Vol. 92 issue 1, 2021

RIDP

Gert Vermeulen, Nina Peršak & Nicola Recchia *(Eds.)*

Artificial Intelligence, Big Data and Automated Decision-Making in Criminal Justice

Revue Internationale de Droit Pénal
International Review of Penal Law
Revista internacional de Derecho Penal
Международное обозрение уголовного права
刑事法律国际评论
المجلة الدولية للقانون الجنائي
Revista Internacional de Direito Penal
Rivista internazionale di diritto penale
Internationale Revue für Strafrecht

MAKLU

Artificial Intelligence, Big Data and Automated Decision-Making in Criminal Justice

Artificial Intelligence, Big Data and Automated Decision-Making in Criminal Justice

Edited by

Gert Vermeulen
Nina Peršak
Nicola Recchia

RIDP
Revue Internationale de Droit Pénal
International Review of Penal Law
Revista internacional de Derecho Penal
Международное обозрение уголовного права
国际刑事法律评论
المجلة الدولية للقانون الجنائي
Revista Internacional de Direito Penal
Rivista internazionale di diritto penale
Internationale Revue für Strafrecht

Antwerpen | Apeldoorn | Portland

AIDP – Association Internationale de Droit Pénal | The International Association of Penal Law is the oldest association of specialists in penal law in the world. Since 1924, it is dedicated to the scientific study of criminal law and covers: (1) criminal policy and codification of penal law, (2) comparative criminal law, (3) international criminal law (incl. specialization in international criminal justice) and (4) human rights in the administration of criminal justice. The Association's website provides further information (http://www.penal.org).

RIDP – Revue Internationale de Droit Pénal | The International Review of Penal Law is the primary publication medium and core scientific output of the Association. It seeks to contribute to the development of ideas, knowledge, and practices in the field of penal sciences. Combining international and comparative perspectives, the RIDP covers criminal law theory and philosophy, general principles of criminal law, special criminal law, criminal procedure, and international criminal law. The RIDP is published twice a year. Typically, issues are linked to the Association's core scientific activities, ie the AIDP conferences, Young Penalist conferences, world conferences or, every five years, the International Congress of Penal Law. Occasionally, issues will be dedicated to a single, topical scientific theme, validated by the Scientific Committee of the Association, comprising high-quality papers which have been either presented and discussed in small-scale expert colloquia or selected following an open call for papers. The RIDP is published in English only.

Peer review: All contributions are subject to double-layered peer review. The primary scientific and peer review responsibility for all issues lies with the designated Scientific Editor(s). The additional scientific quality control is carried out by the Executive Committee of the Editorial Board, which may turn to the Committee of Reviewers for supplementary peer review.

Disclaimer: The statements and opinions made in the RIDP contributions are solely those of the respective authors and not of the Association or MAKLU Publishers. Neither of them accepts legal responsibility or liability for any errors or omissions in the contributions nor makes any representation, express or implied, with respect to the accuracy of the material.

ISSN – 0223-5404
ISBN 978-90-466-1130-2
D/2021/1997/46
NUR 824
BISAC LAW026000

Maklu- Publishers
Somersstraat 13/15, 2018 Antwerpen, Belgium, info@maklu.be
Koninginnelaan 96, 7315 EB Apeldoorn, The Netherlands, info@maklu.nl
www.maklu.eu

USA & Canada
International Specialized Book Services
920 NE 58th Ave., Suite 300, Portland, OR 97213-3786, orders@isbs.com, www.isbs.com

Editorial Board

Summary

Preface: Capabilities and Limitations of AI in Criminal Justice
by Gert Vermeulen, Nina Peršak and Nicola Recchia 7

Setting the Scene

Algorithmic Decisions within the Criminal Justice Ecosystem and their Problem Matrix, *by Krisztina Karsai* 13

AI and Big Data in Predictive Detection and Policing

Applying the Presumption of Innocence to Policing with AI, *by Kelly Blount* 33

Click, Collect and Calculate: The Growing Importance of Big Data in Predicting Future Criminal Behaviour, *by Julia Heilemann* 49

Augmented Reality in Law Enforcement from an EU Data Protection Law Perspective: The DARLENE Project as a Case Study, *by Katherine Quezada-Tavárez* 69

On the Potentialities and Limitations of Autonomous Systems in Money Laundering Control, *by Leonardo Simões Agapito, Matheus de Alencar e Miranda and Túlio Felippe Xavier Januário* 87

Crimes Involving AI: Liability Issues and Jurisdictional Challenges

AI Crimes and Misdemeanors: Debating the Boundaries of Criminal Liability and Imputation, *by Anna Moraiti* 109

AI and Criminal Law: The Myth of 'Control' in a Data-Driven Society
by Beatrice Panattoni 125

The Impact of AI on Corporate Criminal Liability: Algorithmic Misconduct in the Prism of Derivative and Holistic Theories, *by Federico Mazzacuva* 143

The Challenges of AI for Transnational Criminal Law: Jurisdiction and Cooperation
by Miguel João Costa and António Manuel Abrantes 159

AI-Assisted and Automated Actuarial Justice or Adjudication of Criminal Cases

Lombroso 2.0: On AI and Predictions of Dangerousness in Criminal Justice
by Alice Giannini 179

The Use of AI Tools in Criminal Courts: Justice Done and Seen to Be Done?
by Vanessa Franssen and Alyson Berrendorf.....199

Automated Justice and Its Limits: Irreplaceable Human(e) Dimensions of Criminal Justice, *by Nina Peršak*.....225

PREFACE:
CAPABILITIES AND LIMITATIONS OF AI IN CRIMINAL JUSTICE

By Gert Vermeulen[*]*, Nina Peršak*[**] *and Nicola Recchia*[***]

Artificial intelligence (AI) is impacting our everyday lives in a myriad of ways. The use of algorithms, AI agents and big data techniques also creates unprecedented opportunities for the prevention, investigation, detection or prosecution of criminal offences and the efficiency of the criminal justice system. Equally, however, the rapid increase of AI and big data in criminal justice raises a plethora of criminological, ethical, legal and technological questions and concerns, eg about enhanced surveillance and control in a pre-crime society and the risk of bias or even manipulation in (automated) decision-making. In view of the stakes involved, the need for regulation of AI and its alignment with human rights, democracy and the rule of law standards has been amply recognised, both globally and regionally. The lawfulness, social acceptance and overall legitimacy of AI, big data and automated decision-making in criminal justice will depend on a range of factors, including (algorithmic) transparency, trustworthiness, non-discrimination, accountability, responsibility, effective oversight, data protection, due process, fair trial, access to justice, effective redress and remedy. Addressing these issues and raising awareness on AI systems' capabilities and limitations within criminal justice is needed to be better prepared for the future that is now upon us.

This special issue on 'Artificial intelligence, big data and automated decision-making in criminal justice' presents topical and innovative papers on the above issues, selected following a call for papers.

Krisztina Karsai (Algorithmic Decisions within the Criminal Justice Ecosystem and their Problem Matrix) sets the scene with a critical socio-legal paper drawing from both criminology and criminal law. After identifying and defining needs and possibilities of deploying algorithmic decision-making solutions in the various stages of the criminal justice system, she warns against the technology-driven use of AI, big data and algorithms in criminal justice, mapping six main overarching incompatibility issues or challenges.

[*] Senior Full Professor of European and international Criminal Law and Data Protection Law, Director of the Institute for International Research on Criminal Policy (IRCP), of the Knowledge and Research Platform on Privacy, Information Exchange, Law Enforcement and Surveillance (PIXLES) and of the Smart Solutions for Secure Societies (i4S) business development center, all at Ghent University; General Director Publications of the AIDP; Editor-in-Chief of the RIDP. For correspondence: <gert.vermeulen@ugent.be>.

[**] Scientific Director and Senior Research Fellow, Institute for Criminal-Law Ethics and Criminology, Ljubljana; Advanced Academia Fellow, CAS, Sofia; Member of the European Commission's Expert Group on EU Criminal Policy; Independent Ethics Adviser; Co-Editor-in-Chief of the RIDP. For correspondence: <nina.persak@criminstitute.org>.

[***] Postdoc Researcher in Criminal Law, Goethe-University Frankfurt; member of the Young Penalists Committee and of the Scientific Committee of the AIDP. For correspondence: <recchia@jur.uni-frankfurt.de>.

The next four papers centre around *AI and big data in predictive detection and policing*.

Kelly Blount (Applying the Presumption of Innocence to Policing with AI), positing that predictive policing is comparable to traditional criminal investigations in both substance and scope, argues that the presumption of innocence as a fair trial right may be nullified by predictive policing that relies upon former arrest records without taking account of possible dismissal of charges or acquittal.

Julia Heilemann (Click, Collect and Calculate: The Growing Importance of Big Data in Predicting Future Criminal Behaviour) takes a critical stance *vis-à-vis* the underregulated and customer-unfriendly private to public (big) data transfer feeding predictive policing software.

Katherine Quezada-Tavárez (Augmented Reality in Law Enforcement from an EU Data Protection Law Perspective: The Darlene Project as a Case Study), in an applied exercise, examines AI-based augmented reality solutions in law enforcement through the lens of EU data protection law, with a focus on data minimisation, the processing of special categories of data and automated decision-making.

A sectoral application is provided by *Leonardo Simões Agapito, Matheus de Alencar e Miranda* and *Túlio Felippe Xavier Januário* (On the Potentialities and Limitations of Autonomous Systems in Money Laundering Control). They analyse the pros and cons of autonomous or AI mechanisms to prevent, detect and investigate money laundering, particularly in receiving and processing reports of suspicious activities at FIU level, and propose solutions to address challenges relating to both the insufficiency, low quality and inaccuracy of the data that feed the systems and the difficulties in understanding, explaining and refuting the resulting automated conclusions.

Another four papers address *liability issues and jurisdictional challenges prompted by crimes involving AI.*

Anna Moraiti (AI Crimes and Misdemeanors: Debating the Boundaries of Criminal Liability and Imputation), exploring the implications that robots and AI agents create under the scope of the general part of substantive criminal law, argues that, while negligent criminal liability of programmers, producers and/or users may be effectively addressed, the criminal liability of (autonomous) robots and AI agents requires reconsidering anthropocentric legal presumptions and reflecting on the rights of nonhuman agents as well as on the value of non-retributive approaches to crime and punishment.

Beatrice Panattoni (AI and Criminal Law: The Myth of 'Control' in a Data-Driven Society) continues the discussion, pointing at the responsibility gap that is likely to arise when AI agents themselves cannot be held responsible, and human agents, lacking full control over AI systems' autonomous functioning, neither. Against the backdrop of the proposed EU Artificial Intelligence Act, she describes the possible and future criminal policies that will allow avoiding such responsibility gap.

Federico Mazzacuva (The Impact of AI on Corporate Criminal Liability: Algorithmic Misconduct in the Prism of Derivative and Holistic Theories) shifts the discussion away from criminal liability of AI or for AI crimes, and focuses on algorithmic corporate liability, discussing corporate liability and compliance issues resulting from the use of new technologies by corporations. He addresses strict and vicarious liability, the principle of identification, and responsibility based on organizational fault or corporate culture.

Miguel João Costa and *António Manuel Abrantes* (The Challenges of AI for Transnational Criminal Law: Jurisdiction and Cooperation) highlight the inflated level of multi-jurisdictional competence that is likely to result from the complex liability issues for crimes involving AI. They posit that the varying liability models underlying such positive jurisdiction conflicts require rethinking traditional international cooperation concepts such as the dual criminality principle. They also see a renewed role for the executive in the requested state to refuse cooperation in criminal matters on fundamental rights grounds where the use of AI has possibly affected the fairness of the procedure in the requesting state.

The last three papers deal with *AI-assisted and automated actuarial justice or adjudication of criminal cases.*

Alice Giannini (Lombroso 2.0: On AI and Predictions of Dangerousness in Criminal Justice) sketches how the development of new AI and machine learning techniques and their application in both medical and criminal justice settings spark traditional discussions on the use and acceptability of clinical and especially actuarial violence risk assessment tools in criminal courts. She critically assesses the pros and cons of AI-based neuropredictions and virtual forensic experts in criminal justice.

Vanessa Franssen and *Alyson Berrendorf* (The Use of AI Tools in Criminal Courts: Justice Done and Seen to Be Done?) focus on the current and future role of AI in the adjudication of criminal cases. Distinguishing between AI systems that facilitate adjudication and those that could, in part or wholly, replace human judges, they sketch and evaluate the possible (dis)advantages of such systems when used in criminal courts.

Nina Peršak (Automated Justice and Its Limits: Irreplaceable Human(e) Dimensions of Criminal Justice) further advances the discussion on potential drawbacks of automated justice by addressing two dimensions of criminal justice that automated decision-making – if it were ever to be fully implemented – would upend, namely, the affective dimension and the human (interactive) dimension, which encompass essential elements, requirements and values of many contemporary (and traditional) criminal justice systems.

SETTING THE SCENE

ALGORITHMIC DECISIONS WITHIN THE CRIMINAL JUSTICE ECOSYSTEM AND THEIR PROBLEM MATRIX

By Krisztina Karsai[*]

Abstract

This paper highlights the social-legal environment of criminal justice through identifying and defining the different needs and possibilities of deploying algorithmic decision-making solutions in the distinct stages of the criminal procedure. A peculiar paradox prevails in this area: although no comprehensive policy on the use of algorithms and algorithmic decision-making exists in the justice process, the application of tools using such technology is almost universal. The objective of this paper is to introduce the main challenges in this regard and to present arguments as to why the applications of algorithms within criminal justice is not evidential simply because technology enables it. The paper follows theoretical criminology methods and addresses issues and principles both from criminology and from criminal law perspectives. Six main criteria are identified, which support explaining both the lack of necessity and the lack of compliance with system-relevant values and characters of criminal justice regarding the application of algorithms. The following criteria are discussed: adaptation traps (the interplay perceived between algorithmizing data and information relevant for criminal justice); the myth of objective truth and of convincing the judge (identifying the main goal of the criminal procedure and describing the goals are to be achieved if algorithms are to play any role in the procedure); the very theoretic paradigms of criminal law and criminology (how these system-shaping paradigms will be eroded – or revolutionized – by algorithmic thinking); the immanent non-mathematisable values of criminal justice (how non-coded values can or cannot play a role in algorithmic solutions); the 'bad' subjectivity (whether or not subjectivity of the judge should be excluded), and the purity of the data (why specific data related to criminal justice are simply unusable as training datasets for algorithmic solutions).

1 Introduction and Objectives

This paper highlights the social-legal environment of criminal justice by identifying and defining the different needs and possibilities of deploying algorithmic decision-making solutions (ADM)[1] in the distinct stages of criminal procedures. The objectives of this paper are to introduce the main conceptual challenges in this regard and to present arguments as to why the application of algorithms is far from evidential within the criminal justice system, even though technology at hand would provide means for us to do so. The paper follows theoretical criminology methods and addresses issues and principles

[*] Full Professor of Law, University of Szeged, Faculty of Law and Political Sciences, Institute of Criminal Law and Criminal Science. For correspondence: <karsai.krisztina@juris.u-szeged.hu>.

[1] 'Algorithms need not be software: in the broadest sense, they are encoded procedures for transforming input data into a desired output, based on specified calculations. The procedures name both a problem and the steps by which it should be solved.' Tarleton Gillespie, 'The Relevance of Algorithms' in Tarleton Gillespie and others (eds), *Media Technologies Essays on Communication, Materiality, and Society* (MIT Press Scholarship Online 2014).

from both criminology and criminal law. The theoretical framework of this study is shaped by the (Central) European continental legal system and by both static and dynamic characteristics of criminal justice; therefore, application of my conclusions to other legal systems would first require wise adaptation and further research. Finally, I elaborate on some of the issues – or purported 'traps' that push most criminal law professionals into the so-called 'uncanny valley'[2] about the rise of algorithmic or machine decision making, so that identification can facilitate further research and discussion of the issues involved.

In 1963, Lawlor stated that 'given a chance, computers can help the legal profession in at least three very important ways. Computers can help find the law, they can help analyse the law and they can help lawyers and lower court judges to predict or anticipate decisions.'[3] He believed that the trustworthy prediction of judicial decisions is dependent upon a scientific understanding of the functioning of the law and how facts and legal norms impact judges and judicial decisions. Indeed, almost sixty years later we have made immense advances in this field, but the latent infiltration of algorithmic solutions without clear scientific reasoning has guided development in a direction different from Lawlor's prediction. Moreover, the Collingridge dilemma[4] is undeniably evident – the dilemma of control over new technologies versus innovation has yet to be; rather, the development was simply allowed to flow. The 'latent infiltration' of such technologies is obvious.[5] As expected, the technological and legal advent of 'big data' as a cultural, technological and scholarly phenomenon can be identified as key factors in the changed landscape. As Boyd and Crawford stated, 'big data' is the interplay of '[t]echnology: maximizing computation power and algorithmic accuracy to gather, analyse, link, and compare large data sets. Analysis: drawing on large data sets to identify patterns in order to make economic, social, technical, and legal claims. Mythology: the widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the *aura of truth, objectivity, and accuracy'*.[6]

---

[2] 'Uncanny valley' is a specific psychological phenomenon in connection with robots: humans develop unsettling feelings if robots are too human-like, introduced in 1970. Bibi van den Berg B, 'The Uncanny Valley Everywhere?' in Simone Fischer-Hübner and others (eds), *Privacy and Identity Management for Life* (Springer 2010).

[3] Reed C. Lawlor, 'What computers can do: analysis and prediction of judicial decisions' [1963] ABAJ 49.

[4] '[A]ttempting to control a technology is difficult…because during its early stages, when it can be controlled, not enough can be known about its harmful social consequences to warrant controlling its development; but by the time these consequences are apparent, control has become costly and slow.' The dilemma descripted by David Collingridge is cited and analysed by Audley Genus, Andy Stirling, 'Collingridge and the dilemma of control: Towards responsible and accountable innovation' [2017] Research Policy < http://dx.doi.org/10.1016/j.respol.2017.09.012>.

[5] Rebecca Wexler, 'Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System' [2018] SLRV 70.

[6] Danah Boyd, Kate Crawford, 'Critical questions for Big Data. Provocations for a cultural, technological, and scholarly phenomenon' [2012] Information, Communication & Society 662.

2 Algorithms in the Criminal Justice Ecosystem

The global conceptualisation of the broad infiltration and deep intrusion of algorithms into the everyday life of our societies speaks about 'governing algorithms'[7] and about 'algorithmic governmentality'[8] and calls for ethical and legal opportunities of applying ADM solutions within the justice sectors of societies. Moreover, 'we are witnessing a gradual movement away from the traditional, retrospective, individualized model of criminal justice, which prioritises a deliberated and personalised approach to pursuing justice and truth, towards a prospective, aggregated model, which involves a more ostensibly efficient, yet impersonal and distanced, approach. "Actuarial justice" is based on a "risk management" or "actuarial" approach to the regulation of crime and the administration of justice.'[9]

The utilization of tools applying such technologies is almost universal, but their applications are fragmented. In many situations, the impact of the decision on people can be significant, especially as regards access to credit, employment,[10] medical treatment[11] and judicial sentences,[12] among other things.[13] Although the emergence of these devices is considered a technological advancement, we have failed to consider systemic differences and specifics, and have perhaps even neglected to realise them. The novelty of the problems combined with the exclusion of analogue solutions (not algorithmic decision-making) have led to the present situation of running into confrontation due to their presence and worldwide massive penetration. The application of ADM solutions can be fundamentally different depending on the stage of the criminal justice ecosystem (or 'pipeline') at which they are implemented. ADM solutions can be deployed for prevention, detection, investigation, prosecution of crimes, in courts' proceedings and during the penal execution.

---

[7] Solon Barocas and others, 'Governing Algorithms: A Provocation Piece' [2013] SSRN <http://dx.doi.org/10.2139/ssrn.2245322> accessed 15 October 2021.

[8] Antoinette Rouvroy, Thomas Berns, 'Gouvernementalité algorithmique et perspectives d'émancipation' *Réseaux* [2013] 177 <http://doi.org10.3917/res.177.0163> accessed 15 October 2021.

[9] Amber Marks and others, 'Automatic Justice? Technology, Crime, and Social Control' in Roger Brownsword and others (eds) *The Oxford Handbook of the Law and Regulation of Technology* (OUP 2017).

[10] Sifting through personal emails for personality profiling in Finland. See more in Brigitte Alfter and others, '*Automating Society: Taking Stock of Automated Decision-Making in the EU*' (AlgorithmWatch 2019)

[11] Allocating treatment for patients in the public health system in Italy. See more in Brigitte Alfter et al. (2019) 88.

[12] Police officers apply facial recognition algorithms to identify suspects (or victims) appearing in recordings from a crime scene, and judges use risk assessment ADM solutions for bail, sentencing and parole decisions based on an individual's demographic characteristics and criminal history (and in order to predict recidivism).

[13] Some further examples include the ADM solutions used by airport security for assessing risks posed by airline passengers (no-flight lists); the automated processing of traffic offences in France or algorithmic identification of children possibly vulnerable to be neglected (Denmark). Konrad Lischka, Anita Klinger, *Wenn Maschinen Menschen Bewerten* (Bertelsmann 2017).

3 Problem Matrix of the Criminal Justice Application of Algorithms

The purpose of criminal justice is to punish the perpetrator of a crime, ie one who violates the coexistence rules of society (retaliation), and to prevent that person or anyone else from committing another (new) crime (prevention goal and deterrence objective). At the system level, we believe this punishment ensures the functioning of society and the fight against crime and, where appropriate, the reduction of crime. It is undeniable that ADM solutions are to some extent present in the full spectrum of criminal justice, in the most general sense, helping human decision-making with a purpose appropriate to their use. In light of the Collingridge dilemma, however, we are here today because the latent infiltration of these devices has left us incapable of answering the original questions. Summarising these questions is appropriate, even if we are yet unable to reassuringly answer them. Of course, a narrative exists in which meaningful answers can be presented, but that does not fall within the frame of modern, rule of law criminal justice that humanity has spent the past two centuries developing and fighting for – instead, this is something completely different – and I will return to this at the end of this paper. In my opinion, the original questions are based along the lines of six main criteria; these problems stem directly from the purpose, function and internal structure of criminal justice as a social subsystem (six criteria of the problem matrix).

Algorithmic solutions can be used for description (recognition of patterns), for the exploration of correlations and for prediction. The resulting new information and facts and their application in the criminal justice chain can be identified along specific sub-objectives. The descriptive algorithm may be suitable for:

- identifying the perpetrator, the victim, the witness (facial recognition, linking personal datasets);
- establishing a pattern for certain offences committed;
- establishing a pattern for scenes;
- establishing probability of who the perpetrator of a committed crime was;
- determining how the court 'usually' decides.

By all means, the results of the descriptive analysis can be based on a variety of theoretical constructions, ie what data are used and why; thus, for example, past similar characteristics or identical conditions; factors to which the outcome is 'presumably' related, etc. The resulting outcome can serve as a basis for actions of the authorities, ie official decision-making, thus supporting a decision to order police action and law enforcement operations. The last two possible outcomes, although this has not yet appeared in criminal cases, would present the possibility for the results of the algorithm to support the court's decision or possibly even replace it.

As such, based on established patterns, a predictive algorithm would be capable of foreseeing (for example):

- the likelihood and location of certain criminal offences;

- the likelihood of recidivism;
- the possible location of a sought person or object;
- the likelihood of becoming a victim;[14]
- the likelihood of becoming an offender;
- how the court is likely to decide;
- whether an incarcerated person is at risk for attempted suicide.[15]

While the results obtained in this way can be used in a manner similar to prescriptive systems, this method allows for the imposition of protective measures (specific crime prevention) and recidivism quality to be taken into consideration. On the other hand, classification can be distinct based on whether we are discussing big data-based algorithms or 'traditional' risk assessment-based, statistical algorithms, where the latter are characterised by the inclusion of variables and data verified by the criminological methodology in the algorithm, whereas in the former case, the algorithm works with data outside of criminal justice, and affirmations are often not delivered by criminology or another type of social research.

3.1 Adaptation traps

The basic premise of criminology is that crime as a social phenomenon and as the individual phenomenon (the commission of a crime by an individual) can depend on factors outside of criminal justice, ie it is based not only on what constitutes a crime, how thorough the police are or how well-functioning crime prevention is, but also on the individual's own circumstances, which is why the use of big data (almost always personal or person-related data) seems revolutionary and promising. That is not the problem. However, the appearance of big data and the essentially unlimited possibilities for its analysis means that a wide variety of data can be examined collectively by algorithms, namely the exploration of data that has not been researched in relation to its own characteristics or to crime, due to the absence of a basic theoretical model. If by connecting large amounts of data, an algorithm is able to detect correlations, *we tend to think that this pattern indicates a causal link***.** In many cases, however, accepted scientific methodology cannot be used to establish the correctness and explanatory power of the correlation, yet we still have a tendency to assume that if there is an abundance of data, there must also be a

[14] According to Perry and his research team, predictive policing, as part of the criminal justice ecosystem, can be divided into four broad categories: 1. Methods for predicting crimes: These are approaches used to predict places where and times at which there is an increased risk of crime. 2. Methods for predicting offenders: These approaches identify individuals likely to become offenders in the future. 3. Methods for predicting perpetrators' identities: These techniques are used to create profiles that accurately match likely offenders with specific past crimes. 4. Methods for predicting victims of crimes: Like those methods that focus on offenders, crime locations and times of heightened risk, these approaches are used to identify groups or, in some cases, individuals who are likely to become victims of crime. Walter R. Perry, "*Hollywood: Predictive policing: The Role of Crime Forecasting in Law Enforcement Operations"* [Rand Corporations 2013].

[15] A historical overview provided by Kevin Ashley, 'A Brief History of Changing Roles of Case Prediction in AI and Law' [2019] *Law in Context* 36.

pattern (correlation). Moreover, the mentioned so-called 'aura'[16] surrounds the myth of big data; however, this so-called aura is not scientifically proven.

