw the bias in judicial discretion. However, the evidence of the past thirty years is different: racial disparity in the criminal justice system has worsened, and mandatory sentencing laws and guidelines have put sentencing on autopilot [9, p. 81].

Lastly, the impacts of AI systems extend beyond human rights. Their impacts may have distorting effects on the fundamental cornerstones and architecture of liberal democracies, i.e., regarding the principle of the separation of powers and the limitation of political power by the rule of law.

References

1. Agüera y Arcas, B., Mitchell, M., Todorov, A.: dPhysiognomy's New Clothes. Medium (2017). Available at: https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a

2. Brkan, M.: Do Algorithms Rule the World? Algorithmic Decision-Making in the Framework of the GDPR and Beyond. SSRN Scholarly Paper (2017)

3. Calling Bullshit: Case study. Criminal Machine Learning (2017). Available at: https://callingbullshit.org/case_studies/case_study_criminal_machine_learning.html?fbclid=IwAR3dfUkn5nY0RR54fcAmyASQK9LmC-n4LWRn3wlcL2eguB3Whd14mzEsEdE

4. Committee of Experts on Human Rights Dimensions of Automated Data Processing and Different Forms of Artificial Intelligence (MSI-AUT): Draft Recommendation of the Committee of Ministers to Member States on the human rights impacts of algorithmic systems (26 June 2019)

5. Cummings, M.L.: Automation Bias in Intelligent Time Critical Decision Support Systems. American Institute of Aeronautics and Astronautics (2014). Available at: https://web.archive.org/web/20141101113133/http://web.mit.edu/aeroastro/labs/halab/papers/CummingsAIAAbias.pdf

6. Dewan, S.: Judges Replacing Conjecture with Formula for Bail. The New York Times (2015)

7. Eck, J.E., Chainey, S., Cameron, J.G., Leitner, M., Wilson, R.E.: Special Report. Mapping Crime: Understanding Hot Spots. U.S. Department of Justice, Office of Justice Programs, National Institute of Justice (2005). Available at: http://discovery.ucl.ac.uk/11291/1/11291.pdf

8. Eckhouse, L.: Big data may be reinforcing racial bias in the criminal justice system. Washington Post (2017)

9. Eubanks, V.: Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor, pp. 80–81. St. Martin's Press, New York (2018)

10. European Commission for the Efficiency of Justice (CEPEJ): European Ethical Charter on the use of artificial intelligence in judicial systems and their environment (2018). Available at: https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c

11. European Union Agency for Fundamental Rights: #BigData: Discrimination in data-supported decision making (2018)

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

13. Freeman, K.: Algorithmic injustice: how the Wisconsin Supreme Court failed to protect due process rights in state V. Loomis. NC J. Law Technol. 18(5), 75–106 (2016)

14. Gless, S., Silverman, E., Weigend, T.: If robots cause harm, who is to blame? Self-driving cars and criminal liability. New Crim. Law Rev. 19(3) (2016)

15. Goode, E.: Data-Crunching Program Guides Santa Cruz Police Before a Crime. The New York Times (2011)

16. Gorner, J.: Chicago Police Use 'Heat List' to Prevent Violence. The Chicago Tribune (2013). Available at: www.policeone.com/chiefs-sheriffs/articles/6403037-Chicago-police-use-heat-list-toprevent-violence/

17. Groff, E.R., La Vigne, N.G.: Forecasting the future of predictive crime mapping. Crime Prev. Stud. 13, 29-58 (2002)

18. Harris, S.: The social laboratory. Foreign Policy (2014)

19. Izdebski, K. (ed.): alGOVrithms—State of Play. ePa´nstwo Foundation (2019). Available at: https://epf.org.pl/en/projects/algovrithms/

20. Kadar, C., Maculan, R., Feuerriegel, S., Public: Decision support for low population density areas: an imbalance-aware hyper-ensemble for spatio-temporal crime prediction. Decis. Support Syst. 107 (2019)

21. Kanduˇc, Z.: Družbena kriza, nacionalna država in 'varnostno vprašanje' v kriminološki perspektivi. J. Crim. Criminol. 62(2), 141-154 (2011)

22. Kehl, D.L., Kessler, S.A.: Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing (2017). Available at: dash.harvard.edu/handle/1/33746041

23. Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., Mullainathan, S.: Human decisions and machine predictions. Q. J. Econ. 133, 237 (2018). More on positive uses: Sunstein, C.R.: Algorithms, Correcting Biases (December 12, 2018). Social Research. Available at: https://ssrn.com/abstract=3300171