The most significant and most common purpose of utilising ADM systems has been and continues to be the exploitation of algorithms *to improve human capabilities*, as such systems are capable of processing *much more* information in *much less time* than humans. The efficiency factor is thus the most paramount theoretical justification for the application of these systems. Coupled with this is the aforementioned – scientifically unproven – fallacy that more data will lead to the discovery of new layers of reality that have hitherto been hidden from the human mind. Drawing upon all of these factors, we may falsely arrive at the conclusion that the application of such systems will provide us with due assistance within criminal justice, so as to say that deconstructing the past (the criminal act committed) in an attempt to reconstruct the future (the application of punishment) will result in novel findings.

As I mentioned, ADM tools that are utilized in criminal justice have two functions. Even if covariance causes are unknown, identifying the correlations can be useful in building an understanding of the functioning of a given crime phenomenon. Then, coupled with the results of predictions based on the former, we may be able to influence the overall social patterns of crime (obviously in the direction of decline), and accordingly, from a crime control perspective, the lack of a scientific foundation could then be acceptable.

3.2 The myth of objective truth of the past and of the conviction of the judge

The conviction of the perpetrator and the coercive execution of penalties are based on the assumption that under criminal proceedings, the decision made is in accordance with the truth, that the crime was committed by the accused and was committed as stated in the judgement. Judicial certainty establishes truth, that is, it describes what happened. Unveiling the past in its entirety is not actually possible, nor can criminal justice rely on complete certainty, so the approach used is approximate: it requires judicial conviction (European systems) or certainty beyond reasonable doubt (Anglo-Saxon systems). And although professional regulations try to minimise the risk of its occurrence, the possibility of error is an inherent part of the system; on the one hand, in terms of the limits of perception about the past,[17] and on the other, in terms of its (human) evaluation process. This, in fact, also means that the truth accepted by a judicial decision (the exploration of a past act) can also be perceived on a probabilistic basis in the sense of how close it is to the real events. However, we have no means for measuring this; the judge or jury making the decision must assume 100% certainty. Everyone else – depending on their procedural position or on the lack of it – would have a different estimate if asked. This also means that from an external objective point of view, a probabilistic decision is made in any case. If algorithms were developed to calculate the probability of commission by the accused

[16] Boyd and Crawford, Critical questions for Big Data (n 6).
[17] Fenyvesi Csaba, 'World Tendencies of Forensic Sciences in XXI Century Criminalist' [2014] Journal of Yaroslav the Wise National Law University 9.

based on all available data (data pertaining to the act, the perpetrator, investigative actions, etc.), then, depending on the scaling, the resulting outcomes would be, for example: "It's more than 75% likely to have been committed by XY" or "it's more than 50% likely to have been…" or possibly "it's not more than 60% likely to have been…". This is quite incongruous to the current paradigm of thinking, even if we see that a 100% conviction, as mentioned, actually refers to the subjective probability of the decision-making judge (or members of the body). In other words, from a different perspective, we can pose the following question: on an overall societal level, which probability would we rather accept to be the decisive one – the approximate truth established by the court or the probability offered by the algorithm?

For the sake of completeness, the criterion according to which we should choose among the two mentioned options of decisive probability should be accompanied by two supplementing observations. First, probabilistic decision-making appears prominently in criminal justice in the course of expert activity: as the relative probability of occurrence of the alleged fact or the probability of the plausibility of the claim[18] (for example, in determining from which weapon a bullet was fired, whose fingerprint, whose DNA, who the father is, who signed it, whose voice can be heard on a recording, could the driver have stopped, had he been travelling at the permitted speed, what is the active substance content of the confiscated drug based, etc..). Moreover, on account of advances in modern technology, we are witnessing a 'rise in the level of probability'[19] in criminal justice. The development of this field has been facilitated by several notorious cases in which the probability established by the expert had filtered into the judge's decision through the 'the prosecutor's fallacy or error'.[20]

The second observation supplementing this critical criterion regards the fact that *in other branches of justice*, such as in property matters, certainty may not be required, in the light of general life experience, to rule out conflicting alternatives and the reasonable doubt they create. In civil litigation, depending on the subject matter of the litigation, the expected level of probability required for the formation of a judicial conviction may even vary from case to case.[21] The role of probability in relation to causation, culpability and the amount of compensation is particularly characteristic for civil litigation.[22]

---

[18] Orbán József, 'Comparison of Applicability of Bayesian and Frequentist Statistics in Criminal Law' [2013] *Internal Security* 1; Michael J Saks, 'History of the Law's Reception of Forensic Science' in Jay A Siegel, Pekka J Sauko (eds) *Encyclopaedia of Forensic Sciences* [2013].

[19] Fenyvesi, 'World tendencies' (n 17).

[20] See more William C. Thompson, Edward L. Schumann, 'The Prosecutor's fallacy and the defence attorney's fallacy' [1987] *Law and Human Behavior* 3.

[21] Mark Schweizer, 'The civil standard of proof – what is it actually?' [2016] *The International Journal of Evidence & Proof Volume* 3; Marco Di Bello, 'Plausibility and probability in juridical proof' [2019] *The International Journal of Evidence & Proof* 1; Alex Biedermann, Joelle Vuille, 'The decisional nature of probability and plausibility assessment in juridicial evidence and proof' [2018] *International Commentary on Evidence* 1.

[22] Julia Mortera, Philip Dawid, 'Probability and Evidence' and Basil C. Bitas, 'Probability in the Courtroom' in Tamás Rudas (ed.) *Handbook of Probability* [SAGE 2008].

To sum up, in relation to the theory of past truth and its role in criminal proceedings, it must be decided whether we seek to capture the probability offered by the algorithm, guided by the idea that it will be somehow more advantageous than the current paradigm in which the judicial decision (and any mistakes) is accepted as an approximate framework of past truth.

3.3 The twin towers of criminal science

Criminal law and criminology provide basic theoretical constructs that work towards the exclusion of algorithmic solutions, namely due to the existence of a general consensus and to a lack of refutation otherwise contradicting their inclusion. Here, I discuss four major paradigms.

Crime is a *complex social phenomenon*; this phenomenon represents the unity of two mutually inseparable yet conceptually distinct components. One element is the violation of criminal law norms, ie the human behaviour; the other is the person violating the criminal law norm. Quantitatively, the two components are also distinct. The reasons for this can be found both in the design of the legal systems and in the activities of law enforcement authorities. The difference between the two sides is strengthened by the fact that the number of persons and acts are not the same, because a person can commit several acts to be prosecuted, or an act that physically appears to be a single occurrence of crime can lead to more people being prosecuted.[23] It is a criminologically well-known fact that the whole crime, the crime as a whole, cannot be known (eg due to underreporting etc.) and therefore cannot be statistically captured to its full extent.[24] The known part of the crime, the crimes that have become known, and the range of offenders detected are the ones to whom statistical methods can be applied. The difference between total and known crime is *latent crime*, the extent, structure, temporal and spatial changes of which are unknown. The branch of criminology concerned with latent crime seeks to explore an unknown set, but even at a societal level, this is only suitable for an approximate description of the phenomenon. It is also clear that this is not a matter of technology – at least to the best of our knowledge today, as accuracy or certainty would only be provided by real-time recording and subsequent 'traceability' of all past events. The dark figure of crime (latent crime) therefore remains a black box[25], as algorithms, information and data cannot be extracted from this segment. This also means that future estimates based on revealed crime data are not based on reality, but only on a part of it, and are therefore

[23] See more John MacDonald, '*Measuring Crime and Criminality*' (Routledge 2017).

[24] The opposite opinion will be represented by Perry and his team by mentioning that criminals and victims follow common life patterns; overlaps in those patterns indicate an increased likelihood of crime. Geographic and temporal features influence the where and when of those patterns; as they move within those patterns, criminals make 'rational' decisions about whether to commit crimes, taking into account such factors as the area, the target's suitability and the risk of getting caught. Perry and others, Hollywood… (n 14).

[25] See more Ales Zavrsnik, 'Algorithmic justice: Algorithms and big data in criminal justice settings' EJC [2019] 1.

necessarily biased, so hardly any argument can be made in favour of accepting their 'truth'.

Central to criminal justice is the crime, the commission of which triggers the machinery, which is then directly aimed at proving the commission of a crime and establishing responsibility. *Crime is a normative category*, having different content through time and space. What constitutes a crime in a given country is a moral social issue embedded in the 'technical' framework of criminal law, which sometimes has a political connotation. It follows that, although a significant proportion of behaviours treated as crimes here and now are and will always be crimes (murder, theft, sexual violence, robbery etc.); their legal classification may nonetheless change, not to mention other 'non-classical' offences (tax evasion, computer fraud, abortion, etc.). This means that even the data available on the crimes committed cannot form a future estimate precisely, because of the capacity of the legislator to bring about changes.[26]

Modern (rule of law-based) criminal justice systems focus *on the perpetrator's act;* criminal justice or the application of criminal law must focus on the act committed – the perpetrator and his or her characteristics cannot be decisive factors in assessment. Meanwhile, the dominance and acceptance of criminal law does not limit the use of ADM tools that do not predict an individual's future actions. For example, the latest 'star weapons'[27] in crime prevention are those that basically predict individual future crime based either on individual or community data and undermine the criminal law provisions on the criminal act. However, this development can logically be understood as meaning that if prediction were to be viewed as a means for crime prevention rather than law enforcement, then criminal procedural guarantees could be otherwise construed – but in reality, such reclassification could not be a means of circumventing human rights.

If we want an algorithm that works with factual data related to specific crimes and that it thus 'identifies' the perpetrator or 'predicts' the commission of a crime, we also face the limitations of perceiving the reality, as *each criminal case is unique*, and the cases (the acts of crime) are shaped by many unknown causes or variables. Individual cases with their unique sets of factual circumstances are less processable algorithmically. Here, too,

[26] Data are further distorted by internal theoretical and practical rules of criminal law. For example, formerly, the crime of child pornography (essentially any conduct related to recordings of under-18s) had for years been recorded in statistics in the order of tens of thousands of commissions in Hungary, because if the perpetrator had owned and possessed multiple recordings, then the case was recorded as counts based on the number of recordings that were seized, each recording being equivalent to one count. This practice was modified to tallying the number of minors involved – based on a statistical approach; from a statistical perspective, we could say that this type of crime has drastically decreased in Hungary – but in reality, it has not.

[27] Leo Kelion, 'Crime Prediction Software "Adopted by 14 UK Police Forces"' *BBC News* (UK, 4 February 2019) <https://www.bbc.com/news/technology-47118229> accessed 15 October 2021.
See Charles Raab and others, *Ethics Advisory Report for West Midlands Police* (The Alan Turing Institute 2017) <https://www.turing.ac.uk/research/publications/ethics-advisory-report-west-midlands-police> accesssed 15 October 2021.

the question that could be raised is whether or not we would accept the tendency-based decisions of algorithms.

3.4 Non-mathematisable system-identical values

The operation of the rule of law, the peaceful coexistence of people and, of course, the changing world all hold many values that cannot be directly expressed in the specific legal norm in the "legal algorithm".[28] At best – in terms of algorithmisation – the written constitution of a country or binding international legal instruments underpin these values. At worst, the historical constitution or customary international law would have to be added to the sources of interpretation in order for the 'legal formula' to properly function. Justice, fairness, fundamental and human rights are values that should be treated as *constant 'variables' in the 'legal algorithm'*. Unfortunately, the content outlines of these values are not constant; at best, their core can be provided with an interpretation that is clear and thus can be subject to algorithmisation. If this path is followed, the recorded content can, of course, be coded into any 'legal algorithm', but this is likely to narrow the scope (content) of the principle.

Moreover, these values affect not only the meaning of the 'legal algorithm' but also the application of the 'legal algorithm' within criminal justice. In particular, this affects the right to individual liberty or human dignity; therefore, the algorithmisation of the requirement of due process seems to be an impossible undertaking.

It should also be pointed out that it was the software used in the US to predict recidivism (old-fashioned systems based on psychological risk analysis) that shed light on the fact that social coexistence values have developed in the 21st century that make the emerging correlations and patterns unacceptable, or at least render the use of the resulting correlations in decision-making (or in decision support) impossible. This is because these relationships are associated with protected characteristics such as gender, race, religion, and so on. And while correlations, and possibly causal relationships may be statistically true, in a social – and possibly political – context, we do not want this to be the case. Meanwhile, from a statistical perspective, algorithms are precisely designed to 'discriminate' on the basis of social consensus on certain values; in particular, the human rights requirement of the prohibition of discrimination makes certain forms of discrimination unacceptable.[29]

In principle, of course, it is also conceivable that our algorithm can also recognise patterns that are output by other variables that are not directly coded into the 'legal algorithm' and thus have an impact on fairness, due process, and fundamental rights requirements in a particular case. The problem, however, is that compiling enough of such cases

[28] A legal norm as a tool of algorithmic problem solving is: if in case A (hypothesis), XY behaves in a manner of B (disposition), then C will be the consequence (sanction, compulsory measure which is a legal effect or legal disadvantage).

[29] Lilian Edwards, Michael Veale, 'Slave to the Algorithm' DLTR [2017] 12.

is highly unlikely, even in legal systems serving the peaceful coexistence of a larger population, so their use as learner data would not lead to adequate results.

Regarding this criterion, it is worth returning once again to the question of objective truth and the legal consequences associated with it. The 'litmus test'[30] of criminal justice is the achievement of the aforementioned judicial certainty, so the role of doubt is therefore crucial: if there is no certainty, if doubt remains, it can only lead to acquittal. However, this key feature of criminal justice cannot be used if the result of the calculation of the algorithm is scalable, ie if it shows the result as a percentage or on any scale instead of a clear 'yes' or a 'no' (eg the probability of guilt or the probability of the crime having been committed, or perhaps whether the person perpetrated the act or not). Under the current principle (if doubt exists, conviction will not be made) would certainly not be applicable in such a scalable paradigm, but this could be circumvented by still accepting some 'degree' of quantified doubt within 'judgement certainty'. This would mean, for example, an 80% or 90% probability of a conviction.[31]

3.5 The 'bad' subjectivity

Exclusion of the subjective component of the human factor in relation to automatic data processing and consequent decision-making may arise as an advantage. Throughout the entire ecosystem of the criminal justice, achieving the set goals requires a series of human decisions.

More specifically, a judicial decision is a human decision, so it is evident that the subject of the judge influences the content of his or her decision. An important system-shaping element is that the legal education, the professional conditions, and the socialisation of becoming a judge guarantee non-subjective professionalism; therefore, it is assumed that undesirable subjectivity does not appear in decisions. It is critical to clearly define what we consider to be a subjective component that we would prefer to banish from algorithmic decision-making. The subjectivity of the judge — in modern criminal justice — is key to humanity, values that are central to democracies based on the rule of law. Emphasis should be placed on the importance of life experience for the judiciary, which comprises the totality of the actions or other manifestations of different people observed in different life situations and includes their comparison and the ability to build upon and draw conclusions from these. The Subject-Judge helps to recognise the problem and then translate the decision into real life. The judge mediates community content and community values and integrates the procedure and the decision into the social coexistence. Based on all this, the subjectivity of the judge – the human decision-maker – actually serves as a kind of control or a factor that allows for the decision to remain within a lucid and logical framework and enables the decision to be made at all. On the other hand, clearly other features of subjectivity also influence the decision, but not at the professional level.[32] The

[30] Zavrsnik, Algorithmic justice (n 25).

[31] Further details see Marco Di Bello, 'Trial by statistics: is a high probability of guilt enough to convict?' *Mind* [2019] 512.

[32] See more, for example Tania Sourdin, 'Judge v. robot?' [2018] *UNSW Law Journal* 4.

basic premise of modern criminal procedure is the exclusivity and omnipotence of the intellect, along with the promise that there will be no room for emotion alongside reason.[33]

Thus, the idea that automatic decision-making could eliminate the 'bad' subjective component is understandable. However, this is neither practical nor possible in our current context of criminal justice, especially in the phases of the criminal procedure that follow the detection stage. For if this were the goal, we would then merely replace the subject-relevant psychological risk of judicial decision-making with the risk logic of algorithms. This, of course, could be pure objectivity and thus an acceptable new paradigm, but if algorithms in criminal justice cannot work with pure data (see the next criterion – under section 3.6.), no added value would result from swapping the two types of risk. Furthermore, the possibility of judicial discretion capable of dealing with the uniqueness of cases would be lost, and the consideration of the essential values already mentioned in the fourth criterion would be undermined.

Professional competence is treated as an axiom of judicial (human) decision-making, as one of the most important safeguards and, at the same time, as a key element of a fair, non-arbitrary and humane justice system. One might also think that if this component were taken away from the formula, incompetent and unacceptable decisions would be made. However, the combination of probabilistic decision-making and criteria that focus on the human decision-maker yields surprising results that may override this paradigm, where appropriate. Scientific research also deals with whether it is possible to foresee the expected decision of a court, even without a detailed legal examination of the cases, ie whether the professional component in decision-making could as such be waived.

In a 2016 study, decisions of the European Court of Human Rights (ECtHR) were examined using the *natural language processing method* and, based on these, the researchers made predictive conclusions about the decisions, the existence of the alleged violation or its exclusion.[34] Research was carried out on Articles 3, 6 and 8 of the ECHR in approximately 600 cases: relevant information and textual relationships were filtered out of the text file and transformed into learning data, and then the output became the binary code of the decision, ie whether there was an infringement in that case. This was then compared to the actual decision, based on which it was concluded that the algorithm worked with an average accuracy of 79% (in an average of 79 cases out of 100 cases, the decision of the algorithm was in line with the decision made by the human judge), which is a fairly good result. Interesting results could be acquired through examination of the cases that received a 'wrong' prediction, and whether the conditions that set them apart from

[33] See more Russel Cropanzano and others, '*Social Justice and the Experience of Justice*' [Routledge 2011]. Moreover, in a research study published in 2017, the researchers used the US Supreme Court as an example and demonstrated that the judges implicitly reveal their leanings during oral arguments, even before arguments and deliberations had been concluded. Bryce J. Dietrich and others, 'Emotional Arousal Predicts Voting on the U.S. Supreme Court' [2018] *Political Analysis* 2.

[34] Nikolaos Aletras and others, 'Predicting judicial decisions of the European Court of Human Rights' *PeerJ Computer Science* [2016] < https://doi.org/10.7717/peerj-cs.93> accessed 1 October 2021.

the rest are identifiable. Unusual constellations may result depending on the outcome of the study, for example, cases in which a pattern could not be determined by the algorithm, and with this, whether some circumstances had existed that 'diverted' the decision (eg, political or 'harm-reducing' is a less legitimate argument).

A surprisingly similar study, published in 2017, is also worth mentioning. The study concerned the decisions of the US Supreme Court, and did not involve examining the legal reasoning of the underlying cases.[35] The researchers examined the votes of judges in 240,000 decisions (28,000 cases between 1816 and 2015), judicial factual data (subject matter, fact of cited legislation, date of submission, the deciding court, procedural acts in the main proceedings, etc.), judges' appointment data (identifiable political ideology), and decision-making characteristics (such as the likelihood of a dissenting opinion, etc.). The applied algorithm estimated the votes of each judge with an accuracy of almost 70–72%.

Professionalism and professional competence, as key components of human judicial decision-making, were not included in the variables in a targeted and meaningful way in any of the mentioned research. And yet, in most cases, the algorithm resulted in predictions in line with the judicial decisions. This can be a good basis for further research and a possible paradigm shift in machine decision-making and leaves us to decide whether – combining these results with the second criterion – the 72% or 79% accuracy can be a rate sufficient enough to replace decision-making. If based on the studies, we were to accept those rates and that decisions made by the algorithms are to a vast degree the same as those made by the human court, would we forgo court decision-making (efficiency, human resources, cost factors, time factor)?

These research studies were conducted with non-criminal case analysis; moreover, the algorithms delivered predictions of courts, whereas the examination of these was limited to the questions of law – as both the ECtHR and the US Supreme Court cases were aimed at reviewing the legal compliance of an earlier decision. However, the discovery and identification of relevant facts is also necessary for the decisions of the ECtHR and the US Supreme Court, and thus is still comparable to judicial activity in criminal matters. Considering these, the degree of accuracy of the research results becomes surprising – these are the results that *could be used to resolve cases* if the second criterion of the matrix discussed here is resolved and accepted. The question of what would happen if we were to compare the results of algorithms with different logic working with the same data set will continue to remain open for some time to come. It is also questionable whether algorithms for predicting the decisions of courts dealing with facts and law and those of courts dealing with questions of law alone should choose different paths, or whether the difference between the two types of judicial activity necessarily disappears during the process of algorithmisation.

[35] Daniel Martin Katz and others, 'A General Approach for Predicting the Behaviour of the Supreme Court of the United States' [2017] *Plos ONE* 4.

3.6 Purity of algorithms – 'you are what you eat'

The responses that result from the analysis of data and information (to support or replace human decisions) depend on the input data. For both algorithms following traditional methodologies and big data-based ADM solutions, the key issue is what data the algorithm should be allowed to work with (ie learn from). It is a basic requirement that both the training and further datasets used be clear and unbiased; otherwise, the output, either the pattern or the prediction, will certainly be objectionable. Within the justice field, this requirement raises several issues that are rather difficult to resolve.

If the data used to 'feed' the algorithm are the previous court decisions themselves (meaning human decisions), then we assume and accept that all previous judicial (or possibly other official) decisions were legally correct, as we allow the algorithm to be based on drawn patterns. This starting point is obviously correct in legal terms, but a retrospective *possibility to remedy* errors exists in all legal systems, typically in the form of extraordinary legal remedies, therefore, changes as subsequent rewritten input modifies the database and changes the output. And this can have an impact on the decisions based on it, so such solutions should hardly be avoided.

If past judicial decisions serve as the basis for the use of the ADM tool in criminal justice, the patterning potential of the *subjective* factors that may have appeared in past judicial decisions (discrimination, racism, etc.) may be filtered through the 'uncovered' contexts of the algorithm, ie we would exclude the infiltration of the human subject from the individual case, but allow it in its cumulative effect.[36] A 'shining' example of this is the software COMPAS, which is used in many US jurisdictions for predicting reoffending patterns and to support judicial sanctioning. COMPAS was accused of *being biased* against black defendants because it classified a greater share of black defendants as high-risk reoffenders than white defendants. Meanwhile, the algorithm assigned defendants scores from 1 to 10 that indicated how likely they were to reoffend based on more than 100 factors, including age, sex, and criminal record, while race was not used (!) as an indicator.[37] And yet, the differences incurred may have been the result of an interplay between the indicators, which represented (perhaps biased) former human decisions as criminal records or prior arrests (eg heavier policing in predominantly black neighbourhoods in certain areas).

The question of *data distillation* as a fundamental activity in any database construction may arise. However, if the data entered as input are judicial decisions made based on the appropriate rules, representing 'truth' and legal correctness, it would be challenging to find a legitimate basis for cleaning the data.

---

[36] Zavrsnik, Algorithmic justice (n 25).

[37] Anthony W. Flores and others, 'False Positives, False Negatives, and False Analyses' *Community Resources for Justice (US)* 2017 <https://www.crj.org/assets/2017/07/9_Machine_bias_rejoinder.pdf> accessed 15 October 2021.

Changing the reality of social coexistence is accompanied by a change in the law; it seems easy to incorporate legislative changes into the algorithms, but legal practice will not change, as the algorithms will not look for a new direction or new interpretation based on the 'old' pattern. This type of algorithmisation *freezes the practice of law* because even if human factors remain in the decision-making process, they do not receive a new or different impulse (the ADM solution provides results shaped by the 'old' pattern). And if human factors were to be fully excluded, the 'algorithmic jurisprudence' necessarily remains unchanged; it will no longer be possible to change it in an organic way as adapted by judges to the changing world.

4 Further Discussion

The present problem matrix contains *inherently interrelated factors* – the professional and scientific paradigms of the rule of law in criminal justice fundamentally lead the way, ie whether and to what extent there is room for algorithmic decision-making throughout the chain. It is also conceivable that, if we accept the correctness of previous decisions, the algorithms can ponder the chances of litigants winning the lawsuit, and the law will allow the parties to accept the higher probability and give legally binding status to the decisions so calculated. This, of course, does not override the specifics of criminal justice, but it can bring significant efficiencies and better resource management in other areas of law. The disengaged capacity can then be used in the human-intensive decision-making processes of criminal justice. Of course, it is also possible that the criteria problems discussed should be bridged through other solutions with a different basic paradigm and *allow the twin towers of criminal science to collapse*, in order to replace them with a new philosophy and a new type of criminal justice. It is possible that in the new paradigm, our knowledge of crime will be provided by databases, the relationships between information or facts will be delivered by algorithms, crime action will shift towards prediction and crime prevention will be the main activity. If this is not successful, then automated justice will make the judgment.[38] However, we have yet to arrive there. Until we can deliver proper answers to the questions raised by the stealth infiltration of algorithms, my position is the non-application of algorithmic systems within the criminal justice ecosystem (pipeline), and I propose for academia and legislative bodies to re-examine (in their own jurisdiction) the application of any of ADM solutions in the light of these criteria and formulate doubts and new limitations they may bear.

References

Aletras N, Tsarapatsanis D, Preoţiuc-Pietro D and Lampos V, 'Predicting judicial decisions of the European Court of Human Rights' (2016) PeerJ Computer Science' <https://doi.org/10.7717/peerj-cs.93>

[38] See Zavrsnik, Algorithmic justice (n 25).

Alfter B, Müller-Eiselt R and Spielkamp M, 'Automating Society: Taking Stock of Automated Decision-Making in the EU' (AlgorithmWatch 2019) <https://algorithmwatch.org/de/wp-content/uploads/2019/02/Automating_Society_Report_2019.pdf>

Ashley KD, 'A Brief History of Changing Roles of Case Prediction in AI and Law' [2019] Law in Context

Barocas S, Hood S and Malte Ziewitz A, 'Governing Algorithms: A Provocation Piece' [2013] <https://www.semanticscholar.org/paper/Governing-Algorithms%3A-A-Provocation-Piece-Barocas-Hood/5a518d93366180456130e7d003b4aaf3a0a2bae7>

Biedermann A and Vuille J, 'The Decisional Nature of Probability and Plausibility Assessment in Juridicial Evidence and Proof' (2018) 16 International Commentary on Evidence 1.

Bitas BC, 'Probability in the Courtroom' in Rudas T (ed.), *Handbook of Probability*: *Theory and Applications* (SAGE 2008)

Boyd D and Crawford K, 'Critical questions for Big Data. Provocations for a cultural, technological, and scholarly phenomenon' (2011) 15 Information, Communication & Society 662

Cropanzano R, Stein JH and Nadisic T, *Social Justice and the Experience of Justice* (Routledge 2011)

Di Bello M, 'Plausibility and Probability in Juridical Proof' (2019) 12 The International Journal of Evidence & Proof 161

—— 'Trial by Statistics: Is a High Probability of Guilt Enough to Convict?' (2019) 128 Mind 1045

Dietrich BJ, Enos RD and Sen M, 'Emotional Arousal Predicts Voting on the U.S. Supreme Court' (2018) 27 Political Analysis

Edwards L and Veale M, 'Slave to the Algorithm: Why a "right to an explanation" is probably not the remedy you are looking for' (2017) 16 Duke Law & Technology Review 18

Fenyvesi C, 'World Tendencies of Forensic Sciences in XXI Century' (2014) Criminalist. Journal of Yaroslav the Wise National Law University, Apostille Publishing House LLC, Ukraine 9/2014, 10-21

Flores AW, Lowenkamp CT and Bechtel K, '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"' *Community Resources for Justice* (US) 2017 <http://www.crj.org/>

Genus A and Stirling A, 'Collingridge and the Dilemma of Control: Towards Responsible and Accountable Innovation' [2017] Research Policy <http://dx.doi.org/10.1016/j.respol.2017.09.012>

Gillespie T, 'The Relevance of Algorithms' in Gillespie T, Boczkowski PJ and Foot KA (eds), *Media Technologies Essays on Communication, Materiality, and Society* (MIT Press Scholarship Online 2014), 167-94 <https://www.researchgate.net/publication/281562384>

Katz DM, Bommarito MJ, Blackman J, 'A General Approach for Predicting the Behaviour of the Supreme Court of the United States' (2017) 12 Plos ONE 1

Lawlor RC, 'What Computers Can Do: Analysis and Prediction of Judicial Decisions' (1963) 49 American Bar Association Journal 337

Lischka K and Klingel A, 'Wenn Maschinen Menschen bewerten'. Internationale Fallbeispiele für Prozesse algorithmischer Entscheidungsfindung. Arbeitspapier, May 2017, Bertelsmann Stiftung. DOI 10.11586/2017025

Marks A, Bowling B and Keenan C, 'Automatic Justice? Technology, Crime, and Social Control' in Brownsword R, Scotford E and Yeung K (eds), *The Oxford Handbook of the Law and Regulation of Technology* (OUP 2017)

MacDonald J, *Measuring Crime and Criminality* (Routledge 2017)

Mortera J and Dawid P, 'Probability and Evidence' in Rudas T (ed.), *Handbook of Probability: Theory and Applications* (SAGE 2008)

Orbán J, 'Comparison of Applicability of Bayesian and Frequentist Statistics in Criminal Law' (2013) 5 Internal Security197

Perry WL, McInnis B, Price CC, Smith SC and Hollywood JS, *Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations* (Rand Corporation 2013)

Rouvroy A and Berns T, 'Gouvernementalité algorithmique et perspectives d'émancipation' (2013) 1 Réseaux (No 177), 163 (DOI 10.3917/res.177.0163) Translated by Elizabeth Libbrecht [Algorithmic Governmentality and Prospects of Emancipation]

Saks MJ, 'History of the Law's Reception of Forensic Science' in Siegel JA and Sauko PJ (eds), *Encyclopaedia of Forensic Sciences* (Academic Press 2013)

Schweizer M, 'The Civil Standard of Proof – What is It Actually?' (2016) 20 The International Journal of Evidence & Proof 217

Sourdin T, 'Judge v. robot? Artificial intelligence and judicial decision making' (2018) 41 UNSW Law Journal1114

The Alan Turing Institute, 'Report on Ethics Advisory Report for West Midlands Police' (2017) <https://www.turing.ac.uk/research/publications/ethics-advisory-report-west-midlands-police>

Thompson WC and Schumann EL, 'The Prosecutor's Fallacy and the Defence Attorney's Fallacy' (1987) 11 Law and Human Behavior 167

Van den Berg B, 'The Uncanny Valley Everywhere?' in Fischer-Hübner S, Duquenoy P, Hansen M, Leenes R and Zhang G (eds), *Privacy and Identity Management for Life* (Springer 2010)

Wexler R, 'Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System' (2018) 70 Stanford Law Review 1343

Zavrsnik A, 'Algorithmic Justice: Algorithms and Big Data in Criminal Justice Settings' (2019) European Journal of Criminology <https://doi.org/10.1177/1477370819876762>

AI AND BIG DATA IN PREDICTIVE DETECTION AND POLICING

APPLYING THE PRESUMPTION OF INNOCENCE TO POLICING WITH AI

By Kelly Blount[*]

Abstract

This paper argues that predictive policing, which relies upon former arrest records, hinders the future application of the presumption of innocence. This is established by positing that predictive policing is comparable to traditional criminal investigations in substance and scope. Police records generally do not clarify whether former charges result in dismissal or acquittal, or conversely, conviction. Therefore, police as state actors may unlawfully act in reliance on an individual's former arrest record, despite a favourable disposition. Accordingly, it is argued that the presumption of innocence as a fair trial right may be effectively nullified by predictive policing.

1 Introduction

Artificial intelligence (AI) is a highly disruptive technology in all manners of daily life, its uses ranging from life-saving and convenient, to suspect and even dangerous. Its growing use in criminal justice processes is no different. While it allows for real time intervention in diffusing deadly situations and finding missing persons, it also opens the door to myriad, unprecedented types of control. As a result, it changes the way that we as citizens interact with the law and the way in which we define criminal justice. This paper will examine the right to the presumption of innocence as applies to criminal defendants in the early criminal justice stages. It will argue that the use of risk assessments for predictive policing is incompatible with the presumption of innocence (Presumption), per the case law of the European Court of Human Rights (ECtHR).

The paper will begin by unpacking the use of AI in policing, namely through the use of risk assessments, used interchangeably here with 'predictive policing.' It will compare traditional criminal investigations to predictive policing to demonstrate that the advances made possible by AI change the character of policing and increase the relative position of the state. In the following section, it will define the Presumption per the European Charter of Human Rights (ECHR). The Presumption, though an important procedural, fair trial right, is often debated as to its scope and application. After a brief review of the doctrine, this section will address the Presumption as treated by the ECtHR.

The remainder of the paper will apply the Court's interpretation of the Presumption to the use of predictive policing to support the assertion that they are contradictory in practice. First, it will be demonstrated that the use of police records for risk assessments necessarily include individuals whose former charges may have been dismissed or acquit-

[*] JD; doctoral researcher of criminal law, University of Luxembourg, supported by the Luxembourg National Research Fund (FNR) (PRIDE15/10965388); For correspondence: <kelly.blount@uni.lu>.

ted. Second, the role of police as public authorities per the Presumption will be addressed. And finally, the use of risk assessments will be argued as comparable in scope to aspects of traditional criminal investigations, which are subject to the Presumption.

2 Risk Assessments and Predictive Policing

Policing practices have changed in their very essence as responding to crime gives way to prevention. Consequently, not only do practices and technologies evolve, but also the boundary between crime control and investigation is blurring.[1] This trend is made possible and powerful by harnessing AI, and preventing crime is achieved by the use of surveillance and risk assessments.[2] Though these practices may originate in the legitimate goals of curbing crime and increasing public safety, they also change the relationship between the state to the individual, as well as the application of the law to the individual.[3] Traditional criminal investigation is a study of 'facts presented by a criminal act or pattern of criminal conduct'.[4] Using this information police may identify and locate individuals believed to be involved with a crime. Risk assessments, which also assess the facts of criminal acts or patterns, albeit historical, are similarly used to identify and locate potential offenders. Though the processes are temporally distinct, similar methods are employed.

Risk assessments are designed to forecast the probability of future crime based on data that reflect past criminal acts, according to an environmental theory of crime.[5] By assessing the correlation and prevalence of repeat factors that coincide with crime, it should be possible to infer the likelihood of future crime based on the co-existence, or lack of, particular attributes.[6] These include situational and environmental factors, but largely centre on the temporal and locational details of crime statistics, as recorded in past arrest and crime reports.[7] AI is critical to the process of identifying correlations because it vastly expands the ability to ascertain connections between large, disparate sets of data. Connections that are virtually unrecognizable to the human eye are reduced to calculations performed in seconds. With the results, policing agencies may then determine how to

[1] Trevor Jones, 'Governing Security: Pluralization, Privatization, and Polarization in Crime Control and Policing' in Mike Maguire, Rod Morgan and Robert Reiner (eds) *The Oxford Handbook of Criminology* (5th edn OUP 2007).
[2] Wim Hardyns and Anneleen Rummens, 'Predictive Policing as a New Tool for Law Enforcement? Recent Developments and Challenges' [2018] 24 Eur J Crim Policy Res 201.
[3] David Garland, *The Culture of Control* (University of Chicago Press 2001).
[4] Daniel Reilly, *Finding the Truth with Criminal Investigation* (Rowman & Littlefield 2019) 3-4.
[5] Nina Brown and Donald Janelle, 'Robert Park and Ernest Burgess: Urban Ecology Studies, 1925' (2002) CSISS Classics UC Berkeley <https://escholarship.org/uc/item/6f39q98d> accessed 7 October 2021.
[6] Walter L Perry and others, 'Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations' (RAND 2013) <http://ebookcentral.proquest.com/lib/unilu-ebooks/detail.action?docID=1437438> accessed 9 January 2020.
[7] David Weisburd, 'The Law of Crime Concentration and the Criminology of Place.' [2015] 53 Criminology 133.

best allocate their resources accordingly, allowing efficiency gains over traditional investigations that are generally responsive to a single crime event.[8]

Risk assessments are not considered to be proof of criminal behavior nor standalone justification for an arrest or stop.[9] Regardless, as they guide policing, they factor heavily into real-time judgments on preventative stops and arrests. Though most commonly used for predicting high crime locations, some jurisdictions have also employed risk assessments to profile individuals.[10] This form of patrol has been described as policing by 'automated suspicion'.[11] Pre-emptive patrols alter if not circumvent the standards applied to policing, for instance, forming individualized suspicion. Risk assessments collate a vast quantity of proxy information and make it actionable for policing by creating a suspect profile. This mirrors a traditional criminal investigation, but rather than seeking out an individual, a class or category of suspects is produced. It is asserted here that the use of investigatory tools for crime prevention constitutes a process parallel to a criminal investigation.

Policing formally sits outside the scope of a trial and is not considered a pre-trial process. It therefore is not subject to fair trial guarantees per the European Convention on Human Rights (ECHR). However criminal investigations subsequent to a charge do fall within the scope of pre-trial processes and are subject to the requirements set forth in Article 6. The argument against applying the Presumption to policing generally ends at this point, at which all parties agree that there is a clear divide between trial procedure and policing, as well as the respective standards.[12] However they remain two parts of the same process and it is problematic to disassociate the effects of policing from individuals' right to a fair trial.

3 Interpreting the Presumption of Innocence

The Presumption receives varied and at times confusing treatment across legal systems. Its application ranges from an evidentiary and procedural standard, to a more expansive approach rooted in normative justifications.[13] The Presumption is universally defined as

[8] Martin Innes, 'The Art, Craft, and Science of Policing' in (Peter Cane and Herbert M. Kritzer eds), *The Oxford Handbook of Empirical Legal Research* (OUP 2010).

[9] Andrew Ferguson, 'Predictive Policing and Reasonable Suspicion' [2012] 62 Emory Law Journal 261.

[10] Amnesty International, 'Trapped in the Matrix: Secrecy, Stigma, and Bias in the Met's Gangs Database' (Amnesty International 2018); David Weisburd, 'Does Hot Spots Policing Inevitably Lead to Unfair and Abusive Police Practices, or Can We Maximize Both Fairness and Effectiveness in the New Proactive Policing?' [2016] University of Chicago Legal Forum 661.

[11] Sabine Gless, 'Automated Suspicion - and Evidence?' (Facial recognition vs. Criminal Justice, Council of Europe AI & Law Webinar Series, 2 February 2021) <https://www.coe.int/en/web/artificial-intelligence/-/ai-law-webinar-9-facial-recognition-vs-criminal-justice?fbclid=IwAR3Y-uYro9TEcar-qr1sJAMJUaThTv9Izhs5L9oNSitOIVaFJuR4Z5sE5Jw> accessed 2 February 2021.

[12] Kevin Cyr, 'The Police Officer's Plight: The Intersection of Policing and the Law' [2015] 52 Alberta Law Review 889.

[13] RA Duff, 'Who Must Presume Whom to Be Innocent of What? [2013] 13 Netherlands Journal of Legal Philosophy 12.

prescribing the state to treat an individual as though he/she is factually innocent, despite simultaneously establishing the burden to produce sufficiently convincing evidence to the contrary.[14] It is a legal principle set in juxtaposition to the prosecution's prima facie assertion that a charged individual has engaged in an act which meets the statutory elements of a crime.[15] Though a presumption is by definition a rebuttable logical inference based in fact, in this case the Presumption acts as a procedural instruction to the court based in legal, if not actual fiction.[16] Whether the defendant is legally or factually innocent is not relevant to applying the Presumption, which may be inferred by the fact that the trial process is initiated solely on a belief in factual guilt. Instead, the Presumption sets the basis from which the trial will proceed.[17] It requires a formal stance that an individual is innocent to avoid insinuating or causing prejudgment, which may lead to undue deprivation of liberty. It is not an assertion that he/she is innocent, but rather ensures the rule of law is maintained in the criminal process.[18] Therefore the concept underlying the Presumption is in practice, often counterfactual and complicated to define with precision.[19]

In common law systems the Presumption is interpreted in terms of procedure and ensures the burden of proof is correctly applied between parties.[20] In inquisitorial systems it likewise provides protection to the individual as regards official public treatment and investigatory measures, however it lacks the hard procedural shell it is afforded in the adversarial context.[21] Jurisdictions subject to the ECHR, which is the focus of this work, apply the Presumption according to its role within the suite of fair trial rights guaranteed in Article 6. Article 6.2 stipulates that 'Everyone charged with a criminal offence shall be presumed innocent until proved guilty according to law'.[22] Though it is a guaranteed, non-derogable right, its scope is hard to define comprehensively. According to the Article, the Presumption does not *de facto* manifest until the issuance of a charge nor does it

---

[14] Thomas Weigend, 'Assuming That the Defendant Is Not Guilty: The Presumption of Innocence in the German System of Criminal Justice' (2014) 8 Criminal Law and Philosophy 285 <https://doi.org/10.1007/s11572-013-9271-4, 286-287> accessed 10 May 2021.

[15] "The presumption of innocence does not have any cognitive pretensions but prescribes the hypothetical starting point of due process." Van Sliedregt, 'A Contemporary Reflection on the Presumption of Innocence [2009] 80 RIDP 1, 264; Antonella Galetta, 'The Changing Nature of the Presumption of Innocence in Today's Surveillance Societies: Rewrite Human Rights or Regulate the Use of Surveillance Technologies?' [2013] 4 European Journal of Law and Technology.

[16] Carl-Friedrich Stuckenberg, 'Who Is Presumed Innocent of What by Whom?' [2014] 8 Criminal Law and Philosophy 301, 305.

[17] Sherman Clark, 'The Juror, the Citizen, and the Human Being: The Presumption of Innocence and the Burden of Judgment' [2014] 8 Crim Law and Philos 421, 424.

[18] Hamish Stewart, 'The Right to be Presumed Innocent' [2014] 8 Crim Law and Philos 407, 407.

[19] Ferry de Jong and Leonie van Lent, 'The Presumption of Innocence as a Counterfactual Principle' [2016] 12 Utrecht Law Review 32.

[20] Elies Van Sliedregt [15] 247–67; Liz Campbell, 'Criminal Labels, The European Convention on Human Rights and The Presumption of Innocence' [2013] 76 The Modern Law Review 4, 681.

[21] Weigend [14] 290-291.

[22] European Convention for the Protection of Human Rights and Fundamental Freedoms as amended by Protocols Nos. 11 and 14 [1950] ETS 5 art 6.2.

de facto apply following the trial.[23] But the bounds of the criminal process may be malleable and as described in the following section, the Presumption's interpretation is often on a case-by-case base, in accordance with preserving its core aims.[24]

At its core, the Presumption acts as a protective umbrella over the trial process to secure the space necessary for other fair trial principles to be effective. For instance, in recognizing the general disparity between parties in access to resources and the ability to conduct fact-finding, the Presumption offers the defense equal opportunity to present its case to an unprejudiced court.[25] It further shields the defendant from required self-incrimination.[26] It is nearly impossible to sever the normative implications of the Presumption from its wider role in ensuring the fairness of the trial.[27] The question to which this paper aims to offer an answer, is at the point an individual becomes a suspect subject to investigation, what protections guard this process before he/she formally becomes a defendant.[28]

As described above, the ECHR provides the baseline at which the Presumption is guaranteed but leaves much room for interpreting its scope. The ECtHR has provided further guidance on situations in which the Presumption may apply. Series of holdings suggest that the Court interprets the Presumption in the manner least likely to eviscerate its value, taking a holistic, fact-based approach. Firstly, the Court has confirmed that in certain situations the Presumption may be applicable beyond pre-trial processes. This includes both police driven pre-trial procedures, such as pre-trial detention, as well as actions taken following an acquittal. Secondly, it has addressed the categories of individuals who may be considered public authorities with the power to sway the opinion of the public as to one's innocence, as well as the content and form of such actions. And finally, it has supported an expanded reading of the Presumption when necessary to ensure its larger, practical applications in day-to-day trials, by applying other fundamental rights in order to safeguard the efficacy of the Presumption.

Following the Court's treatment of the Presumption, it may be suggested that a case-by-case analysis is necessary for ensuring the aims of the Presumption are fulfilled. Though the Court appears hesitant to apply the Presumption to determining permissive uses of coercive pre-trial processes such as pre-trial detention and intrusive criminal investigatory measures, the rationale underlying the Presumption is invoked.[29] In the use of pre-trial detention it has held that there is no comprehensive prohibition, but rather the essence of Article 6.2 should be a means of guidance on the limited use of the practices

[23] Van Sliedregt [15] 260.
[24] Galetta [15].
[25] Campbell, 'Criminal Labels, The European Convention on Human Rights and The Presumption of Innocence' [2013] 76 The Modern Law Review 4, 683.
[26] Van Sliedregt [15].
[27] Weigend [14].
[28] Duff [13] 4.
[29] Directive (EU) 2016/343 on the strengthening of certain aspects of the presumption of innocence and of the right to be present at the trial in criminal proceedings [2016] OJ L65 art 2.

deemed acceptable prior to trial.[30] This must necessarily be true. As the Presumption protects the innocent against wrongful convictions and subsequent punishment, the innocent must also be protected against undue harsh treatment prior to the trial.[31] If the trial were to function according to the requirements of the Presumption but a gap in application allows that a suspect may languish in pre-trial detention for an unreasonable period, it is arguable that punishment has already been enacted and to apply the Presumption only at the beginning of a trial is completely arbitrary and meaningless.[32] This approach extends the Presumption's legitimacy at all points throughout the criminal justice process in order to ensure its value is solid, rather than an itinerant procedural principle.[33]

Similarly, the Court has held that the Presumption is also applicable to those who have not been found guilty. In the case of acquittals, the ECtHR has held that the Presumption applies against actions, statements, and manifestations of a belief by state authorities that the acquitted individual is in fact guilty. In *Asan Rushiti v. Austria* the Court held that the issuance of a final acquittal makes even the 'voicing of suspicions regarding the accused's innocence' by a public authority, 'incompatible' with the Presumption.[34] In this case, the applicant's compensation claim was denied following his narrow acquittal due to a judicial determination that there was still a reasonable, and credible suspicion that the individual was in fact guilty, regardless of his acquittal.[35] By extending the Presumption beyond the scope of trial the Court is not only upholding its procedural uses, but ensuring that its underlying rationale retains legitimacy. Were the Presumption to end with the trial, the state may informally and publicly treat the acquitted as guilty and the court's power to find an individual innocent would be totally hollow in practice.

Next, the Court has clarified what actions taken by which parties are subject to Article 6.2. It has recognized that police and comparable authorities are state authorities or agents, who have in their control the capacity to make public manifestations that would have the same adverse, substantive effects on an individual's right to be viewed as innocent as if spoken by a courtroom official.[36] The public use of coercive measures by police in the pursuit of an arrest arguably represents a high degree of belief in the guilt of an

[30] Lonneke Stevens, 'Pre-Trial Detention: The Presumption of Innocence and Article 5 of the European Convention on Human Rights Cannot and Does Not Limit Its Increasing Use' [2009] 17 European Journal of Crime, Criminal Law and Criminal Justice, 165, 167-168; Van Sliedregt [15] 263.
[31] Herbert Packer, 'Two Models of the Criminal Process' [1964] 113 University of Pennsylvania Law Review 1, 16; Stewart [18].
[32] ibid 411.
[33] Marco Mendola, 'One Step Further in the "Surveillance Society": The Case of Predictive Policing' [2016] Leiden University Tech and Law Center.
[34] *Case of Asan Rushti v. Austria* App no 28389/95 (ECHR, 21 March 2000) para 31.
[35] ibid.
[36] Council of Europe, *Guide on Article 6 of the European Convention on Human Rights; Right to a Fair Trial (Criminal Limb)* (2020) 62.

individual and may affect the determination of legal culpability in a system where verdict is reached by a jury of one's peers.[37] It would diminish the value of the Presumption to not apply its weight to all state authorities with influence on public opinion. The unique role of police is indicated by the Court's bright line distinctions between the police and other influential actors, such as media, political actors, and prosecutors.[38]

Finally, the Court has demonstrated the importance of maintaining the Presumption's essence, even when necessary to do so by pursuing avenues legally based in other fundamental rights. For instance, it has stopped short of extending the Presumption to informal manifestations of suspicion, but it has recognized the power of a label or categorization of suspicious for the practical effects it has on an individual's rights. [39] In its reasoning that was very much a justification supported by the Presumption, it held that such acts are unlawful according to Article 8 of the ECHR.[40]

The category of suspect, or even defendant, brings with it some degree of deprivation of liberty as well as other unavoidable varieties of treatment to which an innocent person will not be subjected.[41] Similarly, an individual deemed to be in a class of persons likely to commit a crime, becomes subject to a different type of treatment than the individual considered innocent.[42] The stigma of an arrest *de facto* labels and separates the guilty from the rest of society.[43] The Court in *S. and Marper v The United Kingdom* accordingly supported the value of an individual's right to be seen as innocent and asserted that there is a 'reputational' aspect to the Presumption, such that a finding of innocence should not be undermined by a stigma of guilt.[44]

4 Applying the Presumption to Policing by Risk Assessment

As addressed above, the Presumption is intended to guard against impediments to a fair trial, holding in place procedural standards and supporting additional fundamental rights. As stated by the ECtHR, there is no value to the Presumption should its application be so narrow in scope that myriad processes surrounding the trial may cause the same injustices it is intended to shield.[45] This section asserts that predictive policing should be subject to the Presumption for three reasons. First, the use of criminal records as the main source of data in risk assessments is problematic in that it does not properly

[37] Campbell [23] 685; Clark [17].
[38] Council of Europe [34] 65.
[39] Galetta [15].
[40] *Case of S. and Marper v. The United Kingdom* App nos 30562/04 and 30566/04 (ECHR 4 December 2008).
[41] Peter DeAngelis, 'Racial Profiling and the Presumption of Innocence' [2014] 43 Netherlands Journal of Legal Philosophy 1, 43, 54.
[42] Pamela Ferguson, 'The Presumption of Innocence and Its Role in The Criminal Process' [2016] Criminal Law Forum 27, 131, 141.
[43] David Wolitz, 'The Stigma of Conviction: Coram Nobis, Civil Disabilities, and the Right to Clear One's Name' [2009] Brigham Young University Law Review 5, 1277, 1276.
[44] Liz Campbell [23]; O.J. Gstrein, A. Bunnik, and A. Zwitter, 'Ethical, Legal and Social Challenges of Predictive Policing' [2019] 3 Catolica Law Review 77, 10.
[45] *Case of Allenet de Ribemont v France* App no 15176/89 (ECHR 1995) para 94.

extract individuals' records whose arrest was not followed by a conviction. Should a risk assessment assign an individual the status of suspect based on categorical profiling, a violation of the Presumption may occur. Secondly, police act as public authorities with the ability to prejudice the fairness of a trial if making a statement or acting in a way that causes the prejudgment of a suspect. Finally, predictive policing with the aid of risk assessments is a *de facto* use of investigatory techniques for preventing crime. This level of intrusion on the individual without targeted suspicion is parallel in process to a pre-trial criminal investigation subsequent to a charge and should therefore also require the same level of procedural protection as found in the Presumption.