24. Mari, A.: DCMS announces new funding for prison coding skills. Computer Weekly (15 March 2019). Available at: https://prisonstudies.us14.list-manage.com/track/click?u=cb51806b184b825cd5f587a8a&id=da236c42f8&e=134997c2cd

25. Newcomb, A.: Finland is Using Inmates to Help a Start-Up Train Its Artificial Intelligence Algorithms (2019). Available at: http://fortune.com/2019/03/28/finland-prison-inmates-train-ai-artificialintelligence-algorithms-vainu/

26. Niiler, E.: Can AI be a Fair Judge in Court? Estonia Thinks so. Wired (2019). Available at: www.wired.com/story/can-ai-be-fair-judge-court-estonia-thinks-so/amp?__twitter_impression =true

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

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

29. Plesniˇcar, M.M., Završnik, A., Šarf, P.: Fighting impunity with new tools: how big data, algorithms, machine learning and AI shape the new era of criminal justice. In: Marin, L., Montaldo, S. (eds.) The Fight Against Impunity in EU Law. Hart Publishing, Oxford (2020, forthcoming)

30. Saunders, J., Hunt, P., Hollywood, J.S.: Predictions put into practice: a quasi-experimental evaluation of Chicago's predictive policing pilot. Journal of Experimental Criminology 12(3), 1-25 (2016)

31. Schermer, B.W., Georgieva, I., Van der Hof, S., Koops, B.J.: Legal Aspects of Sweetie 2.0. Tilburg Institute for Law, Technology, and Society, Tilburg (2016)

32. Shalev, S.: Supermax: Controlling Risk Through Solitary Confinement. Willan Publishing, Milton Park (2009)

33. Siegel, E.: Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die, 1st edn. Wiley, Hoboken (2103)

34. The Committee of Experts on Internet Intermediaries (MSI-NET): Study on the Human Rights Dimensions of Automated Data Processing Techniques (in Particular Algorithms) and Possible Regulatory Implications (6 October 2017). Available at: https://rm.coe.int/study-hr-dimension-of-automateddata-processing-incl-algorithms/168075b94a

35. The Council of Europe Commissioner for Human Rights: Recommendation Unboxing Artificial Intelligence: 10 steps to protect Human Rights (May 2019). Available at: https://rm.coe.int/unboxingartificial-intelligence-10-steps-to-protect-human-rights-reco/1680946e64

36. Turkle, S.: Life on the Screen: Identity in the Age of the Internet. Simon & Schuster, New York (1995)

37. Veale, M., Edwards, L.: Clarity, surprises, and further questions in the Article 29 Working Party draft guidance on automated decision-making and profiling. Comput. Law Secur. Rev. 34(2), 398-404 (2018)

38. Wu, X., Zhang, X.: Automated Inference on Criminality using Face Images. arXiv (2016). Available at: http://arxiv.org/abs/1611.04135

39. Yan, S.: Chinese High-Security Jail Puts AI Monitors in Every Cell 'to Make Prison Breaks Impossible'. The Telegraph (2019). Available at: www.telegraph.co.uk/news/2019/04/01/chinese-prisonrolls-facial-recognition-sensors-track-inmates/

40. Završnik, A.: Algorithmic justice: algorithms and big data in criminal justice settings'. Eur. J. Criminol. (2019). https://doi.org/10.1177/1477370819876762

Original source - Završnik Aleš – ERA Forum - https://link.springer.com/article/10.1007/s12027-020-00602-0

Завршник А.♦

DOI: 10.1007/s12027-020-00602-0
УДК: 343.1

Уголовное правосудие, системы искусственного интеллекта и права человека

Аннотация: Автоматизация, вызванная аналитикой больших данных, машинным обучением и системами искусственного интеллекта, заставляет нас пересмотреть фундаментальные вопросы уголовного правосудия. В статье описывается автоматизация, которая произошла в сфере уголовного правосудия, и дается ответ на вопрос, что автоматизируется и кто при этом заменяется. Затем анализируются коллизии между системами искусственного интеллекта и законом, их воздействие на прецедентное право и права человека. В заключении статьи предлагаются некоторые мысли о предлагаемых решениях по устранению рисков, связанных с системами искусственного интеллекта в сфере уголовного правосудия.

Ключевые слова: уголовное правосудие; права человека; алгоритмы; искусственный интеллект; справедливое судопроизводство.

♦ **Завршник Алеш** – профессор, д-р, Институт криминологии Юридического факультета Университета Любляна (Словения). E-mail: ales.zavrsnik@pf.uni-lj.si