4.1 Criminal records as incomplete data

Risk assessments are as effective as the data which inform their output, the more data that are used the more accurate the crime forecast.[46] In the case of risk assessments to predict likely crime, historical crime data are the most important information available.[47] Some software rely almost exclusively upon static crime data as collected and recorded by police agencies,[48] and include any combination of arrest records, calls for assistance, or non-custodial stops.[49] Due to the emphasis on historic data, prior offenders and those who share their characteristics, are often considered to be likely future offenders.[50] Indeed, data on prior offenders inform both geographic and individual predictive profiles, based on a perception that these shared traits indicate a 'propensity to commit harmful behaviour'.[51] The ECtHR has held that the Presumption is not violated by the mere retention of acquitted individuals' data in a law enforcement database.[52] However, in the case of predictive policing, it is not the database at issue but rather its use to inform police action toward individuals labelled as 'suspect'. [53]

This article uses the term 'static' to describe crime data, referring to the fact that data based on arrests, for instance, do not account for the subsequent disposition of charges. For the purposes of a predictive software using historical arrest data, the arrest stands alone as a historical fact, regardless of whether a conviction follows.[54] For the purposes of police records this is a legitimate means of recording data. Police do not collect data for the sole purpose of risk assessments and so these data are not recorded in a manner

---

[46] Matteo Pasquinelli, 'How a Machine Learns and Fails - a Grammar of Error for Artificial Intelligence' [2019] Journal for Digital Culture (Spectres of AI no 5).
[47] Perry [5].
[48] Hardyns and Rummens [2].
[49] Kristian Lum and William Isaac, 'To Predict and Serve?' [2016] 13 Significance 14.
[50] For more information on the inaccuracies and the inability to correct police records, Interview with Phillip Atiba Goff, 'How Police Reports Became Bulletproof' (*All Things Considered*, 26 May 2021) <https://www.npr.org/2021/05/26/1000598495/how-police-reports-became-bulletproof> accessed 7 October 2021.
[51] Campbell [23] 25; Mendola [31] 15.
[52] Campbell [23] 698; *Marper v. UK* [38].
[53] Elizabeth Joh, 'Policing by Numbers: Big Data and the Fourth Amendment' [2014] 89 Washington Law Review 35, 55.
[54] Aleš Završnik, 'Algorithmic Justice: Algorithms and Big Data in Criminal Justice Settings' [2019] 18 European Journal of Criminology 5.

aligned with the legal implications of a risk assessment.[55] In this way, an inference is made that an individual who was charged, even if later found by a court to be innocent, will be algorithmically processed as one found guilty.

This misapplication of data directly invokes the Court's assertion that authorities may violate the tenets of the Presumption when ascribing guilt in the absence of a guilty finding.[56] The Court makes a clear distinction between found not guilty and not found guilty, determining the Presumption applies to both. In the case of *Clive v. Germany* the Court held that in order to ensure the Presumption is 'practical and effective', it applies beyond the context of pending criminal proceedings, including following an acquittal, dismissal, or discontinuance.[57] The Court further held in *Minelli v. Switzerland* that in the absence of a guilty finding a judicial opinion which suggests that an individual is guilty will violate the Presumption, even in the absence of a formal finding.[58] In *Allen v. United Kingdom*, the Court offered this:

> in keeping with the need to ensure that the right guaranteed by Article 6 § 2 is practical and effective … Its general aim … to protect individuals who have been acquitted of a criminal charge, or in respect of whom criminal proceedings have been discontinued, from being treated by public officials and authorities as though they are in fact guilty of the offence charged. In these cases, the presumption of innocence has already operated, through the application at trial of the various requirements inherent in the procedural guarantee it affords, to prevent an unfair criminal conviction being imposed. Without protection to ensure respect for the acquittal or the discontinuation decision in any other proceedings, the fairtrial guarantees of Article 6 § 2 could risk becoming theoretical and illusory.[59]

In the instant case of predictive policing, in which police records include charges which resulted in conviction, acquittal, dismissal or some other discontinuance of proceedings, there must be a clear distinction between the data used. In relying upon risk assessments to generate suspicion, police may be acting in reliance on data relative to a prior arrest, despite the fact that a court may have subsequently found him/her not guilty of the offense. It therefore follows that predictive policing predicated on the use of any arrest record that resulted in anything other than a conviction may cause a violation of the Presumption when relied upon by policing authorities to identify potential 'suspects'.[60]

---

[55] Phillip Goff and Kimberly Barsamian Kahn, 'Racial Bias in Policing: Why We Know Less Than We Should' [2012] 6 Social Issues and Policy Review 177; Joh [51].

[56] *Case of Sekanina v. Austria* App no 13126/87 (ECHR 25 August 1993) para 37; Galetta [15] citing *Sekanina v. Austria.*

[57] *Case of Cleve. v. Germany* App no 48144/09 (ECHR 15 January 2015) para 35.

[58] *Case of Minelli v. Switzerland* App no 8660/79 (ECHR 25 March 1983) para 37.

[59] *Case of Allenet de Ribemont v. France* [43] para 94.

[60] *Case of Minelli v. Switzerland* [57] 37.

4.2 Clarifying prejudicial acts

If the police data used are incomplete as described above, acts taken by the police may constitute prejudicial behavior by a public authority or an individual of public standing.[61] The ECtHR has held that expressive acts by public authorities which have the effect of prejudicing an individual in the course of a criminal trial may constitute a violation of the Presumption. In *Allenet de Ribemont v. France*, the Court held that the obligations imposed by the Presumption are not limited to officials of criminal courts, but also other authorities.[62] In its assessment, the Court held that a proclamation of guilt by a senior police officer 'firstly, encouraged the public to believe him guilty and, secondly, prejudged the assessment of the facts by the competent judicial authority'.[63] It further determined that statements made by police in the course of an investigation parallel to arrest and detention, have the 'foreseeable' effect of prejudicing the defendant in a public manner, amounting to prejudgment.[64] Therefore the Presumption can be infringed by any public authorities who in their official capacity may have some bearing on the outcome of a trial.[65]

The Court distinguishes legitimate acts, such as the factual notification to the public of the existence of a criminal investigation.[66] The Court has further clarified that acts which stand in as indirect statements on believed guilt may amount to a violation of the Presumption, whereas a passive utterance of suspicion alone will unlikely constitute a violation.[67] The Court has further indicated that, '…a fundamental distinction must be made between a statement that someone is merely suspected of having committed a crime and a clear declaration'.[68] It is then a reasonable assertion that the public arrest of an individual in the presence of bystanders is an inherently clear declaration of official suspicion against an individual that supersedes 'mere' suspicion.[69] In its analysis in *Marper*, the Court held that though not a formal declaration of guilt, the retention of acquitted individuals' DNA equates to treating innocent people as guilty, in practical contradiction of an acquittal and outside the spirit of the Presumption.[70] Predictive policing which is predicated on the use of data, similarly relies on the retention of records of arrests which may have also ended in acquittal.[71] Inclusion in a DNA database, though

---

[61] Dovydas Vitkauskas and Grigoriy Dikov, *Protecting the Right to a Fair Trial Under the European Convention on Human Rights; A Handbook for Legal Practitioners* (2nd ed Council of Europe 2017) 113 <https://rm.coe.int/ protecting-the-right-to-a-fair-trial-under-the-european-convention-on-/168075a4dd> accessed 10 Febru-ary 2021.

[62] *Case of Allenet de Ribemont v. France* [43] para 33.

[63] ibid 41.

[64] *Case of Allenet de Ribemont v. France* [43] 37; Campbell [23] 694.

[65] Galetta [15] citing *Case of Allenet de Ribemont v. France* [43].

[66] Council of Europe [34] 65.

[67] Liz Campbell, 'A Rights-Based Analysis of DNA Retention' [2012] Criminal Law Review 12, 7.

[68] *Case of Ismoilov and Others v. Russia* App no 2947/06 (ECHR 1 December 2008) para 166.

[69] DeAngelis [39] 56.

[70] Campbell [65] 902-905.

[71] Campbell [23] 5-6, 21-23.

certainly more intrusive, is only used to compare against evidence. Risk assessments used by police are continuously and proactively inducing police action. This clearly meets the standard as described in *Marper*.

The Court in *Marper* further differentiated between official manifestations of perceived guilt by the state and stigma as a result of state action. In addressing the inclusion of acquitted individuals in a DNA repository, the Court further held that to be treated as guilty after having been found not guilty of an offence leads to a presumption against innocence and risks stigmatization.[72] In addition, it found that inclusion in a database used to locate criminals '…enlarges the category of 'suspect"',[73] and that this could not be considered necessary in light of the undue consequences on individuals' reputations.[74] It is therefore argued that risk assessments relying on historic crime data for the purpose of building suspicion similarly leads to categorizing individuals according to their degree of guilt, creating new distinctions between citizen, suspect, and defendant.[75] It is well documented that increased police encounters and scrutiny result in 'evidence-based' stigmatization,[76] which extends well beyond the criminal justice system into applications for jobs, housing, and credit.[77] Studies demonstrate that in both the U.K. and U.S. criminal justice systems, which utilize predictive policing at increasing levels, criminal stigmatization often extends to whole communities.

Because predictive policing is the identification of relevant factors to anticipate future crimes, it is logical and necessary that police utilize their own records to fuel risk assessments. However as established, police records are not often updated to reflect the disposition of legal processes and this creates an inaccuracy as regards the disposition of charges.[78] When risk assessments are used to generate suspicion and induce police action, acquitted or otherwise cleared individuals may become subject of police scrutiny based on disposed of charges. In this case, the Presumption of Innocence has been violated.

4.3 Prevention as investigation

Finally, it is argued here that the act of predictive policing mirrors a criminal investigation much more closely than a police patrol and therefore should be subject to the pro-

[72] Mendola[31] 15; Katerina Hadjimatheou, 'Surveillance, the Moral Presumption of Innocence, the Right to Be Free from Criminal Stigmatisation and Trust' (SURVEILLE Seventh Framework Programme, 30 September 2013).

[73] The ECHR refers to this as the 'pérennisation de la catégorie de "suspect"' Galetta [15].

[74] *Case of S. and Marper v. United Kingdom* [38]; ibid.

[75] Amber Marks, Benjamin Bowling and Colman Keenan, 'Automatic Justice? Technology, Crime and Social Control' (Roger Brownsword, Eloise Scotford and Karen Yeung eds) *The Oxford Handbook of the Law and Regulation of Technology* (OUP 2017); Andrew Ashworth, 'Four Threats to the Presumption of Innocence' [2006] 10 The International Journal of Evidence & Proof 241.

[76] Gstrein, Bunnik, and Zwitter [42] 10; Duff [13] 13.

[77] Amnesty International [10].

[78] Mendola [31] 17.

tection of the Presumption afforded subjects of criminal investigations. Criminal investigations, a recognized pre-trial process, require that police resources be utilized to determine the perpetrator of a crime which has been committed and catalogued by police. Criminal investigation is a process which is organized and methodical, using all available information.[79] The use of tools that may generate a profile based on factual evidence allow for the targeting of individuals who are in some way connected to the details of a crime.[80] Predictive policing uses a similar method, whereby the evidence and elements of former crimes are used to build the profile of individuals who may be similar to past offenders and therefore potentially likely to commit future crimes.[81] This thereby creates a categorical profile for suspects who warrant heightened police attention in the form of what may equate to a criminal investigation.

Whereas earlier forms of sophisticated techniques, or first-generation forensic technology, were used to confirm or deny suspicion, the second generation is capable of being used for proactive investigations, perfectly illustrated by risk assessments.[82] The main difference between the two methods is the commission of a crime, a temporal place holder at which a charge may be filed, and the pre-trial protections are afforded the suspect. Without the crime there is no charge and the potential suspect, so treated by police, may not also garner the protections of the legal process. As alluded to above, this has real, detrimental effects on an individuals' rights. It is argued that as action against the individual is increasingly executed in a preventative context, the protections against arbitrary state action should follow accordingly.[83]

5 Conclusion

Predictive policing, risk assessments, and other uses of surveillance technology for law enforcement show no signs of declining, but instead are becoming increasingly prominent. This article has argued that as these powerful enforcement tools become more ubiquitous, the attendant due process protections provided for within criminal justice must also evolve. Further, it asserted that the events leading to a criminal trial are set in motion by policing and that as a result policing practices carry numerous implications to the fairness and outcome of a trial. In the case of predictive policing, the legitimacy of the Presumption of Innocence is tested at its very core. Due to the imprecise manner in which police data are used to include charges later dismissed for forming suspicion; the ability of the police to act as state authorities with the potential to prejudice a trial outcome; and the intensive investigation processes invoked by predictive policing, a deep misalignment is forming between the application of Article 6.2 of the ECHR and policing. Without

[79] Garland [3].
[80] DeAngelis [39].
[81] Marks, Bowling and Keenan [73].
[82] ibid.
[83] Lucia Sommerer, 'The Presumption of Innocence's Janus Head in Data-Driven Government' (Emre Bayamlioglu, Irina Baraliuc, Liisa Janssens, Mireille Hildebrandt eds) *Being Profiled* (Amsterdam University Press 2018).

accounting for the new realities inherent in building a criminal charge, applying the Presumption to the resulting criminal trial process is likely to be hollow in effect.

References

'Guide on Article 6 of the European Convention on Human Rights; Right to a Fair Trial (Criminal Limb)' (Council of Europe 2020) <https://www.echr.coe.int/documents/guide_art_6_criminal_eng.pdf> accessed 7 October 2021

'Trapped in the Matrix: Secrecy, Stigma, and Bias in the Met's Gangs Database' (Amnesty International 2018) <https://www.amnesty.org.uk/files/reports/Trapped%20in%20the%20Matrix%20Amnesty%20report.pdf> accessed 7 October 2021

Ashworth A, 'Four Threats to the Presumption of Innocence' [2006] 10 The International Journal of Evidence & Proof 241

Benbouzid B, 'To Predict and to Manage. Predictive Policing in the United States' [2019] Big Data & Society 1

Brown N, 'Robert Park and Ernest Burgess: Urban Ecology Studies, 1925' (2002) CSISS Classics UC Santa Barbara <https://escholarship.org/uc/item/6f39q98d> accessed 7 October 2021

Campbell L, 'A Rights-Based Analysis of DNA Retention' [2012] Criminal Law Review 12, 889

Campbell L, 'Criminal Labels, The European Convention on Human Rights and The Presumption of Innocence' [2013] 76 The Modern Law Review 4, 681

Case of Allenet de Ribemont v. France App no 15175/89 (ECHR 10 February 1995)

Case of Asan Rushti v. Austria App no 28389/95 (ECHR 21 March 2000)

Case of Cleve. v. Germany App no 48144/09 (ECHR 15 January 2015)

Case of Minelli v. Switzerland App no 8660/79 (ECHR 25 March 1983)

Case of S. and Marper v. The United Kingdom App no 30562/04 and 30566/04 (ECHR 4 December 2008)

Clark S, 'The Juror, the Citizen, and the Human Being: The Presumption of Innocence and the Burden of Judgment' [2014] Criminal Law and Philosophy 8, 421

Cyr K, 'The Police Officer's Plight: The Intersection of Policing and the Law' [2015] 52 Alberta Law Review 4, 889

DeAngelis P,' Racial Profiling and the Presumption of Innocence' [2014] 43 Netherlands Journal of Legal Philosophy 1, 43

Duff RA, 'Who Must Presume Whom to Be Innocent of What?' [2013] 13 Netherlands Journal of Legal Philosophy 12

Ferguson A, 'Predictive Policing and Reasonable Suspicion' [2012] 62 Emory Law Journal 259, 261

Ferguson P, 'The Presumption of Innocence and Its Role in The Criminal Process' [2016] Criminal Law Forum 27, 131

Galetta A, 'The Changing Nature of the Presumption of Innocence in Today's Surveillance Societies: Rewrite Human Rights or Regulate the Use of Surveillance Technologies' [2013] 4 European Journal of Law and Technology 2

Garland D, *The Culture of Control* (University of Chicago Press 2001)

Gless S, 'Automated Suspicion - and Evidence?' (Facial recognition vs. Criminal Justice, Council of Europe AI & Law Webinar Series, February 2021) <https://www.coe.int/en/web/artificial-intelligence/-/ai-law-webinar-9-facial-recognition-vs-criminal-justice?fbclid=IwAR3Y-uYro9TEcar-qr1sJAMJUaThTv9Izhs5L9oNSitOIVaFJuR4Z5sE5Jw> accessed 2 February 2021

Goff P A, 'How Police Reports Became Bulletproof' (*All Things Considered*, May 2021) <https://www.npr.org/2021/05/26/1000598495/how-police-reports-became-bulletproof> accessed 7 October 2021

—— and Barsamian Khan K, 'Racial Bias in Policing: Why We Know Less Than We Should' [2012] 6 Social Issues and Policy Review 1, 177

Gstrein O, Bunnik A, and Zwitter A, 'Ethical, Legal and Social Challenges of Predictive Policing' [2019] 3 Catolica Law Review 3, 77

Hadjimatheou K, 'Surveillance, the Moral Presumption of Innocence, the Right to Be Free from Criminal Stigmatisation and Trust' (SURVEILLE Seventh Framework Programme, September 2013)

Hardyns, W, and Rummens A, 'Predictive Policing as a New Tool for Law Enforcement? Recent Developments and Challenges' [2018] 24 Eur J Crim Policy Res 1, 201

Innes M, 'The Art, Craft, and Science of Policing' (Peter Cane and Herbert M. Kritzer eds), *The Oxford Handbook of Empirical Legal Research* (OUP 2010)

Joh E, 'Policing by Numbers: Big Data and the Fourth Amendment' [2014] 89 Washington Law Review 35

Jones T, 'Governing Security: Pluralization, Privatization, and Polarization in Crime Control and Policing' (Mike Maguire, Rod Morgan and Robert Reiner eds) *The Oxford Handbook of Criminology* (5th edn OUP 2007)

Jong F, and van Lent L, 'The Presumption of Innocence as a Counterfactual Principle' [2016] 12 Utrecht Law Review 32

Lum K, and Isaac W, 'To Predict and Serve' [2016] Significance

Marks A, Bowling B, and Keenan, C, 'Automatic Justice? Technology, Crime and Social Control' (Roger Brownsword, Eloise Scotford and Karen Yeung eds) *The Oxford Handbook of the Law and Regulation of Technology* (OUP 2017)

Mendola M, 'One Step Further in the "Surveillance Society": The Case of Predictive Policing' [2016] Leiden University Tech and Law Center

Packer H, 'Two Models of the Criminal Process' [1964] 113 *University of Pennsylvania Law Review* 1

Pasquinelli M, 'How a Machine Learns and Fails - a Grammar of Error for Artificial Intelligence' [2019] Journal for Digital Cultures 5

Perry W, McInnis B, Price C, Smith S, Hollywood J, and Perry W, 'Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations' (RAND 2013) <http://ebookcentral.proquest.com/lib/unilu-ebooks/detail.action?docID=1437438> accessed 9 January 2020

Reilly D, *Finding the Truth with Criminal Investigation* (Rowman & Littlefield 2019)

Sommerer L, 'The Presumption of Innocence's Janus Head in Data-Driven Government' (Emre Bayamlioglu, Irina Baraliuc, Liisa Janssens, Mireille Hildebrandt eds) *Being Profiled* (Amsterdam University Press 2018)

Stevens L, 'Pre-Trial Detention: The Presumption of Innocence and Article 5 of the European Convention on Human Rights Cannot and Does Not Limit Its Increasing Use' [2009] 17 European Journal of Crime, Criminal Law and Criminal Justice 165

Stewart H, 'The Right to Be Presumed Innocent' [2014] 8 Criminal Law and Philosophy 407 <https://doi.org/10.1007/s11572-013-9233-x> accessed 25 March 2021

Stuckenberg C, 'Who Is Presumed Innocent of What by Whom?' [2014] 8 Criminal Law and Philosophy 301

Van Sliedregt E, 'A Contemporary Reflection on the Presumption of Innocence' [2009] 80 Revue Internationale de Droit Penal 1, 247

Vitkauskas D, and Dikov G, *Protecting the Right to a Fair Trial Under the European Convention on Human Rights; A Handbook for Legal Practitioners* (2nd ed Council of Europe 2017) 113 <https://rm.coe.int/protecting-the-right-to-a-fair-trial-under-the-european-convention-on-/168075a4dd> accessed 10 February 2021

Weigend T, 'Assuming That the Defendant Is Not Guilty: The Presumption of Innocence in the German System of Criminal Justice' [2014] 8 Criminal Law and Philosophy 285 <https://doi.org/10.1007/s11572-013-9271-4> accessed 10 May 2021

Weisburd D, 'Does Hot Spots Policing Inevitably Lead to Unfair and Abusive Police Practices, or Can We Maximize Both Fairness and Effectiveness in the New Proactive Policing' [2016] University of Chicago Legal Forum 16, 661

—— 'The Law of Crime Concentration and the Criminology of Place' [2015] 53 *Criminology* 2, 133

Wolitz D, 'The Stigma of Conviction: Coram Nobis, Civil Disabilities, and the Right to Clear One's Name' [2009] Brigham Young University Law Review 5, 1277

Završnik A, 'Algorithmic Justice: Algorithms and Big Data in Criminal Justice Settings' [2019] European Journal of Criminology

CLICK, COLLECT AND CALCULATE: THE GROWING IMPORTANCE OF BIG DATA IN PREDICTING FUTURE CRIMINAL BEHAVIOUR

By Julia Heilemann[*]

Abstract

The age of digitalization, able to offer new opportunities to states and citizens in all areas of life, has also reached law enforcement authorities. In the realm of criminal justice, new technologies can now provide police officers with software able to predict the future commission of a criminal offence and thereby assist in their prevention and reduction. However, in order to enhance the effective functioning of these predictive policing software, authorities require large amounts of data, usually gathered by private corporations. These companies play an increasing role in transferring so-called 'big data', often contrary to the liking of their customers. Whereas private to public data sharing is encouraged at the European level, the European Union has failed to take a clear stance and adopt a framework on the use of such practices for predictive policing purposes.

1 Introduction

Like, comment, tag, share – and suddenly, your post goes viral. The rise of the technological era has impacted everyone – people are constantly connected online, checking Facebook, messaging through WhatsApp, listening to music on Spotify or shopping on Amazon. The information shared by consumers on these platforms is extremely valuable and therefore retained and analysed by the respective companies so as to enhance their services.[1] This phenomenon is called data analytics and the information gathered therefrom can have an impact on the products of tomorrow.

But what if this information was also used for other purposes than those initially intended, for example to predict crime? Wouldn't this be beneficial for society as a whole if available data could help make cities more secure? There are no additional steps that citizens would have to take in order to be able to assist law enforcement authorities, and yet, the answers are not that simple.

So-called predictive policing software has originally relied on historical data and crime reports to predict the occurrence of future offences.[2] The availability of big data would allow law enforcement authorities the opportunity to adapt this software to the current technological age, thereby enhancing the software and its findings. At the European level,

[*] Alumna of Utrecht University (2021); LL.M. in European Criminal Justice in a Global Context. For correspondence: <julia.heilemann@hotmail.com>.

[1] Richard Hurley, *Big Data: A Guide to Big Data Trends, Artificial Intelligence, Machine Learning, Predictive Analytics, Internet of Things, Data Science, Data Analytics, Business Intelligence, and Data Mining* (Independently Published 2019) 27-40.

[2] Wim Hardyns and Anneleen Rummens, 'Predictive Policing as a New Tool for Law Enforcement? Recent Developments and Challenges' (2017) 24 European Journal on Criminal Policy and Research 201, 202.

more and more countries are reverting to artificial intelligence (AI) in order to prevent and reduce crime rates within their cities,[3] as the systems predict not only the location of a future crime, but also the behaviour of a specific person, ie how likely a person is to commit a crime.[4] Yet, this practice has also faced backlash, notably due to privacy and data protection infringements, discriminatory and bias practices or the involvement of private actors in traditionally law enforcement authorities' activities.

2 Predictive Policing

In simple terms, predictive policing refers to the introducing of data into software which can then predict the likely occurrence of a future crime.[5] The process of analysing data and drawing conclusions from it is also referred to as "predictive analysis".[6] By using certain data, software can predict four types of occurrences: (i) who will be a perpetrator, (ii) what a perpetrator's identity or profile is like, (iii) who will be a victim, or (iv) the time and place when a criminal offence is likely to occur.[7] Although all these predictions are possible, software mainly has recourse to the latter option,[8] as the former ones (and especially predicting offenders) have faced considerable criticism from human rights activists, for example because of unfair profiling.[9] Nevertheless, recent developments are witnessing a growing shift towards predicting offenders, especially as more person-related data is made available each day.[10]

Predictive policing software essentially relies on two elements – AI and data. Whereas traditional software used historical crime data, as well as information about the weather, seasons and times of day to make their predictions,[11] the 21st century and technological developments have brought additional large amounts of data to the forefront – so-called "big data". This involves the registration of every click that is made when online shopping, each post and message that is sent on social media networks, or each use of smart home gadgets (also known as "internet of things"). In short, big data refers to the collection of mass data that may relate to anything individuals do/click/like/post/shop/listen

[3] Martin Degeling and Bettina Berendt, 'What is wrong about Robocops as consultants? A technology-centric critique of predictive policing' (2018) 33 AI & Soc 347, 348.
[4] Oskar Gstrein, Anno Bunnik and Andrej Zwitter, 'Review of ethical, legal & social issues impacting Predictive Policing' (2018) Cutting Crime Impact, 17.
[5] Orla Lynskey, 'Criminal justice profiling and EU data protection law: precarious protection from predictive policing' (2019) 15 International Journal of Law in Context 162, 162.
[6] Febe Liagre, 'Predictive Policing Recommendations paper' (2016) European Crime Prevention Network 3.
[7] Gloria González Fuster, 'Artificial Intelligence and Law Enforcement: Impact on Fundamental Rights' (2020) European Parliament 22.
[8] Hardyns and Rummens (n 2) 203.
[9] Degeling and Berendt (n 3) 348.
[10] Andrew Guthrie Ferguson, 'Policing Predictive Policing' (2017) 94 Washington University Law Review 1109, 1137.
[11] Fieke Jansen, 'Data Driven Policing in the Context of Europe' (2018) Cardiff University <https://data-justiceproject.net/wp-content/uploads/sites/30/2019/05/Report-Data-Driven-Policing-EU.pdf> accessed 26 March 2021.

to etc.[12] The information as such may not be of much importance on its own, however, when brought together with other data and analysed significant conclusions can be drawn from it.[13] Based on the data inserted in the software, AI systems are then able to make certain predictions using an algorithm.[14] While business corporations will be able to optimize their services, health authorities can improve the allocation of staff in hospitals, and banks can analyse the risks related to customers requesting a banking loan.[15]

In terms of predictive policing, big data can imply the improving of software and thus increase the offences that could be prevented. By including and analysing additional information, such as Twitter and Facebook posts for example, more predictions could be made, and new patterns can be identified by law enforcement authorities.[16]

Predictive policing software can be used for all types of crime provided sufficient information about the risks of a crime occurring and data related to it can be collected.[17] Accordingly, predictions can be made if the following conditions are met (i) there is a willingness from victims to report crimes and the police adequately registers those, (ii) the crime occurs at a high frequency thus ensuring that enough data is available, and (iii) the location and time when a crime occurs can be established relatively precisely, for example by using geo-coding and time intervals.[18] At the moment, predictive policing software has been used for high-impact crimes – crimes that occur regularly and have a strong impact on the victims – such as burglaries, theft or drug-related crimes.[19] Money laundering and terrorism financing have also recently been added to the list.[20]

The objective of predictive policing is, of course, to predict and thereby prevent crimes by intervening in an area before the offence is committed. Whereas European countries are deploying such software more and more,[21] the US has already had a multitude of projects able to show vast effectiveness. According to reports, Richmond was able to decrease random gunfire rates on New Year's Eve by 47% by anticipating the location and

[12] Hurley (n 1) 3-16.
[13] ibid 4.
[14] Tim Lau, 'Predictive Policing Explained: Attempts to forecast crime with algorithmic techniques could reinforce existing racial biases in the criminal justice system' (*Brennan Center for Justice*, 1 April 2020) <https://www.brennancenter.org/our-work/research-reports/predictive-policing-explained> accessed 1 June 2021.
[15] Hurley (n 1) 35-40.
[16] Rosamunde van Brakel, 'Pre-emptive Big Data Surveillance and its (Dis)empowering Consequences: The Case of Predictive Policing' in Van der Sloot et al. (eds), *Exploring the Boundaries of Big Data* (Amsterdam university Press 2016) 4.
[17] Hardyns and Rummens (n 2) 205.
[18] ibid.
[19] Liagre (n 6) 5.
[20] González Fuster (n 7) 23.
[21] Jan-David Franke, 'A year in surveillance' (*About: Intel*, 2020) <https://aboutintel.eu/a-year-in-surveillance/> accessed 26 March 2021.

time of future crimes, thereby saving the police department $15.000 in personnel costs,[22] whereas Santa Cruz reduced property thefts by 19%.[23] These numbers show the benefits of predictive policing in reducing crime, increasing cities' safety, and cutting costs for law enforcement authorities.[24] However, the benefits also exceed the prevention of crime and reduction of costs. Data analysis can reveal new trends and crime patterns that were not originally known to police departments and thus adjust their response mechanisms.[25] Furthermore, lasting benefits of predictive policing may also be achieved through projects similar to the Cedar Grove Initiative. This association of police officers, schools, social workers and civil society organizations has used big data to discern specific areas in a city that need attention to reduce crime rates. Factors taken into consideration by the project include the quality of school education, unemployment rates or how "green" an area is. If an algorithm predicts high numbers of crimes in an area scoring poorly in these categories, it may be a sign for the city administration to invest in urban spaces, improving the schooling system, strengthening family support etc.[26] Predictive policing is thus not only oriented to diminish crime, but also to analyse the factors leading to an area being designated as "high risk" and addressing its root causes.

Nevertheless, predictive policing and especially the use of big data in such software has also faced criticism, notably due to privacy rights concerns. Scholars argue that insufficient information is made available regarding which data and how much of it is being collected to feed the algorithms.[27] Moreover, the intrusion into citizen's private life is severe, particularly when using big data, as limitations and restrictions on data transfers to third parties apply. By being aware that their data may be surveyed and used by governmental authorities, individuals could see themselves restricted in freely expressing themselves and communicating online.[28]

As a still emerging trend subject to few rules, it is also unclear who will bear the responsibility and be accountable in case of unforeseen incidents. It is uncertain whether the company creating the software (often private corporations), the data scientist who developed the algorithm or the law enforcement officers who used the service should bear the

[22] Beth Pearsall, 'Predictive Policing: The Future of Law Enforcement?' (2010) 10 National Institute of Justice Journal 266, 17.
[23] Miriam Jones, 'Predictive Policing a Success in Santa Cruz, Calif - City sees significant reduction in property theft thanks in part to predictive crime modeling' (*Government Technology*, 8 October 2012) <https://www.govtech.com/public-safety/predictive-policing-a-success-in-santa-cruz-calif.html> accessed 9 October 2021.
[24] Franke (n 21).
[25] Jaevon George, 'How Social Media Is Changing Law Enforcement: Policing Big Data in the Connected World' (*LinkedIn*, 19 March 2019) <https://www.linkedin.com/pulse/how-social-media-changing-law-enforcement-policing-big-jaevon-george/> accessed 26 March 2021.
[26] van Brakel (n 16) 13.
[27] Hardyns and Rummens (n 2) 214.
[28] Sergio Carrera, Deirdre Curtin and Andrew Geddes, *20 Years Anniversary of the Tampere Programme: Europeanisation Dynamics of the EU Area of Freedom, Security and Justice* (European University Institute 2020) 301-302.

blame.[29] As long as the rules remain blurred in such cases, a high possibility remains that mistakes will not be accounted for, and that responsibility will, instead, be blamed on the technology itself, thus leaving victims with no redress.[30]

A third disadvantage of predictive policing is the outsourcing of police activities to private companies.[31] In addition to involving private actors in public service activities, software is often developed and owned by private corporations who sell their services to law enforcement authorities. This means that the algorithms used are developed by a corporation and often cannot be amended by the authorities themselves.[32] This not only increases the responsibility of the private actor providing the service, but it also limits the possibility of law enforcement authorities to adapt the software to their own needs or to have sufficient information concerning how the algorithm operates (and thus a lack of transparency).

Finally, a concern that has been prevalent in human-led as well as machine-learning policing is police-bias. Police officers have often been criticized for pursuing especially vulnerable citizens and profiling minorities from lower socio-economic demographics.[33] Similarly, machine learning using big data is said to have the same disadvantages.[34] This is manifested through the availability of an abundance of data relating to high-impact crimes (such as robberies, drug crimes etc., often committed by poorer citizens) and a lack of available data concerning white-collar crimes (often committed by richer citizens). Due to the discrepancy in available data, software focus, among others, on statistics mostly involving vulnerable citizens, and thus target the prevention of crime committed by less-privileged individuals disproportionately.[35]

3 Big Data in Predictive Policing

The "new oil" of the 21st century is nowadays considered to be data[36] – and large corporations are in possession of an abundance of it.[37] In the age of increasing digitalization, most activities related to daily life are registered, whether it is a doctor's appointment saved in an online calendar, presents purchased on Amazon or the time needed to read a book on a Kindle.[38] The involvement of private actors in predictive policing is inevitable

[29] van Brakel (n 16) 7.
[30] ibid.
[31] ibid.
[32] ibid 7-8.
[33] Jansen (n 11).
[34] van Brakel (n 16) 9.
[35] ibid 10.
[36] Kristina Irion and Giacomo Luchetta, 'Online Personal Data Processing and EU Data Protection Reform' (2013) CEPS, 7.
[37] European Commission, 'FAQs on business-to-government data sharing' (*European Commission*, 23 March 2020) <https://ec.europa.eu/digital-single-market/en/faq/faqs-business-government-data-sharing> accessed 15 April 2021.
[38] Kevin Miller, 'Total Surveillance, Big Data, and Predictive Crime Technology: Privacy's Perfect Storm' (2014) 19 Journal of Technology Law & Policy 105, 111-112.

if big data is to be used as these are the entities responsible for collecting, storing, and transferring the information to law enforcement authorities. Whereas data received from one single company may not contribute much to predictive policing, the combination of data gathered from a multitude of players helps creating a clear picture of current trends, individual's preferences, and behaviours.[39] The benefits from such transfers may be vast, yet they have not remained uncriticised.

Sources of big data in the realm of predictive policing can vary and include different types of data. One important aspect could be gathered from social media. So-called "social media mining" analyses posts, likes, comments, pictures and behaviours of individuals on platforms such as Twitter, Facebook or Instagram in order to draw conclusions.[40] Moreover, posts may be scanned for specific "suspicious" keywords, for example language inciting or encouraging the commission of criminal acts.[41] A differentiation between gathering data from public and private social media accounts can be made, with the former allowing any person to access the full range of information posted by an individual, and the latter restricting access to friends/followers. Whilst users posting information from a privately held account have a strong expectation to remain private, posts deliberately made available for access by the general public are intended for wider dissemination. Privacy protections and safeguards may thus differ depending on the type of account being held.[42] This, however, does not mean that public social media accounts do not afford any privacy protections. According to Edwards and Urquhart, arguments in favour of providing public social media information with privacy protections include the fact that data beyond a post's content relating to the social network (ie the "friends list") must be protected, public posts may publish information and tag persons with privately held accounts whose information is thereby involuntarily shared, and the changing nature of privacy settings making it difficult to "maintain a privacy status quo".[43]

Social media as well as other apps can also provide useful information regarding geolocation. Through the use of GPS, the positions and movements of citizens can be measured, thereby indicating potentially suspicious travel patterns.[44] Furthermore, tracking shopping activities and analysing the products bought by individuals at the supermarket, hardware store or online can also entail relevant information. For example, a person

[39] Anand Paul, Mark Cleverley, William Kerr, Frank Marzolini, Mike Reade and Stephen Russo, 'Smarter Cities Series: Understanding the IBM Approach to Public Safety' (2011) IBM, 8.
[40] Mohammad A Tayebi and Uwe Glässer, *Social Network Analysis in Predictive Policing* (Springer 2016) 10.
[41] Lina Dencik, Arne Hintz and Zoe Carey, 'Prediction, pre-emption and limits to dissent: Social media and big data uses for policing protests in the United Kingdom' (2018) 20(4) New Media & Society 1433, 1441.
[42] Lilian Edwards and Lachlan Urquhart, 'Privacy in public spaces: what expectations of privacy do we have in social media intelligence?' (2016) 24 International Journal of Law and Information Technology 279, 293.
[43] ibid 294-295.
[44] Matthew Williams, Pete Burnap and Luke Sloan, 'Crime sensing with Big Data: the Affordances and Limitations of Using Open-Source Communications to Estimate Crime Patterns' (2017) 57 British Journal of Criminology 320, 324.

purchasing scales, zip-lock bags and rubber bands may be flagged as a potential drug dealer.[45] Another data source also includes information gained from the emerging "internet of things" products,[46] which contribute to the development of "smart homes" or "smart cities" – ie areas in which activities related to everyday life are connected online and monitored. Products can range from smart security systems, toilets, sprinklers, fridges and light sensors.[47] The data gathered in this realm not only provides a great opportunity to learn how people behave in the comfort of their own home, when used in public they can also be seen as surveillance devices. While smart sensors and gadgets advertise their life-simplifying value, the data gathered can be of large significance in learning and predicting human behaviour.[48]

The data used for predictive policing can thus be extended from mere police records and crime statistics as traditionally used, and encompass social media data, smartphone activity, surveillance footage, shopping habits or GPS tracking.[49] In other words, the police will increasingly rely on "crime data, personal data, gang data, associational data, locational data, environmental data, and a growing web of sensor and surveillance sources" in the future.[50] According to González Fuster, the information contained in this data includes biographical, biometric, financial, location, associates and affiliations, employment and business, visa and immigration, travel and criminal and investigative history information.[51] The advantage of using big data lies in its volume and immediacy as it concerns real-time activities.[52] Acquiring this data significantly facilitates the creation of profiles able to predict criminal behaviours.

4 Obtaining Data through Public-Private Cooperation

As the above data is (in most cases) initially gathered by private entities, some form of cooperation is necessary between these bodies and public authorities in order for law enforcement officers to gain access to and use big data in predictive policing software. Although some content may be openly accessible (such as some social media content), this may not be enough to feed the algorithms.

---

[45] Andrew Guthrie Ferguson, *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement* (New York University Press 2017) 19.

[46] Alexander Babuta, 'Big Data and Policing: An Assessment of Law Enforcement Requirements, Expectations and Priorities' (2017) Royal United Services Institute for Defence and Security Studies, xi.

[47] Neil Wilkins, *Internet of Things: What You Need to Know About IoT, Big Data, Predictive Analytics, Artificial Intelligence, Machine Learning, Cybersecurity, Business Intelligence, Augmented Reality and Our Future* (Independently Published 2019) 25-32.

[48] ibid 51-57.

[49] Ferguson, *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement* (n 45) 2-3.

[50] ibid.

[51] González Fuster (n 7) 22.

[52] van Brakel (n 16) 118.

Necessary big data can be obtained by authorities imposing an obligation on private corporations to share their data.[53] Thus, a government body may explicitly require an entity to share specific contents. However, within the EU, the imposition of a wide-ranging obligation to share data was prohibited by the European Court of Justice, the Court holding in *Tele2* that the general and indiscriminate retention of data concerning all subscribers was contrary to the European Union (EU) Charter rights to private life and data protection.[54] Instead, such measures should be targeted and limited in their scope to what is strictly necessary.[55]

On the other hand, companies may also be required to save/retain certain data with a view to sharing it with the authorities in the future if necessary. An example of obligatory private to public data and information sharing can be found in the EU anti-money laundering framework.[56] According to Article 33 of the 5th Anti-Money Laundering Directive, so-called "obliged entities" (private parties) must share information with overseeing authorities in case of a suspicion of money laundering or terrorist financing.[57] Another example are the proposed European Production and Preservation Orders.[58] The proposal stipulates that a Member State would be able to send a Production Order, ie an order to send specific data, to a service provider located in another Member State. This service provider will have to comply and provide the requesting authorities with the requested data. These European developments show the increased active participation of private companies to contribute to securing European criminal justice. Although most (proposed) legislations relate to the gathering of electronic evidence, and not (yet) to the prevention of crimes, they demonstrate the possibilities available at EU level to compel private data sharing.

As state authorities cannot always rely on an obligation to transfer data, and certainly not on a general obligation, they may also be required to rely on voluntary data sharing by private actors. At the European level, this is currently advertised for the gathering of electronic evidence, notably through the negotiations of the Second Additional Protocol to the Cybercrime Convention of the Council of Europe. The proposal includes provisions on the direct cooperation with service providers to disclose subscriber and domain name registration information, yet provides service providers with the opportunity to

[53] Iain Cameron, 'European Union Law Restraints on Intelligence Activities' (2020) 33:3 International Journal of Intelligence and Counter Intelligence 452, 456.
[54] Joined Cases C-203/15 and C-698/15 *Tele2 Sverige AB and Watson* [2016] ECLI:EU:C:2016:970, para 112.
[55] ibid para 108.
[56] Directive (EU) 2018/843 of the European Parliament and of the Council of 30 May 2018 amending Directive (EU) 2015/849 on the prevention of the use of the financial system for the purposes of money laundering or terrorist financing, and amending Directives 2009/138/EC and 2013/36/EU [2018] OJ L156/3, arts 32-38.
[57] ibid art 33.
[58] Proposal for a Regulation of the European Parliament and of the Council on European Production and Preservation Orders for electronic evidence in criminal matters [2018] COM(2018) 225 final.

refuse such cooperation.[59] However, requests for such voluntary data sharing are limited to "the purposes of specific criminal investigations or proceedings".[60] As no investigations have been launched prior to the use of predictive policing software, it remains to be seen whether future legislative proposals will adopt a broader stance on the purposes for which data can be voluntarily shared.

In order to enhance and promote voluntary data sharing in the EU, the European Commission has brought forward its B2G initiative – business to government data sharing.[61] Although explanations and guidelines about this concept have been published, they are still not largely used.[62] Reasons for a lack of B2G refer to deficiencies in knowledge about gathering and identifying relevant data, the absence of professionals able to work in the field, ethical concerns and different laws in the Member States as to the validity and legitimacy of such programs.[63] Whilst the European Commission has also identified a lack of incentives for corporations to share their data with public authorities, it has reiterated the possibility of including compensations and tax incentives to induce participation.[64] Moreover, the Commission argues that a business would profit from increased reputation if it actively contributed to B2G.[65]

Finally, criticism levelled against Germany following a deadly terrorist attack on a Christmas market in 2017 provides a further example of data sharing between public and private entities. The German authorities were criticized for failures in the chain of transfer of data which could have predicted the existence of a threat emanating from the attacker. As a result, the Land Hesse acquired US software Palantir, a database allowing individual searches through various police databases, and which also includes information gained from social media received by US authorities.[66]

Despite the added value of involving big data in predictive policing, challenges regarding the gathering and transfer of such data persist. One of the core criticisms in this regard is that private bodies involved in data sharing are effectively taking part in and performing traditionally public tasks.[67] As illustrated by Rasch, individuals may not care if Google Maps collects GPS signals and is aware of their location – yet, when this data is transferred to the police, it can be viewed as unwanted surveillance and intrusion into

[59] Cybercrime Convention Committee, 'Second Additional Protocol to the Convention on Cybercrime on enhanced cooperation and disclosure of electronic evidence – Draft Protocol version 3' (May 2021) Council of Europe, arts 6-7.
[60] ibid paras 42-52.
[61] European Commission, 'Towards a European strategy on business-to-government data sharing for the public interest' (2020) European Union.
[62] European Commission, 'FAQs on business-to-government data sharing' (n 37).
[63] ibid.
[64] ibid.
[65] ibid.
[66] Timo Rademacher, 'Artificial Intelligence and Law Enforcement' in Thomas Wischmeyer and Timo Rademacher, *Regulating Artificial intelligence* (Springer 2020), 230.
[67]. Nadezhda Purtova, 'Between the GDPR and the Police Directive: navigating through the maze of information sharing in public-private partnerships' (2018) 8:1 International Data Privacy Law 52, 52-53.

private lives.[68] Even if data gathered by law enforcement authorities when occasionally using Google Maps to reach a friends' house does not seem to impact citizen's privacy rights to a large extend, the combination with data from other sources may. Privacy rights violations can thus emerge following the quantitative volume of data gathered and analysed together.[69] Furthermore, the shift to private actors performing public tasks may lead to the abuse of legal safeguards applicable to data sharing.[70] Whereas data retention for economic and commercial purposes may be subject to broader rules, law enforcement authorities gathering this data are subject to stricter limitations. By blurring the lines on how data is obtained, these safeguards may be by-passed effectively.[71]

Another major issue related to data collection concerns the lack of meaningful consent given by the consumer, ie the original data sharer.[72] Although social media posts may be considered "public" or "publicly available" many posts (or other activities) are only intended to be shared with friends.[73] Users are often unaware of the subsequent use of their activity, and their actions may be further analysed without accounting for context-specific aspects, thereby leading to misinterpretation of the information.[74] Boyd clearly states that "[j]ust because content is publicly accessible does not mean that it was meant to be consumed by just anyone".[75] Strict ethical guidelines thus need to be considered in assessing if and to what extent private data displayed in public settings can be analysed.[76] Whilst consumers initially agree to the terms and conditions of a corporation stipulating for the collection and further sharing of data, such consent should be "given freely and be specific, informed, and unambiguous".[77] However, the lack of alternative options available if a service wants to be used, coupled with the fact that a large majority of individuals never read these terms and conditions or privacy policies, notably due to their lengths and complexity,[78] compromises these safeguards. Whilst consent is thus technically obtained, questions concerning its real value remain.

---

[68] Mark Rasch, 'Public-Private Partnerships: Sharing Data, Compromising Privacy' (*Security Boulevard*, 23 May 2018) <https://securityboulevard.com/2018/05/public-private-partnerships-sharing-data-compromising-privacy/> accessed 15 April 2021.
[69] Miller (n 38) 135.
[70] Purtova (n 67).
[71] ibid 53.
[72] Babuta (n 46) 36.
[73] Danah Boyd and Kate Crawford, 'Critical Questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon' (2012) 15:5 Information, Communication & Society 662, 672.
[74] Babuta (n 46) 36.
[75] Boyd and Crawford (n 73).
[76] ibid.
[77] European Union Agency for Fundamental Rights and Council of Europe, *Handbook on European data protection law* (European Union Agency for Fundamental Rights and Council of Europe 2018) 365.
[78] David Berreby, 'Click to agree with what? No one reads terms of service, studies confirm' (*The Guardian*, 3 March 2017) <https://www.theguardian.com/technology/2017/mar/03/terms-of-service-online-contracts-fine-print> accessed 5 June 2021.

The scope of terms of services and privacy policies depends greatly among private corporations. Amazon[79] and Apple,[80] for example, have rather open-ended privacy statements subject to vast interpretation as they may share data with law enforcement authorities for the purposes of protecting the rights of others, national security or law enforcement. This broad wording could, indeed, imply that sharing data for predictive policing purposes would be legitimate for these corporations. Similarly, other entities like Booking.com,[81] Facebook[82] or Zoom[83] specifically allow data sharing for the detection and/or prevention of fraud, abuse, or illegal or harmful activities. Only few corporations seem to have a very restrictive privacy policy allowing them to share the relevant data merely in specific circumstances. Telegram, for example, may only share data in instances where users are believed to be terror suspects.[84] Although the wording of most privacy statements does not specifically indicate whether data sharing for predictive policing is allowed, their phrasing can be interpreted as not expressly restricting such purposes. Law enforcement authorities and private corporations could thus legitimately cooperate in sharing the relevant data. Over the last years, however, major corporations with lax privacy statement have seen a shift in consumer behaviour as customers are increasingly migrating to service providers offering stricter privacy policies – as perfectly exemplified in the decline of WhatsApp users and large increase of Telegram customers.[85]

5 An EU Approach to Public-Private Partnerships

Whilst the use of big data in predictive policing may be beneficial, the initial obtaining of the relevant data poses great challenges. In that regard, further research and clarifications at the European level are necessary, both from the Member States and the EU itself. To date, no explicit authorisation or prohibition for the use of predictive policing soft-

[79] Amazon, 'Amazon.com Privacy Notice' (*Amazon*, 12 February 2021) <https://www.amazon.com/gp/help/customer/display.html?nodeId=GX7NJQ4ZB8MHFRNJ#GUID-8966E75F-9B92-4A2B-BFD5-967D57513A40__SECTION_3DF674DAB5B7439FB2A9B4465BC3E0AC> accessed 5 June 2021.

[80] Apple, 'Apple Privacy Policy' (*Apple*, 1 June 2021) <https://www.apple.com/legal/privacy/en-ww/> accessed 5 June 2021.

[81] Booking.com, 'Privacy Statement' (*Booking.com*, 30 September 2020) <https://www.booking.com/general.en-gb.html?aid=356980;label=gog235jc-1DCBQoggJCBXRlcm1zSAdYA2ipAYgBAZgBB7gBF8gBDNgBA-gBAfgBAogCAagCA7gCjLXthQbAAgHSAiQ5NWVjY2UxYS03NjhhLTQ1MTItOWFhYi0zYWY2ZDkwNjU2YzHYAgTgAgE;sid=0b1393ade68be8c2c916ca4e17ef5a0f;sig=v1NHFxMJat;tmpl=docs/privacy-policy#personal-data-3rd-parties-shared-how> accessed 5 June 2021.

[82] Facebook, 'Data Policy' (*Facebook*) <https://www.facebook.com/policy.php#legal-requests-prevent-harm> accessed 5 June 2021.

[83] Zoom, 'Zoom Privacy Statement' (*Zoom*, 4 June 2021) <https://zoom.us/privacy#_70wwucijkyy> accessed 5 June 2021.

[84] Telegram, 'Telegram Privacy Policy' (*Telegram*) <https://telegram.org/privacy#8-who-your-personal-data-may-be-shared-with> accessed 5 June 2021.

[85] Jack Nicas, Mike Isaac and Sheera Frenkel, 'Millions Flock to Telegram and Signal as Fears Grow Over Big Tech' (*The New York Times*, 13 January 2021) <https://www.nytimes.com/2021/01/13/technology/telegram-signal-apps-big-tech.html#:~:text=Telegram%20has%20been%20particularly%20popular,Facebook%20and%20Twitter%20limited%20Mr> accessed 5 June 2021.

ware can be found in EU legislation. To the contrary, opinions remain divided, with European Commission Vice-President for Digital Policy, Vestager, arguing that it should not be allowed at all.[86] Whilst some Directives and Regulations refer to the transfer of data and the protection of privacy and data protection rights,[87] no instruments refer to the use of data for predictive policing. Nevertheless, the EU has taken steps towards regulating the use of AI, algorithms and big data in the past years, notably through its 2018 Communications about AI[88] and the idea of a European data space.[89]

European Commission President Von der Leyen has set a strong agenda for embracing and further developing digitalization.[90] Following its 2018 communication, the Commission published its European Data Strategy in February 2020.[91] The goal is to create a European data space where (personal and non-personal) data can flow freely and be accessible to actors ranging from start-ups to tech giants or governmental bodies.[92] At the same time, this data space would consider the protection of fundamental interests and values as promoted by the EU.[93] The goal of such a strategy is to increase data access (and reuse) to all actors, independently of the amount of data they themselves have generated.[94] This would promote the creation of new infrastructures and initiatives, thereby increasing European competence and dominance in this sector and providing a level-playing field with leaders in this sphere – notably China and the US.[95]

---

[86] Samuel Stolton, 'Vestager warns against predictive policing in Artificial Intelligence' (*EurActiv*, 30 June 2020) <https://www.euractiv.com/section/digital/news/vestager-warns-against-predictive-policing-in-artificial-intelligence/> accessed 15 April 2021.

[87] Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications) [2002] OJ L201/37 (ePrivacy Directive); 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; 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 L119/1.

[88] Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Region on Artificial Intelligence for Europe [2018] COM(2018) 237 final.

[89] Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Region on "Towards a common European data space" [2018] COM(2018) 232 final.

[90] González Fuster (n 7) 8.

[91] Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Region on A European strategy for data [2020] COM(2020) 66 final.

[92] ibid 4-5.

[93] ibid.

[94] ibid 3.

[95] ibid.

Simultaneously to the Data Strategy, the Commission also published its White Paper on Artificial Intelligence.[96] The instrument acknowledged seven key requirements that should be considered when implementing AI as found by the High-Level Expert Group tasked with drafting Guidelines on trustworthy AI. These requirements include the need to have human oversight, a technical robustness and safety, privacy and data governance, transparency, diversity, non-discrimination and fairness, societal and environmental wellbeing, and accountability.[97] The Paper also recognized the lack of a common framework regulating AI in Europe – a deficit pointed out by the Member States themselves.[98] Due to missing overarching rules, Member States are adopting provisions at the domestic level, thereby leading to a fragmentation of the internal market and a lack of trust in new AI systems.[99] In this regard, the Commission set forth that a commonly regulated framework protecting all citizens equally was necessary. This could be achieved by adjusting existing legal frameworks and/or through the creation of a new legislative instrument.[100]

The establishment of a common framework on AI, including the use of predictive policing would indeed be beneficial. The EU and EDPS (European Data Protection Supervisor) have been criticized for failing to extensively address data protection and the use of AI in law enforcement.[101] Through the adoption of a new instrument more discussions could be sparked in that regard. Moreover, the adoption of a common approach would hinder the implementation of differing applications at the domestic level. A fragmented approach with regard to predictive policing could have negative implications for corporations providing the data as a multitude of legal obligations and restrictions would apply, leading to an overall lack of clarity. The EDPS has also acknowledged that predictive policing may not be adequately regulated under existing data protection instruments such as the GDPR.[102] This Regulation aims at protecting the rights of the individual person being affected – yet predictive policing can also negatively impact a community as a whole. A future framework should thus reflect the impact of AI (and predictive policing) not only on citizens but also on the community as a collective. Finally, benefits related to a common legal framework also relate to increased transparency and accountability of

[96] European Commission 'White Paper on Artificial Intelligence – A European approach to excellence and trust' [2020] COM(2020) 65 final.
[97] ibid 9.
[98] ibid 10.
[99] ibid.
[100] ibid 13-22.
[101] González Fuster (n 7) 18-19.
[102] European Data Protection Supervisor, 'Opinion 4/2020 EDPS Opinion on the European Commission's White Paper on Artificial Intelligence – A European approach to excellence and trust' (2020) EDPS, para 19.

allowed AI systems as well as compliance with EU values, interests, and fundamental rights.[103]

6 Conclusion

The new digitalised era has shown the immense value that a costumer's data can bring to all types of services, ranging from private companies to governmental agencies. The use of big data gathered from private entities could improve the abilities of law enforcement authorities to predict and prevent more and more crimes from occurring. This allows police officers to adopt a preventive rather than responsive approach to crime by intervening before the offence has even occurred. Without doubt, this type of policing is more practical and less harmful for victims than subsequent evidence gathering and prosecutions after an incident. Nevertheless, as the software predicts potential offenses and offenders which have not yet occurred, law enforcement authorities must be particularly careful in assessing the results as there is always a possibility to incur wrongful predictions. Especially data transfers undertaken to strengthen and enhance predictive policing software must be carried out with great care as risks to fundamental rights like privacy and data protection persist. As no offence has actually been committed, it remains questionable which data may legitimately be gathered by governmental authorities. An interference by the state into fundamental rights must, after all, be proportionate and necessary, thus requiring authorities to make an abstract assessment based on predicted, yet hypothetical scenarios.

Although EU Member States are increasingly using predictive policing software, the trend is still at the initial stage. Little information is available regarding the exact working of the algorithms and even though discussions about the use of AI in law enforcement have been initiated, an applicable legal framework is still at loss. The absence of rules and regulations in this domain intensifies the lack of confidence in new software by the general public and increases the chances of misuse by the authorities. The adoption of measures regulating the use of new technologies by governmental authorities is thus crucial and should be adopted shortly. In order to ensure uniformity across European law enforcement authorities, it is important to adopt common rules within the EU. Service providers located throughout the Union may be compelled to share their data with the authorities of a requesting state yet cannot be expected to be aware of the applicable domestic legal framework regarding data transfers of 27 Member States. For this reason, a consistent EU legal basis regulating transfer, receipt and use of big data for predictive policing is essential.

[103] European Commission, 'Shaping Europe's Digital Future' (2020) European Union, 6 <https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/shaping-europe-digital-future_en> accessed 15 April 2021.

References

Amazon, 'Amazon.com Privacy Notice' (*Amazon*, 12 February 2021) <https://www.amazon.com/gp/help/customer/display.html?nodeId=GX7NJQ4ZB8MHFRNJ#GUID-8966E75F-9B92-4A2B-BFD5-967D57513A40__SECTION_3DF674DAB5B7439FB2A9B4465BC3 E0AC> accessed 5 June 2021

Apple, 'Apple Privacy Policy' (*Apple*, 1 June 2021) <https://www.apple.com/legal/privacy/en-ww/> accessed 5 June 2021

Babuta A, 'Big Data and Policing: An Assessment of Law Enforcement Requirements, Expectations and Priorities' (2017) Royal United Services Institute for Defence and Security Studies

Berreby D, 'Click to agree with what? No one reads terms of service, studies confirm' (*The Guardian*, 3 March 2017) <https://www.theguardian.com/technology/2017/mar/03/terms-of-service-online-contracts-fine-print> accessed 5 June 2021

Booking.com, 'Privacy Statement' (*Booking.com*, 30 September 2020) <https://www.booking.com/general.en-gb.html?aid=356980;label=gog235jc-1DCBQoggJCBXRlcm1zSAdYA2ipAYgBAZgBB7gBF8gBDNgBA-gBAfgBAogCAagCA7gCjLXthQbAAgHSAiQ5NWVjY2UxYS03NjhhLTQ1MTItOWFhYi0zYWY2ZDkwNjU2YzHYAgTgAgE;sid=0b1393ade68be8c2c916ca4e17ef5a0f;sig=v1NHFxMJat;tmpl=docs/privacy-policy#personal-data-3rd-parties-shared-how> accessed 5 June 2021

Boyd D and Crawford K, 'Critical Questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon' (2012) 15:5 Information, Communication & Society 662

Cameron I, 'European Union Law Restraints on Intelligence Activities' (2020) 33:3 International Journal of Intelligence and Counter Intelligence 452

Carrera S, Curtin D and Geddes A, *20 Years Anniversary of the Tampere Programme: Europeanisation Dynamics of the EU Area of Freedom, Security and Justice* (European University Institute 2020)

Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Region on A European strategy for data [2020] COM(2020) 66 final

Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Region on "Towards a common European data space" [2018] COM(2018) 232 final

Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Region on Artificial Intelligence for Europe [2018] COM(2018) 237 final

Cybercrime Convention Committee, 'Second Additional Protocol to the Convention on Cybercrime on enhanced cooperation and disclosure of electronic evidence – Draft Protocol version 3' (May 2021) Council of Europe

Degeling M and Berendt B, 'What is wrong about Robocops as consultants? A technology-centric critique of predictive policing' (2018) 33 AI & Soc 347

Dencik L, Hintz A and Carey Z, 'Prediction, pre-emption and limits to dissent: Social media and big data uses for policing protests in the United Kingdom' (2018) 20(4) New Media & Society 1433

Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications) [2002] OJ L201/37 (ePrivacy Directive)

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

Directive (EU) 2018/843 of the European Parliament and of the Council of 30 May 2018 amending Directive (EU) 2015/849 on the prevention of the use of the financial system for the purposes of money laundering or terrorist financing, and amending Directives 2009/138/EC and 2013/36/EU [2018] OJ L156/3

Edwards L and Urquhart L, 'Privacy in public spaces: what expectations of privacy do we have in social media intelligence?' (2016) 24 International Journal of Law and Information Technology 279

European Commission, 'FAQs on business-to-government data sharing' (*European Commission*, 23 March 2020) <https://ec.europa.eu/digital-single-market/en/faq/faqs-business-government-data-sharing> accessed 15 April 2021

—— 'Shaping Europe's Digital Future' (2020) European Union, 6 <https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/shaping-europe-digital-future_en> accessed 15 April 2021

—— 'Towards a European strategy on business-to-government data sharing for the public interest' (2020) European Union

—— 'White Paper on Artificial Intelligence – A European approach to excellence and trust' [2020] COM(2020) 65 final

European Data Protection Supervisor, 'Opinion 4/2020 EDPS Opinion on the European Commission's White Paper on Artificial Intelligence – A European approach to excellence and trust' (2020) EDPS

European Union Agency for Fundamental Rights and Council of Europe, *Handbook on European data protection law* (European Union Agency for Fundamental Rights and Council of Europe 2018)

Facebook, 'Data Policy' (*Facebook*) <https://www.facebook.com/policy.php#legal-requests-prevent-harm> accessed 5 June 2021

Ferguson A G, 'Policing Predictive Policing' (2017) 94 Washington University Law Review 1109

—— *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement* (New York University Press 2017)

Franke J-D, 'A year in surveillance' (*About: Intel*, 2020) <https://aboutintel.eu/a-year-in-surveillance/> accessed 26 March 2021

George J, 'How Social Media Is Changing Law Enforcement: Policing Big Data in the Connected World' (*LinkedIn*, 19 March 2019) <https://www.linkedin.com/pulse/how-social-media-changing-law-enforcement-policing-big-jaevon-george/> accessed 26 March 2021

González Fuster G, 'Artificial Intelligence and Law Enforcement: Impact on Fundamental Rights' (2020) European Parliament

Gstrein O, Bunnik A and Zwitter A, 'Review of ethical, legal & social issues impacting Predictive Policing' (2018) Cutting Crime Impact

Hardyns W and Rummens A, 'Predictive Policing as a New Tool for Law Enforcement? Recent Developments and Challenges' (2017) 24 European Journal on Criminal Policy and Research 201

Hurley R, *Big Data: A Guide to Big Data Trends, Artificial Intelligence, Machine Learning, Predictive Analytics, Internet of Things, Data Science, Data Analytics, Business Intelligence, and Data Mining* (Independently Published 2019)

Irion K and Luchetta G, 'Online Personal Data Processing and EU Data Protection Reform' (2013) CEPS

Jansen F, 'Data Driven Policing in the Context of Europe' (2018) Cardiff University <https://datajusticeproject.net/wp-content/uploads/sites/30/2019/05/Report-Data-Driven-Policing-EU.pdf> accessed 26 March 2021

Joined Cases C-203/15 and C-698/15 *Tele2 Sverige AB and Watson* [2016] ECLI:EU:C:2016:970

Jones M, 'Predictive Policing a Success in Santa Cruz, Calif - City sees significant reduction in property theft thanks in part to predictive crime modeling' (*Government Technology*, 8 October 2012) <https://www.govtech.com/public-safety/predictive-policing-a-success-in-santa-cruz-calif.html> accessed 9 October 2021

Lau T, 'Predictive Policing Explained: Attempts to forecast crime with algorithmic techniques could reinforce existing racial biases in the criminal justice system' (*Brennan Center for Justice*, 1 April 2020) <https://www.brennancenter.org/our-work/research-reports/predictive-policing-explained> accessed 1 June 2021

Liagre F, 'Predictive Policing Recommendations paper' (2016) European Crime Prevention Network

Lynskey O, 'Criminal justice profiling and EU data protection law: precarious protection from predictive policing' (2019) 15 International Journal of Law in Context 162

Miller K, 'Total Surveillance, Big Data, and Predictive Crime Technology: Privacy's Perfect Storm' (2014) 19 Journal of Technology Law & Policy 105

Nicas J, Isaac M and Frenkel S, 'Millions Flock to Telegram and Signal as Fears Grow Over Big Tech' (*The New York Times*, 13 January 2021) <https://www.nytimes.com/2021/01/13/technology/telegram-signal-apps-big-tech.html#:~:text=Telegram %20has%20been %20particularly%20popular,Facebook%20and%20Twitter%20limited% 20Mr> accessed 5 June 2021

Paul A, Cleverley M, Kerr W, Marzolini F, Reade M and Russo S, 'Smarter Cities Series: Understanding the IBM Approach to Public Safety' (2011) IBM

Pearsall B, 'Predictive Policing: The Future of Law Enforcement?' (2010) 10 National Institute of Justice Journal 266

Proposal for a Regulation of the European Parliament and of the Council on European Production and Preservation Orders for electronic evidence in criminal matters [2018] COM(2018) 225 final

Purtova N, 'Between the GDPR and the Police Directive: navigating through the maze of information sharing in public-private partnerships' (2018) 8:1 International Data Privacy Law 52

Rademacher T, 'Artificial Intelligence and Law Enforcement' in Wischmeyer T and Rademacher T, *Regulating Artificial intelligence* (Springer 2020)

Rasch M, 'Public-Private Partnerships: Sharing Data, Compromising Privacy' (*Security Boulevard*, 23 May 2018) <https://securityboulevard.com/2018/05/public-private-partnerships-sharing-data-compromising-privacy/> accessed 15 April 2021

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 L119/1

Stolton S, 'Vestager warns against predictive policing in Artificial Intelligence' (*EurActiv*, 30 June 2020) <https://www.euractiv.com/section/digital/news/vestager-warns-against-predictive-policing-in-artificial-intelligence/> accessed 15 April 2021

Tayebi M A and Glässer U, *Social Network Analysis in Predictive Policing* (Springer 2016)

Telegram, 'Telegram Privacy Policy' (*Telegram*) <https://telegram.org/privacy#8-who-your-personal-data-may-be-shared-with> accessed 5 June 2021

van Brakel R, 'Pre-emptive Big Data Surveillance and its (Dis)empowering Consequences: The Case of Predictive Policing' in Van der Sloot et al. (eds), *Exploring the Boundaries of Big Data* (Amsterdam University Press 2016)

Wilkins N, *Internet of Things: What You Need to Know About IoT, Big Data, Predictive Analytics, Artificial Intelligence, Machine Learning, Cybersecurity, Business Intelligence, Augmented Reality and Our Future* (Independently Published 2019)

Williams M, Burnap P and Sloan L, 'Crime sensing with Big Data: The Affordances and Limitations of Using Open-Source Communications to Estimate Crime Patterns' (2017) 57 British Journal of Criminology 320

Zoom, 'Zoom Privacy Statement' (*Zoom*, 4 June 2021) <https://zoom.us/privacy#_70wwucijkyy> accessed 5 June 2021

AUGMENTED REALITY IN LAW ENFORCEMENT FROM AN EU DATA PROTECTION LAW PERSPECTIVE: THE DARLENE PROJECT AS A CASE STUDY

By Katherine Quezada-Tavárez[*]

Abstract

Augmented reality (AR) is gaining popularity given its ability to blend the digital and physical worlds, amplifying human capabilities and improving the performance of tasks. In the law enforcement domain, a promising application is AR combined with artificial intelligence (AI) to improve situational awareness by enabling the optimal and timely delivery of crucial information during tactical decision-making. These technological advances create great opportunities for the fight against crime and terrorism, but they also raise a plethora of legal and ethical concerns. This paper aims to contribute to the growing literature on AI in criminal justice by examining AI-based AR solutions in law enforcement through the lens of EU data protection law. Focus is placed on three major issues, namely data minimisation, processing of special categories of data and automated decision-making. The analysis is concretised by looking at a specific project in this field, the EU-funded Deep AR Law Enforcement Ecosystem (DARLENE) project, to illustrate the possible implications.

1 Introduction

Augmented reality (AR) is a technology that overlays computer-generated information into physical environments, merging real and virtual objects in a complementary manner.[1] Combined with artificial intelligence (AI) techniques, AR capabilities can be greatly improved,[2] offering immense potential for the timely and accurate processing of real-time data streams, facilitating data-driven decisions, and enabling the quick and efficient execution of required tasks.[3] Such functionalities can prove useful in a wide range of domains, and criminal justice is not the exception.

[*] Researcher, KU Leuven Centre for IT & IP Law (CiTiP). This contribution is based on research conducted in the DARLENE project, which has received funding from the EU Horizon 2020 research and innovation programme under grant agreement No 883297. For correspondence: <katherine.quezada@kuleuven.be>.

[1] Ronald Azuma and others, 'Recent Advances in Augmented Reality' (2001) 21 IEEE Computer Graphics and Applications 34.

[2] Chandan K Sahu, Crystal Young and Rahul Rai, 'Artificial Intelligence (AI) in Augmented Reality (AR)-Assisted Manufacturing Applications: A Review' (2020) 59 International Journal of Production Research 4903.

[3] See Dragos Datcu and others, 'On the Usability of Augmented Reality for Information Exchange in Teams from the Security Domain', *2014 IEEE Joint Intelligence and Security Informatics Conference* (IEEE 2014); Stephan Lukosch and others, 'Providing Information on the Spot: Using Augmented Reality for Situational Awareness in the Security Domain' (2015) 24 Computer Supported Cooperative Work (CSCW) 613.

Initiatives are being explored to build AR tools that can help Law Enforcement Agencies (LEAs) rapidly and accurately process and exchange a wealth of information in complex and highly dynamic operations. DARLENE, a Horizon 2020 research and innovation action, is one of such projects developing AR technology underpinned by AI to assist LEAs in rapid scene analysis and real-time processing of information from different sources, thus facilitating more informed and efficient decision-making.[4]

Despite the obvious positive effects to augment human capabilities, cutting-edge AR systems bring along possible negative side-effects for fundamental rights and freedoms of citizens.[5] Given its 'pervasive information capture',[6] major concerns about AR technology relate to privacy[7] and data protection[8]. Even more so because with the developments of big data, particularly in an AI context, personal and non-personal data might indeed increasingly get mixed, making it ever more difficult to completely rule out the possibility of eventually processing personal data. Previous studies have focused on the legal implications of AR in a private and commercial setting, and mostly from a United States law perspective.[9] This contribution unravels the expected data protection issues that may arise when implementing AI-driven AR technology in law enforcement, using DARLENE as a case study.

This paper is structured as follows. Section 2 sets the scene by discussing the potential of AR to improve law enforcement practices, introducing the system considered in this study (DARLENE), and outlining the applicable data protection instrument in such a

[4] DARLENE, 'About DARLENE' (*darleneproject.eu*, 2020) <https://www.darleneproject.eu/about/> accessed 11 October 2021. DARLENE stands for: Deep AR Law Enforcement Ecosystem.

[5] In view of the legal and ethical questions raised by these novel technologies, the DARLENE project includes a work package devoted to the legal and ethical aspects of the research. See Katherine Quezada-Tavárez and Eva Houtave, 'Augmented Reality Technology to Counter Crime and Terrorism: Introduction to DARLENE' (*CiTiP Blog*, 16 April 2021) <https://www.law.kuleuven.be/citip/blog/augmented-reality-technology-to-counter-crime-and-terrorism-introduction-to-darlene/> accessed 16 April 2021.

[6] Mark A Lemley and Eugene Volokh, 'Law, Virtual Reality, and Augmented Reality' (2018) 166 University of Pennsylvania Law Review 1051, 1125.

[7] See Brian Wassom, *Augmented Reality Law, Privacy, and Ethics: Law, Society, and Emerging AR Technologies* (Syngress 2014) ch 3; Lemley and Volokh (n 6); Sally A Applin and Catherine Flick, 'Facebook's Project Aria Indicates Problems for Responsible Innovation When Broadly Deploying AR and Other Pervasive Technology in the Commons' (2021) 5 Journal of Responsible Technology 100010.

[8] See Sayoko Blodgett-Ford and Mirjam Supponen, 'Data Privacy Legal Issues in Virtual and Augmented Reality Advertising' in Woodrow Barfield and Marc Jonathan Blitz (eds), *Research handbook on the law of virtual and augmented reality* (Edward Elgar Publishing 2018); see also European Data Protection Supervisor, 'Technology Report No 1: Smart Glasses and Data Protection' (2019) <https://edps.europa.eu/sites/default/files/publication/19-01-18_edps-tech-report-1-smart_glasses_en.pdf> accessed 11 June 2021.

[9] See Wassom (n 7); Lemley and Volokh (n 6); Woodrow Barfield and Marc Jonathan Blitz, *Research Handbook on the Law of Virtual and Augmented Reality* (Edward Elgar Publishing 2018); however, a chapter in one of the cited works considers the data privacy and data protection regulations and case law of both the United States and Europe, including the General Data Protection Regulation, applicable to advertising in virtual and augmented reality. See Blodgett-Ford and Supponen (n 8).

scenario, namely the Law Enforcement Directive (LED).[10] Section 3 elaborates on data protection issues that may arise when using AI-powered AR technology in the field, focusing on key questions about data minimisation, processing of special categories of data and automated decision-making. It also suggests some ways to deal with those issues based on the insights gained in security research projects, including DARLENE. Section 4 concludes.

2 Cutting-Edge AR Technology in Law Enforcement

To identify legal and data protection issues, an analysis of the technology in question and its functionalities, as well as the applicable rules, is necessary. This section introduces AR technology and its potential benefits when integrated into the domain of policing, focusing on the DARLENE prototype through an overview of the system and its features. It then summarises the applicable data protection instrument to the foreseen uses of the technology in question, namely the LED.

2.1 Current applications

AR is mostly known for its uses in the gaming industry, with Pokémon GO being one of the most popular AR applications thus far,[11] as well as in the advertising and commercial domain. It is also used in other sectors, such as entertainment,[12] education,[13] manufacturing[14] and tourism[15]. In policing, AR has proven useful for training activities and for the fulfilment of certain tasks. AR (and immersive technology more broadly) is used in field staff training through serious games for police officers and first responders that allow for a realistic experience, which can enhance, or in some cases even replace, traditional training approaches.[16] AR applications are also used in an operational context, for

[10] 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 (LED).

[11] Nicola Liberati, 'Phenomenology, Pokémon Go, and Other Augmented Reality Games' (2018) 41 Human Studies 211, 216.

[12] Mariza Dima, Linda Hurcombe and Mark Wright, 'Touching the Past: Haptic Augmented Reality for Museum Artefacts' in Randall Shumaker and Stephanie Lackey (eds), *Virtual, Augmented and Mixed Reality (VAMR) 2014. Applications of Virtual and Augmented Reality* (Springer International Publishing 2014).

[13] George Margetis and others, 'Augmented Interaction with Physical Books in an Ambient Intelligence Learning Environment' (2013) 67 Multimedia Tools and Applications 473.

[14] Sahu, Young and Rai (n 2).

[15] Ryan Yung and Catheryn Khoo-Lattimore, 'New Realities: A Systematic Literature Review on Virtual Reality and Augmented Reality in Tourism Research' (2019) 22 Current Issues in Tourism 2056.

[16] Babak Akhgar (ed), *Serious Games for Enhancing Law Enforcement Agencies: From Virtual Reality to Augmented Reality* (Cham: Springer International Publishing AG 2019).

instance, to support the remote collaboration of investigative teams, by allowing eg investigators on the scene to consult expert colleagues at a distance,[17] or as a tool to annotate crime scenes in forensic investigation.[18]

More recent studies have shown AR's potential as a tool to facilitate the rapid and adequate exchange of information between operational teams, thus fostering situational awareness of LEAs.[19] There are two concrete examples of this in the EU. One is the location-based hotspot policing system under development and testing in the Netherlands.[20] The other is the AR system developed in the EU-funded DARLENE project, used as a case study in this contribution.

2.2 DARLENE system overview

The DARLENE project investigates how AR, combined with AI and other cutting-edge technology, can be deployed in real-time to support the rapid visual understanding of complex real-world scenes in tactical decision-making. The solution developed in the project combines AR capabilities with powerful Machine Learning (ML) algorithms, sensor information fusion techniques, 3D reconstruction and wearable technology (including smart glasses and smart bands) that can help improve end-user situational awareness. It also integrates 5G technology to allow agile processing of large volumes of real-time data.

More specifically, the system components can be grouped into four clusters, namely: i) a wearable edge computing node, comprising the binocular AR glasses and the smart band that will be worn by officers, as well as the micro-processing module and a WI-FI module transceiver, which will provide users with information and services relevant for their contexts and useful for their assigned tasks; ii) a patrol car edge computing node, comprising multiple microprocessors, similar to the ones 'worn' by the officers; iii) the sensors network, consisting of a fusion of both the smart devices and existing sensors in the facilities where the technology will be deployed (such as closed-circuit television (CCTV) cameras, drones, beacons, etc.), for personalised context-aware recommendations and to create a miniature 3D model of the area; and iv) a cloud, comprising servers that will enable the continuous training of intensive ML algorithms to perform the classification of data captured by the sensors or transmitted to the sensors.

Ultimately, the goal is to offer European law enforcement practitioners a proactive security solution that can help them make more informed and rapid decisions by improving their situational awareness. In that way, the solution will contribute to crime detection

[17] Ronald Poelman and others, 'As If Being There: Mediated Reality for Crime Scene Investigation', *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* (2012).
[18] Jan Willem Streefkerk and others, 'The Art of Csi: An Augmented Reality Tool (Art) to Annotate Crime Scenes in Forensic Investigation', *International conference on virtual, augmented and mixed reality* (Springer 2013).
[19] Datcu and others (n 3); Lukosch and others (n 3).
[20] Hendrik Engelbrecht and Stephan Lukosch, 'Dangerous or Desirable: Utilizing Augmented Content for Field Policing' (2020) 36 International Journal of Human–Computer Interaction 1415.

and prevention by enabling end-users the timely reception of crucial information and analysis of massive volumes of data, thus allowing them to anticipate criminal activities and more quickly respond to incidents in progress.

The functionalities will be tested using two use cases. The first is the rapid visual scene analysis for anomaly detection. This scenario focuses on the security monitoring in the crowded areas of an airport, where the system will contribute to the situational awareness of police officers through real-time, quick and accurate information of the situation. It will also facilitate information exchanges and coordination between ground personnel and command and control staff. The second use case is the tactical neutralisation of human adversaries in the presence of friendlies. This scenario concerns incidents involving armed perpetrators to be apprehended while acting in indoor crowded spaces (eg a transit station, hotel, shopping centre, or bank). Hence, it is intended for coordinated response and planning of tactical approach of LEAs addressing an ongoing incident. By using an array of sensors pre-installed in the facilities (eg CCTV), the system will reconstruct and create a miniature 3D model of the area, facilitating the visualisation of the locations of persons in the scene, their movements and orientations in real-time.

2.3 EU data protection law in law enforcement

The use of DARLENE will fall under the scope of the LED, which is the instrument equivalent to the GDPR[21] for the police and criminal justice sector.[22] The LED regulates the processing of personal data by competent authorities for law enforcement purposes, which will most likely be the case of DARLENE given the foreseen use by LEAs in field operations. While not covering the processing of personal data during the research project, the LED does need to be considered in the analysis of data protection issues regarding the solution under development. In particular, the efforts done by system users to comply with their ethical and legal responsibilities when employing the technology can be facilitated and enabled through adequate design choices in the system architecture.

Amongst the requirements mandated by the LED, one that appears particularly relevant for the development of novel security solutions is that of data protection by design and by default.[23] If privacy and data protection are built into the system design, it can be reasonably assumed that end-users would more easily comply with their data protection by design and by default obligations when using the technology. In fact, privacy and data protection should be best tackled from the earliest stage of the system development as it may become harder to prevent or overcome any such issues at a later phase.[24] More

[21] 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 L119/1.
[22] Paul De Hert and Vagelis Papakonstantinou, 'The New Police and Criminal Justice Data Protection Directive: A First Analysis' (2016) 7 New Journal of European Criminal Law.
[23] LED, Article 20.
[24] Laurens Sion and others, 'DPMF: A Modeling Framework for Data Protection by Design' (2020) 15 Enterprise Modelling and Information Systems Architectures (EMISAJ) 1, 1.

than a good practice, implementing privacy and data protection measures in the technology design is also encouraged by data protection provisions that apply at the research and development stage.[25]

3 Key Issues from an EU Data Protection Law Perspective

Automated systems, such as AR technology and AI, assist LEAs in decisions that can have an impact on the life of individuals. As cutting-edge technologies gain popularity in the law enforcement sector given their apparent positive effects, so too are there growing ethical and legal concerns. The multitude of legal and ethical risks associated with novel security solutions, such as DARLENE, include privacy, non-discrimination, due process, fair trial and other fundamental rights and ethical values.[26] Data-intensive technologies, such as AR and AI, typically require the processing of significant amounts of (personal) data, giving rise to many data protection questions as well. Considering the system overview and the foreseen uses, this section addresses major data protection questions associated with AR technology in law enforcement, focusing on data minimisation, processing of special categories of data and automated decision-making. The analysis mostly relates to the processing of data regarding the persons in the field of view of the AR glasses, with limited references to the processing of police officer data through the smart bands.

3.1 Data minimisation

A key concept in data protection, both from a data subject rights and an information security perspective, is data minimisation. The LED stipulates that personal data may only be processed if adequate, relevant, and not excessive in relation to the purposes for which they are collected.[27] The data minimisation principle requires those in charge of processing activities to determine the smallest amount of data required to accomplish

[25] The research and development of law enforcement technology is regulated in the GDPR. According to Recital 78 of the GDPR, '[w]hen developing, designing, selecting and using applications, services and products that are based on the processing of personal data or process personal data to fulfil their task, producers of the products, services and applications should be encouraged to take into account the right to data protection when developing and designing such products, services and applications and, with due regard to the state of the art, to make sure that controllers and processors are able to fulfil their data protection obligations.'

[26] See, amongst others, Andrew Guthrie Ferguson, *The Rise of Big Data Policing* (NYU Press 2017); Alexander Babuta, Marion Oswald and Christine Rinik, 'Machine Learning Algorithms and Police Decision-Making: Legal, Ethical and Regulatory Challenges' (Royal United Services Institute for Defence and Security Studies 2018) 3 <https://rusi.org/publication/whitehall-reports/machine-learning-algorithms-and-police-decision-making-legal-ethical> accessed 5 May 2021; Gloria González Fuster, 'Artificial Intelligence and Law Enforcement - Impact on Fundamental Rights' (European Parliament's Policy Department for Citizens' Rights and Constitutional Affairs at the request of the Committee on Civil Liberties, Justice and Home Affairs 2020) <https://www.europarl.europa.eu/RegData/etudes/STUD/2020/656295/IPOL_STU(2020)656295_EN.pdf> accessed 27 April 2021; Angelika Adensamer and Lukas Daniel Klausner, '"Part Man, Part Machine, All Cop": Automation in Policing' (2021) 4 Frontiers in Artificial Intelligence 29.

[27] LED, Article 4(1)(c).

their goals. As a result, LEAs should only keep that much data, and are not allowed to gather additional data merely because it could be beneficial in the future or because no consideration has been given to whether it is required in a specific circumstance. For instance, it might be excessive to build a profile on every unique person detected through AR glasses used by LEAs in public spaces for the purpose of apprehending some criminals.

Despite the intuitive clash between the data minimisation principle and AI-driven AR solutions, there are ways to implement and enforce this data protection requirement through technical design measures. A way forward is to embed privacy settings within the entire lifecycle of the novel technology, ensuring that, by default, the system avoids the collection and/or the further processing of unnecessary personal data.[28] That means that the DARLENE devices should be designed in way that helps meet that goal, for instance by engineering the AR glasses and the smart bands in a manner that inhibits data collection wherever possible. More concrete technical design strategies to alleviate ethical and legal concerns include the following.

First, it should be ensured that the data collection and processing capabilities of the system are limited to only data required for the proper performance of the technology. In DARLENE, the system design and functionalities are carefully assessed and worked out to ensure that the processing capabilities are limited to what is needed for the stated purposes. For instance, during regular multidisciplinary conversations,[29] which started before the formal start of the project, it was decided that the processing of data through the technology will only take place live and on-site, meaning that no (personal) data captured through the AR devices will be stored.

Second, for systems involving storage of data, it should be ensured that any recording through the wearable devices is only triggered by specific situations. More specifically, the AR device should be programmed to start any recordings preferably upon the human operator's initiative (such as pressing a button, for example). Yet, considering the foreseen uses of DARLENE in high-pressure situations, automated recordings may be more suitable in certain circumstances. Particularly, the human operator might not always be able to undertake active action on the wearable devices due to other activities taking priority (eg using another device to call for backup). In that scenario, automated recordings should be initiated only in strictly defined circumstances, such as when a police officer's physiological state, which can be derived from the measurements processed through the smart band, reaches a certain level of excitement indicating stress or fear, obviating the need for the wearer to manually start the recording.

[28] As suggested by the Article 29 Working Party (currently, European Data Protection Board) in its opinion on the use of drones, which is also a novel technology used for surveillance purposes. See Article 29 Working Party, 'Opinion 01/2015 on Privacy and Data Protection Issues Relating to the Utilisation of Drones' (2015) WP231 13–15 <https://ec.europa.eu/newsroom/article29/items/640602/en> accessed 11 June 2021.

[29] See Quezada-Tavárez and Houtave (n 5).

Third, for systems involving data recording, local storage of the data should be preferred, and the information should be removed as soon as reasonably practicable. The data retention period of any recordings should be strictly defined, in line with data protection rules (particularly the principles of data minimisation and storage limitation), as well as the specific provisions for storage periods with regards to surveillance material under national law.[30] It is for LEAs, as data controllers, to define the retention period in line with the principles of necessity and proportionality, and to demonstrate compliance with the applicable data protection rules. Still, technical design measures in the system can help in the enforcement of those periods through the automated erasure of the data after the lapse of this period, for example. In any case, the long-term storage of collected data on a device is highly unadvisable as it may generate unnecessary privacy risks (eg data loss or theft on a later use of the system), and might also entail a breach of data retention rules.

Fourth, the system should only enable the processing of a limited set of attributes that are strictly needed for the foreseen uses. Regarding the DARLENE solution, for instance, no face or audio data collection or processing features will be incorporated in the technology. Such technical design measure helps mitigate the risks related to the devices' potential to capture large amounts of unintended sensitive information.[31] In this context, it should be noted that, despite the potential of facial recognition for the real-time detection and prevention of crime,[32] the system will not embed facial recognition features due to the ethical and legal risks concerning the implementation of that technology in law enforcement. Facial recognition has generated much debate and controversy about privacy and other fundamental rights,[33] and its use in law enforcement was ruled unlawful

[30] European Data Protection Board, 'Guidelines 3/2019 on Processing of Personal Data through Video Devices - Version 2.0' (2020) 13 and 28 <https://edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_201903_video_devices_en_0.pdf> accessed 28 June 2021.

[31] See Article 29 Working Party, 'Opinion 8/2014 on Recent Developments on the Internet of Things' (2014) WP223 17 <https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp223_en.pdf> accessed 11 June 2021.

[32] Cameron Martin, 'Facial Recognition in Law Enforcement' (2020) 19 Seattle Journal for Social Justice 309, 329.

[33] See eg Philip Brey, 'Ethical Aspects of Facial Recognition Systems in Public Places. Journal of Information, Communication, & Ethics in Society' (2004) 2 Journal of Information, Communication, & Ethics in Society 97; Sharon Naker and Dov Greenbaum, 'Now You See Me: Now You Still Do: Facial Recognition Technology and the Growing Lack of Privacy' (2017) 23 BUJ Sci. & Tech. L. 88; Katherine Quezada-Tavárez, 'Law Enforcement AI in the Spotlight as EDPB Cast Doubt on Legality of Facial Recognition Tech' (*KU Leuven CiTiP Blog*, 20 July 2020) <https://www.law.kuleuven.be/citip/blog/law-enforcement-ai-in-the-spotlight-as-edpb-cast-doubt-on-legality-of-facial-recognition-tech/> accessed 17 June 2021.

by a European court in 2020.[34] Also, it is worth mentioning that, in the ongoing discussions prompted by the 2021 proposal for an EU AI regulation,[35] facial recognition is one of the most hotly debated aspects.[36]

3.2 Processing of special categories of personal data

Once fully operational, the DARLENE system will enable the processing of behavioural information about the citizens in the field of view of the AR glasses, which might entail the processing of special categories of personal data, such as biometric data. Special categories of personal data are subject to a specific regime established in Article 10 of the LED. That provision allows the processing of special categories of data, including biometric data for unique identification purposes, under three cumulative conditions: i) where strictly necessary; ii) subject to appropriate safeguards for the rights and freedoms of the data subject; and iii) only a) where authorised by Union or Member State law; b) to protect the vital interests of the data subject or of another natural person; or c) where such processing is related to data which are manifestly made public by the data subject. In that the way, the LED reverses the logic applied under the GDPR, where the processing of biometric data for uniquely identifying a natural person is, as a general rule, prohibited.[37]

The LED defines biometric data as 'personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of a natural person, which allow or confirm the unique identification of that natural person, such as facial images or dactyloscopic data'.[38] The processing of biometric information does not automatically involve biometric data processing as defined under data protection law. Three criteria must be considered in the determination of biometric data processing, namely: i) the nature of data (ie data relating to physical, physiological or behavioural characteristics of a natural person); ii) the means and way of processing (ie data 'resulting from a specific technical processing'; and iii) the purpose of processing (ie data must be used for the purpose of uniquely identifying a natural person.

DARLENE will involve the processing of data relating to the physical or behavioural characteristics of individuals in the field of view of the smart glasses. Whereas those data may allow the identification of a natural person, they only lead to biometric data processing when certain conditions are met, including the purpose of uniquely identifying

[34] R (Bridges) v Chief Constable of South Wales Police [2020] EWCA CIV 1058.

[35] European Commission, Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021) 206 final).

[36] See eg Natasha Lomas, 'Ban Biometric Surveillance in Public to Safeguard Rights, Urge EU Bodies' *TechCrunch* (21 June 2021) <https://techcrunch.com/2021/06/21/ban-biometric-surveillance-in-public-to-safeguard-rights-urge-eu-bodies/> accessed 22 June 2021.

[37] Els Kindt, 'Having Yes, Using No? About the New Legal Regime for Biometric Data' (2018) 34 Computer Law & Security Review 523, 528.

[38] LED, Article 3(13).

natural persons.[39] The processing operations in the system are intended to detect specific pre-defined activities (eg signs of dangerousness or suspicious activity), as opposed to the unique identification of persons. Therefore, it can be concluded that the processing operations to be carried out through DARLENE are not likely to qualify as biometric data processing as defined under data protection law. This is mainly because such processing does not meet all the three criteria to be considered as such since the physical and behavioural information will not be processed in a biometric sense. A different conclusion could contribute to the perpetuation of the terminological confusion between the legal definition and the scientific definition of the notions of 'biometrics' and 'biometric data'.[40]

Regardless, the physical and behavioural characteristics processed through the system are subject to safeguards to the rights and freedoms of the individuals, in line with the second criterion in Article 10 LED, which is the one of the three conditions that can be addressed through technology design choices. The live and on-site processing of data are technological safeguards in the system ensuring that no physical or behavioural data captured through the AR devices will be stored.[41] Also, despite the wireless communications involved, the processing of data on the edge (ie through wearable computational units carried by police officers) will help limit data transmissions. These measures should contribute to the mitigation of privacy and data protection risks, as well as help avoid cybersecurity threats such as eavesdropping attacks.[42]

3.3 Automated decision-making

Finally, given the uses of AR technology for the purpose of making law enforcement decisions, particular attention should be paid to the issues raised by automated decision-making. According to Article 11 LED, LEAs must not take a significant decision based solely on automated processing, including profiling, which significantly affects or produces adverse legal effect on an individual, unless that decision is required or authorised by law. That same provision grants individuals the right to obtain human intervention as a safeguard against the risk that a potentially damaging decision is taken by solely automated means.

Automated decision-making only comes into play where a *significant* decision producing adverse legal effects concerning the individual is taken based solely upon automated

[39] Catherine Jasserand, 'Avoiding Terminological Confusion between the Notions of "Biometrics" and "Biometric Data": An Investigation into the Meanings of the Terms from a European Data Protection and a Scientific Perspective' (2016) 6 International Data Privacy Law 63; Kindt (n 37) s 1.2; Erik Zouave and Jessica Schroers, 'You've Been Measured, You've Been Weighed and You've Been Found Suspicious: Biometrics and Data Protection in Criminal Justice Processing' in Ronald Leenes and others, *Data Protection and Privacy: The Internet of Bodies* (Hart Publishing 2019) 6.

[40] Jasserand (n 39).

[41] See section 3.1.

[42] See Article 29 Working Party, 'Opinion 8/2014 on Recent Developments on the Internet of Things' (n 31) 18.

processing, meaning 'the ability to make decisions by technological means without human involvement'.[43] Considering the foreseen uses of DARLENE, it can be said that the system is likely to involve decisions significantly affecting or producing adverse legal effects on individuals. The processing of data through the smart glasses may contribute to the apprehension of persons in the field of view of the AR devices, and eventually lead to a criminal prosecution. Such decisions are likely to involve adverse legal effects, or at a minimum can be expected to have a significant impact on the life of the person(s) in question. Yet, such decisions are unlikely to qualify as *solely* automated since the system will not operate in a manner that it can act by itself without human intervention. The smart glasses will provide human operators with augmented content to help them make more informed decisions when taking relevant action. Hence, it can be concluded that DARLENE does not involve automated decision-making within the meaning of the LED.

3.3.1 Profiling in light of the use cases

It is important to also determine whether the system will involve profiling, which is another aspect regulated by Article 11 LED. Article 3(4) of the LED defines 'profiling' as 'any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements'. Considering the LED's definition, it can be said that profiling consists of three specific elements, namely: i) an automated form of processing; ii) carried out on personal data; iii) for the purpose of evaluating personal aspects about a natural person to predict their behaviour and take decisions regarding him or her.[44]

At this point, it should be clarified that, although the definition of profiling refers to a form of 'automated processing', this should not be confused with the notion of 'automated decision'. While automated decisions producing legal effects concerning data subjects could possibly be based on profiling, it is not always the case. This is emphasised in the opinion of the Article 29 Working Party (currently European Data Protection Board) on profiling, which states that simply assessing or (in some cases) classifying individuals based on known characteristics could be considered profiling, with or without predictive purpose.[45]

---

[43] Article 29 Working Party, 'Opinion on Some Key Issues of the Law Enforcement Directive (EU 2016/680)' (2017) WP 258 11.

[44] Article 29 Working Party, 'Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679' (2018) WP251rev.01 6–7 <https://ec.europa.eu/newsroom/article29/document.cfm?doc_id=49826> accessed 29 June 2021; while the referred opinion relates to the GDPR, most aspects of that guidance also apply in a law enforcement context, see Article 29 Working Party, 'Opinion on Some Key Issues of the Law Enforcement Directive (EU 2016/680)' (n 43) 11.

[45] Article 29 Working Party, 'Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679' (n 44) 7.

Considering the definition and analysis of the relevant provisions, and when applying the three conditions triggering the qualification of profiling, it can be said that DARLENE is likely to involve profiling within the meaning of the LED. First, DARLENE can be considered an automated system, which will thus enable the automated processing of data. Second, the processing operations will be carried out on personal data, such as a person's precise positioning and orientation in a crowded space (first use case) and tracking the movements of persons in the scene (second use case). Third, the purpose of the processing will consist of evaluating behaviour to determine whether a potentially unlawful activity has unfolded. For instance, DARLENE will assess the movements of the individuals within crowded areas of an airport to highlight their silhouettes in the scene with specific labels (eg 'injured', 'armed suspect'), thus rendering them more salient in a crowded area (first use case). Another example is the situation where incidents involving armed individuals take place within internal, crowded spaces (eg transit station, hotel, shopping centre, bank). In that context, the system will facilitate viewing individual movements and orientations in real-time, even through solid walls, with the possibility for additional filters to highlight each 'skeleton' in a different colour, indicating whether occupants inside a building are friends or foes (second use case).

In light of the foregoing, it can be said that the LED does not prohibit profiling itself (ie automated processing of personal data for the purpose of making a decision) but rather grants individuals with a right to avoid being subject to a 'decision based solely on automated processing, including profiling'.[46] The LED clarifies that the decision itself may, nonetheless, be made provided it is authorised by Union or Member State law that the controller (ie LEAs using the system) is subject to and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests. Determining whether any such laws exist would require an exhaustive analysis of relevant EU law, as well as every applicable national legislation, which goes beyond the scope of this contribution, and thus not further explored in the analysis.

3.3.2 Human oversight to mitigate risks

Human oversight (also known as 'human-in-the-loop') plays a major role in automated decision-making.[47] This is particularly the case in the context of AI-driven systems given the limitations of the technology.[48] Even the most statistically accurate ML system may occasionally reach the wrong decision in an individual case, which highlights the need to ensure human review of the system's outcomes. Many circumstances may require a human to overturn an AI-derived decision, for instance the fact that an individual is an 'outlier', as well as the possibility to challenge the assumptions in the design of the automated system.

[46] LED, Article 11.

[47] See High-Level Expert Group on Artificial Intelligence, 'Ethics Guidelines for Trustworthy AI' (2019) 16 <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai> accessed 9 February 2021.

[48] See some examples in the introductory remarks of the present section 3.

This requirement has been linked to the more general and complex challenge of explaining AI systems and their outcomes. For instance, according to the High-Level Expert Group on AI, one of the principles that should inspire the development of AI is explainability (or explicability), along with respect for human autonomy, prevention of harm, and fairness.[49] In particular, explainability can contribute to the meaningful human oversight of automated systems, such as AI-driven technology.

Explainability involves two aspects, namely interpretability, which entails the communication of ML functions to the user in a way that the person can understand it, and completeness, meaning the need to explain the system operations accurately and in an auditable manner.[50] Other discussions on explainability revolve around two main types of explanations, namely process-based explanations and outcome-based explanations.[51] While the explanatory techniques and models developed by computer science experts seem to be more suited for individuals with a technical background, social scientists have focused on the goal of making explanations accessible to lay people. In particular, some of the proposed approaches include contrastive, selective, causal, and social explanation.[52]

These approaches appear to be more relevant in an ex-post explanation context, which could hardly be influenced from the technology development. An ex-ante approach might be more adequate to address at the research and development stage, where specific steps can be taken in support of human oversight. The approach used in security research projects involving automated systems (including DARLENE) consists of tackling the completeness and interpretability aspects of explainability during the technology design process and through end-user training. These measures can be considered of a technical and organisational nature respectively.

During the technical design process, completeness may be tackled by creating comprehensive technical documentation, including a general description of the system and its components, information about its capabilities and limitations, risk management measures implemented during the system's lifecycle, amongst others.[53] In the context of DARLENE, for instance, reports produced during the lifetime of the project (technical and otherwise) contain detailed information about the technical and other aspects of the solution, including the items listed in this paragraph. This should help ensure the completeness aspect of explainability.

---

[49] See High-Level Expert Group on Artificial Intelligence (n 47) 12–13.

[50] Leilani H Gilpin and others, 'Explaining Explanations: An Overview of Interpretability of Machine Learning', *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)* (IEEE 2018).

[51] See Riccardo Guidotti and others, 'A Survey of Methods for Explaining Black Box Models' (2019) 51 ACM Computing Surveys.

[52] See Brent Mittelstadt, Chris Russell and Sandra Wachter, 'Explaining Explanations in AI', *In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)* (2019).

[53] This is very much in line with a condition that forthcoming EU legislation may soon require from high-risk AI systems soon (see proposal for EU AI Act, Article 11).

However, the technical documentation might not be understandable to non-specialised audiences, such as system users of the AI-driven security solution. Hence, to tackle the interpretability aspect of explainability, before implementing the technology in their operations, end-users should at a minimum be provided with information on the system's general functioning, limitations and any other aspects that may somehow have an impact on the automated decision-making process, such as: i) the input data that the system takes into consideration (eg to label an individual as a potential 'threat', the position of the individual and carrying objects classified as weapons); ii) the target values that the system is meant to compute (eg the type of object that the individual is likely to be carrying); iii) the envisaged consequence of the automated assessment or system recommendation (eg prompting the officer to apprehend an individual).[54] This can be achieved by complementing the technical documentation with specific training materials and activities for system users. In DARLENE, for example, training materials are being produced, providing detailed information on the system, its functionalities, limitations, and other relevant aspects suggested earlier in this paragraph. The training package includes guidelines and recommendations for users about key legal and ethical aspects relevant to the automated system and its implementation in law enforcement practices. The aim is to raise awareness amongst system users on key legal and ethical issues, including the importance of human oversight in automated decision-making and ways to address risks such as automation bias.[55]

4 Concluding Remarks

Novel security solutions can be particularly useful in the fulfilment of law enforcement tasks. For instance, AI-powered AR technology can improve the situational awareness of agents in the field, thus contributing to a more rapid and effective detection and prevention of potentially unlawful behaviour. However, just like any disruptive technology, security solutions bring with them a set of considerations of a legal, ethical and of other nature. And those willing to contribute to the present and future of law enforcement practices would be wise to pause and ask the hard questions to address them on time.

This contribution provides an analysis of AR, combined with other cutting-edge technology, in law enforcement from a data protection perspective. Key concerns likely to arise in the foreseen uses consist of data minimisation, special categories of personal data and automated decision-making. Some of the measures considered in this study to overcome the associated challenges include the live and on-site processing of data, the careful selection of the system's features and data to be processed, as well as taking specific steps

[54] This is very much in line with a condition that forthcoming EU legislation may soon require from high-risk AI systems (see proposal for EU AI Act, Article 13).

[55] 'Automation bias' is a phenomenon consisting of the human tendency to over-rely on information provided by machines primarily due to the perception that they are generally trustworthy and free from (human) flaws. See Danielle Keats Citron, 'Technological Due Process' (2008) 85 Washington University Law Review 1249, 1271–1277; see also David G Robinson, 'The Challenges of Prediction: Lessons from Criminal Justice' (2018) 14 I/S: A Journal of Law and Policy for the Information Society 151, 164–165.

towards the meaningful human oversight of AR-assisted decisions, such as adequate training for users. Further research is needed to elaborate on other data protection issues, as well as to explore other fundamental rights concerns, such as privacy, non-discrimination, due process and fair trial, and ways to address them.

References

Adensamer A and Klausner LD, '"Part Man, Part Machine, All Cop": Automation in Policing' (2021) 4 Frontiers in Artificial Intelligence 29

Akhgar B (ed), *Serious Games for Enhancing Law Enforcement Agencies: From Virtual Reality to Augmented Reality* (Cham: Springer International Publishing AG 2019)

Applin SA and Flick C, 'Facebook's Project Aria Indicates Problems for Responsible Innovation When Broadly Deploying AR and Other Pervasive Technology in the Commons' (2021) 5 Journal of Responsible Technology 100010

Article 29 Working Party, 'Opinion 8/2014 on Recent Developments on the Internet of Things' (2014) WP223 <https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp223_en.pdf> accessed 11 June 2021

—— 'Opinion 01/2015 on Privacy and Data Protection Issues Relating to the Utilisation of Drones' (2015) WP231 <https://ec.europa.eu/newsroom/article29/items/640602/en> accessed 11 June 2021

—— 'Opinion on Some Key Issues of the Law Enforcement Directive (EU 2016/680)' (2017) WP 258

—— 'Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679' (2018) WP251rev.01 <https://ec.europa.eu/newsroom/ article29/document.cfm?doc_id=49826> accessed 29 June 2021

Azuma R and others, 'Recent Advances in Augmented Reality' (2001) 21 IEEE Computer Graphics and Applications 34

Babuta A, Oswald M and Rinik C, 'Machine Learning Algorithms and Police Decision-Making: Legal, Ethical and Regulatory Challenges' (Royal United Services Institute for Defence and Security Studies 2018) 3 <https://rusi.org/publication/whitehall-reports/machine-learning-algorithms-and-police-decision-making-legal-ethical> accessed 5 May 2021

Barfield W and Blitz MJ, *Research Handbook on the Law of Virtual and Augmented Reality* (Edward Elgar Publishing 2018)

Blodgett-Ford S and Supponen M, 'Data Privacy Legal Issues in Virtual and Augmented Reality Advertising' in Woodrow Barfield and Marc Jonathan Blitz (eds), *Research handbook on the law of virtual and augmented reality* (Edward Elgar Publishing 2018)

Brey P, 'Ethical Aspects of Facial Recognition Systems in Public Places. Journal of Information, Communication, & Ethics in Society' (2004) 2 Journal of Information, Communication, & Ethics in Society 97

Citron DK, 'Technological Due Process' (2008) 85 Washington University Law Review 1249

DARLENE, 'About DARLENE' (*darleneproject.eu*, 2020) <https://www.darleneproject.eu/about/> accessed 11 October 2021

Datcu D and others, 'On the Usability of Augmented Reality for Information Exchange in Teams from the Security Domain', *2014 IEEE Joint Intelligence and Security Informatics Conference* (IEEE 2014)

De Hert P and Papakonstantinou V, 'The New Police and Criminal Justice Data Protection Directive: A First Analysis' (2016) 7 New Journal of European Criminal Law

Dima M, Hurcombe L and Wright M, 'Touching the Past: Haptic Augmented Reality for Museum Artefacts' in Randall Shumaker and Stephanie Lackey (eds), *Virtual, Augmented and Mixed Reality (VAMR) 2014. Applications of Virtual and Augmented Reality* (Springer International Publishing 2014)

Engelbrecht H and Lukosch S, 'Dangerous or Desirable: Utilizing Augmented Content for Field Policing' (2020) 36 International Journal of Human–Computer Interaction 1415

European Data Protection Board, 'Guidelines 3/2019 on Processing of Personal Data through Video Devices - Version 2.0' (2020) <https://edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_201903_video_devices_en_0.pdf> accessed 28 June 2021

European Data Protection Supervisor, 'Technology Report No 1: Smart Glasses and Data Protection' (2019) <https://edps.europa.eu/sites/default/files/publication/19-01-18_edps-tech-report-1-smart_glasses_en.pdf> accessed 11 June 2021

Ferguson AG, *The Rise of Big Data Policing* (NYU Press 2017)

Gilpin LH and others, 'Explaining Explanations: An Overview of Interpretability of Machine Learning', *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)* (IEEE 2018)

González Fuster G, 'Artificial Intelligence and Law Enforcement - Impact on Fundamental Rights' (European Parliament's Policy Department for Citizens' Rights and Constitutional Affairs at the request of the Committee on Civil Liberties, Justice and Home Affairs 2020) <https://www.europarl.europa.eu/RegData/etudes/STUD/2020/656295/IPOL_STU(2020)656295EN.pdf> accessed 27 April 2021

Guidotti R and others, 'A Survey of Methods for Explaining Black Box Models' (2019) 51 ACM Computing Surveys

High-Level Expert Group on Artificial Intelligence, 'Ethics Guidelines for Trustworthy AI' (2019) <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai> accessed 9 February 2021

Jasserand C, 'Avoiding Terminological Confusion between the Notions of "Biometrics" and "Biometric Data": An Investigation into the Meanings of the Terms from a European Data Protection and a Scientific Perspective' (2016) 6 International Data Privacy Law 63

Kindt E, 'Having Yes, Using No? About the New Legal Regime for Biometric Data' (2018) 34 Computer Law & Security Review 523

Lemley MA and Volokh E, 'Law, Virtual Reality, and Augmented Reality' (2018) 166 University of Pennsylvania Law Review 1051

Liberati N, 'Phenomenology, Pokémon Go, and Other Augmented Reality Games' (2018) 41 Human Studies 211

Lomas N, 'Ban Biometric Surveillance in Public to Safeguard Rights, Urge EU Bodies' *TechCrunch* (21 June 2021) <https://techcrunch.com/2021/06/21/ban-biometric-surveillance-in-public-to-safeguard-rights-urge-eu-bodies/> accessed 22 June 2021

Lukosch S and others, 'Providing Information on the Spot: Using Augmented Reality for Situational Awareness in the Security Domain' (2015) 24 Computer Supported Cooperative Work (CSCW) 613

Margetis G and others, 'Augmented Interaction with Physical Books in an Ambient Intelligence Learning Environment' (2013) 67 Multimedia Tools and Applications 473

Martin C, 'Facial Recognition in Law Enforcement' (2020) 19 Seattle Journal for Social Justice 309

Mittelstadt B, Russell C and Wachter S, 'Explaining Explanations in AI', *In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)* (2019)

Naker S and Greenbaum D, 'Now You See Me: Now You Still Do: Facial Recognition Technology and the Growing Lack of Privacy' (2017) 23 BUJ Sci. & Tech. L. 88

Poelman R and others, 'As If Being There: Mediated Reality for Crime Scene Investigation', *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* (2012)

Quezada-Tavárez K, 'Law Enforcement AI in the Spotlight as EDPB Cast Doubt on Legality of Facial Recognition Tech' (*KU Leuven CiTiP Blog*, 20 July 2020) <https://www.law.kuleuven.be/citip/blog/law-enforcement-ai-in-the-spotlight-as-edpb-cast-doubt-on-legality-of-facial-recognition-tech/> accessed 17 June 2021

Quezada-Tavárez K and Houtave E, 'Augmented Reality Technology to Counter Crime and Terrorism: Introduction to DARLENE' (*CiTiP Blog*, 16 April 2021) <https://www.law.kuleuven.be/citip/blog/augmented-reality-technology-to-counter-crime-and-terrorism-introduction-to-darlene/> accessed 16 April 2021

R (Bridges) v Chief Constable of South Wales Police [2020] EWCA CIV 1058.

Robinson DG, 'The Challenges of Prediction: Lessons from Criminal Justice' (2018) 14 I/S: A Journal of Law and Policy for the Information Society 151

Sahu CK, Young C and Rai R, 'Artificial Intelligence (AI) in Augmented Reality (AR)-Assisted Manufacturing Applications: A Review' (2020) 59 International Journal of Production Research 4903

Sion L and others, 'DPMF: A Modeling Framework for Data Protection by Design' (2020) 15 Enterprise Modelling and Information Systems Architectures (EMISAJ) 1

Streefkerk JW and others, 'The Art of Csi: An Augmented Reality Tool (Art) to Annotate Crime Scenes in Forensic Investigation', *International conference on virtual, augmented and mixed reality* (Springer 2013)

Wassom B, *Augmented Reality Law, Privacy, and Ethics: Law, Society, and Emerging AR Technologies* (Syngress 2014)

Yung R and Khoo-Lattimore C, 'New Realities: A Systematic Literature Review on Virtual Reality and Augmented Reality in Tourism Research' (2019) 22 Current Issues in Tourism 2056

Zouave E and Schroers J, 'You've Been Measured, You've Been Weighed and You've Been Found Suspicious: Biometrics and Data Protection in Criminal Justice Processing' in Ronald Leenes and others, *Data Protection and Privacy: The Internet of Bodies* (Hart Publishing 2019)

ON THE POTENTIALITIES AND LIMITATIONS OF AUTONOMOUS SYSTEMS IN MONEY LAUNDERING CONTROL

By Leonardo Simões Agapito [*], *Matheus de Alencar e Miranda*[**] *and Túlio Felippe Xavier Januário*[***]

Abstract

This paper analyses the potential gains and eventual difficulties using autonomous systems – such as artificial intelligence (AI) mechanisms – to prevent, detect and investigate money laundering. As it is well-known, new technologies have been applied in the most varied social contexts, being no different in the case of the FIUs, especially when receiving and processing reports of suspicious activities from obligated entities. However, in addition to the already identified difficulties imposed by new technologies, the specific scope of money laundering presents particular challenges. Potential guidelines are proposed for a better interaction between AI and money laundering prosecution. For that, is is initially analysed what is effectively meant by AI and autonomous systems and how they are effectively used in this scope. Subsequently, some of the difficulties encountered in this context are demonstrated, ranging from insufficiency, low quality and inaccuracy of data that feed the systems, to the difficulties in understanding, explaining and allowing the refutation of the conclusions reached by them. From this analysis and through a deductive methodology, possible solutions are proposed that allow a better and more efficient interaction between humans and autonomous systems in the field of money laundering and its prosecution.

1 Introduction

It is currently undeniable that facing drug trafficking, terrorism and other misconducts related to organized criminality depends to a great extent on public instruments capable of preventing the circulation not only of the proceeds of crime, but also of the capital needed for its commission. It is in this context that criminal prosecution of money laundering reached a larger international consensus in terms of criminal policy in the 20th Century.[1]

[*] PhD Student in Latin American Integration, University of São Paulo. For correspondence: <leoagapito@gmail.com>.

[**] PhD Candidate in Criminal Law, State University of Rio de Janeiro. For correspondence: <matheus.alencarm@gmail.com>.

[***] PhD Fellow, Fundação para a Ciência e a Tecnologia (FCT), University of Coimbra. For correspondence: <tuliofxj@gmail.com>.

[1] See: Pierpaolo Cruz Bottini, 'Aspectos Conceituais da Lavagem de Dinheiro' in Gustavo Henrique Badaró and Pierpaolo Cruz Bottini (eds), *Lavagem de Dinheiro: Aspectos Penais e Processuais Penais: Comentários à Lei 9.613/98, com alterações da Lei 12.683/12* (4th edn, Thomson Reuters Brasil 2019) 25-29; Benjamin Vogel, 'Introduction' in Benjamin Vogel and Jean-Baptiste Maillart (eds), *National and International Anti-Money Laundering Law: Developing the Architecture of Criminal Justice, Regulation and Data Protection* (Intersentia 2020) 1.

The globalization experienced in money laundering, the trend toward crime professionalization and especially the complexity and ability to adapt and create new methods of committing these misbehaviors[2] are some of the challenges that require constant attention from the policymakers, when defining preventive and repressive strategies.[3-4]

As one example of such policies, the unification of security and inspection standards of banking systems became imperative with the intensification of international financial operations. In this scope, while international cooperation treaties on transnational and organized crime took a few years to be properly internalized and operationalized in the different signatory countries, the FATF recommendations and the articulation of the Egmont system were rapidly assimilated, imposing rigorous information standards from obligated entities. However, these recommendations do not always consider the diversified actuation of banking entities in their respective countries and may be ineffective in identifying possible criminal conducts.

The present work starts from the premise that ignoring local particularities may end up affecting the quality of data that foster autonomous systems used by the Financial Intelligence Units (FIUs), reducing their effectiveness and increasing the risk of false positives and false negatives. Furthermore, it is undeniable that autonomous systems and artificial intelligence (AI) mechanisms still have some inherent limitations, such as the opacity of their proceedings and the consequent difficulties in understanding (and sometimes refuting) their conclusions.

In view of these difficulties, the main goal of the present article is to understand how data- and regulation-related issues can affect the efficiency of money laundering prosecution, even with the employment of AI and autonomous systems for its detection and prevention. Aiming at proposing some guidelines for a better interaction between these sectors, we will initially describe what we can understand by AI and autonomous systems and how these technologies are effectively applied in the prevention, detection and

[2] Isidoro Blanco Cordero, *El Delito de Blanqueo de Capitales* (2nd edn, Aranzadi 2002) 51-55.

[3] As pointed out by Nuno Brandão, money laundering shows itself as the dark side of the globalization process, the liberalization of international exchanges and capital movements, the opening of markets, the massive computerization and the electronic commerce. If, on the one hand, there have always been economic criminality and attempts to dissimulate illicit gains, these activities have never been of such proportion and reached so many interests as in the present time. See: Nuno Brandão, *Branqueamento de Capitais: O Sistema Comunitário de Prevenção* (Coimbra Editora 2002) 16-17. For a detailed analysis of the impacts of technological innovations on money laundering, see also: Miguel Abel Souto, 'Blanqueo, Innovaciones Tecnológicas, Amnistía Fiscal de 2012 y Reforma Penal' (2012) 14 Revista Electrónica de Ciencia Penal y Criminología 1 <http://criminet.ugr.es/recpc/14/recpc14-14.pdf> accessed 14 July 2021.

[4] On the global context of money laundering and its characteristics, see also: Anna Carolina Canestraro, 'Compartilhamento de Dados e Persecução do Crime de Branqueamento de Capitais no Âmbito dos Paraísos Financeiros' (2018) 22(35) Revista de Estudos Jurídicos Unesp 135, 137-139 <https://doi.org/10.22171/rej.v22i35.2197> accessed 12 July 2021; Anna Carolina Canestraro, 'Cooperação Internacional em Matéria de Lavagem de Dinheiro: da Importância do Auxílio Direto, dos Tratados Internacionais e os Mecanismos de Prevenção' (2019) 5(2) Revista Brasileira de Direito Processual Penal 623, 626-633 <https://doi.org/10.22197/rbdpp.v5i2.234> accessed 12 July 2021.

prosecution of money laundering. Once these concepts and their respective applications are delimited, we will investigate their main limitations and obstacles. In the end, we will demonstrate how a new structure could be designed to be able not only to mitigate the well-known issues of data bias but also make regulatory enforcement in this matter more effective.

2 Autonomous Systems, AI and their Application to Money Laundering Control

It is currently common to encourage automation of all kinds of tasks, including those related to the criminal justice system. In particular, the automation of repetitive tasks and data analysis for investigations stands out. In this context, automation still occurs mainly through *autonomous systems,* which are previously programmed for autonomous decision-making (without immediate human intervention). However, an undeniable expansion in automation is currently observed due to an increase in employment of AI.[5]

According to the European Commission, AI can be understood as 'systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals'. In the Communication 'Artificial Intelligence for Europe', the Commission explains that these systems 'can be purely software-based, acting in the virtual world' (eg voice assistants) or 'embedded in hardware devices' (eg autonomous cars).[6]

Despite this definition, there is still some confusion in criminal justice system regarding the distinction between automation and AI.[7] For the purposes of this essay, we will consider AI as machine intelligence, capable of solving problems similarly to a human being, as having the ability to understand its environment through data inputs and, based on them, to choose a course of action among several possible others, aimed at solving a posed problem.

Also for the purposes of this essay, autonomous systems are those capable of reacting to the environment without the need for human intervention (therefore, autonomous) but unable to choose a course of action or create a new solution to a problem (therefore, not

[5] According to Fabiano Hartmann and Roberta Zumblick, artificial intelligence operates through the identification of patterns in the available database, prioritizing, from them, behaviors that have positive effects related to the objective sought. Widely used to find patterns and classify documents, this technology has expanded to other functions as well. See: Fabiano Hartmann Peixoto and Roberta Zumblick Martins da Silva, *Inteligência Artificial e Direito* (Alteridade Editora 2019) 63ff.

[6] European Commission, 'Communication From the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions (Artificial Intelligence for Europe)' COM(2018) 237 final.

[7] On these concepts and distinctions, see: Amedeo Santosuosso and Barbara Bottalico, 'Autonomous Systems and the Law: Why Intelligence Matters' in Eric Hilgendorf and Uwe Seidel (eds), *Robotics and the Law: Legal Issues Arising from Industry 4.0 Technology Programme of the German Federal Ministry for Economic Affairs and Energy* (Nomos 2017) 35ff.

intelligent). These systems only present a pre-programmed response according to the environment identified by them.[8]

Having in mind this distinction is important because it enables a better comprehension of the technological *state-of-art* and possible improvements in each kind of system. In this sense, AI can be seen as an 'umbrella concept', which encompasses several subfields such as robotics, machine learning and natural language processing.[9]

With specific regard to the application of these systems in criminal justice, we can observe their consolidated usage in the internationally standardized model for the prevention of money laundering. Preventive regulation of these crimes requires constant data provision from obligated agents. This data provision forms a huge database that is impossible to be manually verified in the scope of prevention, detection and repression of suspect transactions. Thus, there is a rise in demand for automation in the verification of patterns, which greatly encourages the use of autonomous systems and creates conditions for the development of massive and problem-solving AI.

Highlighting the international standardization of regulation in this matter, we must mention the Basel Capital Accord I, of 1988, whose main objectives were: a) 'to strengthen the soundness and stability of the international banking system'; and that b) 'the framework should be in fair and have a high degree of consistency in its application to banks in different countries with a view to diminishing an existing source of competitive inequality among international banks'.[10] In order to achieve these goals, the document devoted special attention to establishing risk assessment standards. It is important to mention that the three subsequent accords aimed to respond to the intensification of financialization and to the economic crises of the 1990s and of 2008, demonstrating that financial systems were already deeply integrated.[11]

Financial systems integration had an undeniable impact on control mechanisms and on the need for their coordination coordination. Following the FATF recommendations, they can be categorized into: (a) information duties; (b) compliance duties.

---

[8] We are aware that the complexity of the distinction is higher than that briefly presented here. However, in terms of juridical and criminal consequences, it is essential to distinguish these two types of technology. This delimitation will directly reverberate in the conclusions that we will reach in this article. For a more detailed analysis of this issue, see: Eric Hilgendorf, 'Recht und autonome Maschinen – ein Problemaufriß' in Eric Hilgendorf and Sven Hötitzsch (eds), *Das Recht vor den Herausforderungen der modernen Technik* (Nomos 2015); Peixoto and Silva (n 5).

[9] Ryan Calo, 'Artificial Intelligence Policy: A Primer and Roadmap' (2017) 51(2) UC Davis Law Review 399, 405 <https://lawreview.law.ucdavis.edu/issues/archive.html> accessed 13 July 2021; Peixoto and Silva (n 5) 75.

[10] Basel Committee on Banking Supervision, 'International Convergence of Capital Measurement and Capital Standards (Basel Capital Accord I)' (1988) 1.

[11] Peter Went, 'Basel III Accord: Where Do We Go From Here?' (2010), 11 <https://dx.doi.org/10.2139/ssrn.1693622> accessed 13 July 2021.

As regards (a) information duties, the following is suggested: i) *national cooperation and coordination mechanisms* (Recommendation n. 2); ii) '*that financial institution secrecy laws do not inhibit implementation of the FATF Recommendations*' (Recommendation n. 9); iii) *customer due diligence,* when certain conditions are observed and following certain procedures (Recommendation n. 10); iv) *record-keeping* (Recommendation n. 11); v) *reporting of suspicious transactions* (Recommendation n. 20) and vi) *transparency and beneficial ownership of legal arrangements* (Recommendation n. 25).[12]

Regarding b) compliance duties, it is recommended that States, when implementing an action model, assess the main sources of risk to which their institutions are subject, promoting programs that effectively curb terrorist financing and money laundering. This analysis is also required from institutions and professionals that operate in the financial system.[13]

With regard to the FIUs, the FATF reinforces the importance of their autonomy and active role, recommending a broad scope of powers and responsibilities of competent authorities (Recommendations 26-35).[14]

The imposition of the security systems resulting from these recommendations encourages regulatory models such as good governance and compliance, in addition to differentiated models of responsibility attribution, aiming at responding to the 'organizational deficits' or states of 'organized irresponsibility' from companies. In order to guarantee the stability of the economic system, collaboration duties are resorted to.[15]

In addition, in sectors that are sensitive to money laundering, some people are selected to act as *gatekeepers* against suspicious activities, preventing the commitment of crimes and reporting it to the central intelligence agency.[16] Taking this preventive potential as a premise, legislations around the world have intensely focused on creating duties for these economic agents. However, this premise is imprecise because the model promotes dependence of control agents on the information provided by *gatekeepers* and the quality

[12] FATF, 'International Standards on Combating Money Laundering and the Financing of Terrorism & Proliferation (The FATF Recommendations)' (2012-2020).

[13] ibid.

[14] '29. Financial intelligence units * Countries should establish a financial intelligence unit (FIU) that serves as a national centre for the receipt and analysis of: (a) suspicious transaction reports; and (b) other information relevant to money laundering, associated predicate offences and terrorist financing, and for the dissemination of the results of that analysis. The FIU should be able to obtain additional information from reporting entities and should have access on a timely basis to the financial, administrative and law enforcement information that it requires to undertake its functions properly'. Ibid.

[15] Eduardo Saad-Diniz, 'Fronteras del Normativismo: a Ejemplo de las Funciones de la Información en los Programas de Criminal Compliance' (2013) 108 Revista da Faculdade de Direito da Universidade de São Paulo 415, 423.

[16] See: John C. Coffee Jr., 'Understanding Enron: It's About the Gatekeepers, Stupid' (2002) 207 Columbia Law School Working Paper 1, 5 <https://dx.doi.org/10.2139/ssrn.325240> accessed 14 July 2021; John C. Coffee Jr, 'The Attorney as Gatekeeper: An Agenda for the SEC' (2003) 103(5) Columbia Law Review 1293, 1296ff <https://doi.org/10.2307/1123838> accessed 14 July 2021; Ana Carolina Carlos de Oliveira, *Lavagem de dinheiro: responsabilidade pela omissão de informações* (Tirant lo Blanch 2019) 30.

of this information is not consistent. Since this dependence can generate a series of problems, this 'surrender' of the supervisory body to interinstitutional cooperation is seemingly no longer adequate, requiring a new framework that can change the correlation of forces through the implementation of new mechanisms.

In any case, one of the main tasks of the FIUs is to receive the Suspicious Activity Reports (SARs) and other information related to money laundering, having access to public (and often private commercial) databases to properly perform their analysis.[17] At this point, the issue regarding the use of data analytics and data mining in these procedures gains relevance.

In Brazil[18], for example, the SARs received by the FIU are submitted to a pre-programmed electronic analysis and distributed individually to technical analysts. Both the communication and its procedure are registered in the same software, so that the database can have an increasing and constructive volume that will serve as subsidy resolutions of subsequent communications.

This logic is the same as that applied to several AI tools: identification of patterns in the database and detection of other similar operations and new patterns. In this case, the patterns are the ones that, based on recorded financial transactions, indicate money laundering. The correlation probability between the operation and the pattern is equal to the probability vector of the money laundering risk matrix.

In the Brazilian system, following the procedure mentioned above, the 'Risk and Priority Management Center (CGRP)' scrutinizes each communication and creates a specific file for each case. The cases are ranked by the CGRP according to the degree of risk, in a procedure that already follows the logic of an autonomous system: the higher the risk assessed by the system, the greater attention will be given to the case.

To summarize, we can affirm that the contemporary model of money laundering and terrorist financing prevention (which is globally standardized) has automation at the basis of its prioritization of investigations and the Brazilian case is a good example of this kind of practice. We can observe that if, on the one hand, the procedure is currently performed by an autonomous system with some human intervention, on the other, the sector has enormous potential for AI application, given the use of large databases for the identification of patterns and subsequent detection of similar operations or new patterns of money laundering.

---

[17] Jean-Baptiste Maillart, 'Anti-Money Laundering Architectures: Between Structural Homogeneity and Functional Diversity' in Benjamin Vogel and Jean-Baptiste Maillart (eds), *National and International Anti-Money Laundering Law: Developing the Architecture of Criminal Justice, Regulation and Data Protection* (Intersentia 2020) 839ff.

[18] For a more detailed analysis of the Brazilian preventive model: COAF, *Casos & Casos: I Coletânea de Casos Brasileiros de Lavagem de Dinheiro: Edição Comemorativa pelos 10 Anos do Conselho de Controle de Atividades Financeiras* (COAF 2011) 10ff.

The consequences of the autonomy of the procedure can be seen in the continuity of the investigation and in the subsequent prosecution of money laundering. Currently, Financial Intelligence Reports can be instigated: i) spontaneously by FIUs; (ii) from exchanges of information with other regulatory agencies; iii) requested by a foreign authority. If the autonomous analysis indicates signs of money laundering, the report must be sent ahead to the competent authorities, alongside with all the evidence collected. From this moment on, the reports often serve as a subsidy for criminal investigations and evidence in criminal proceedings.

Considering international requirements and the particularities of money laundering and its perpetrators, it is evident that investigation and prevention by FIUs demand major financial, technological and personnel resources. However, practical experience demonstrates that these expectations are not fulfilled, specially taking into account the infrastructure of the FIUs in developing countries.[19]

That said, it is possible to point out an interdependence between regulation and automation in the system of money laundering prevention. In other words, it is clear that the feasibility of regulation enforcement depends on automation, given the discrepancy between the volume of SARs to be analyzed and the human resources available to carry out the enforcement. On the other hand, the automation of this system is only possible due to the database provided by the regulation. In short, without database there is no system development (autonomous or AI). Ultimately, the success of automation will always depend on the quality of the data provided by the regulation and the success of the latter will always depend on the quality of automation.

3 Challenges in Using AI to Control Money Laundering

Considering the abovementioned regulatory standards, AI systems involved in money laundering surveillance face at least three different kinds of challenges, which are: a) *inadequacy of data produced by FIUs*; b) *lack of reliability of data produced by FIUs*; c) *opacity of AI*. We will now analyze each of them in detail.

3.1 Insufficiency and inadequacy of data

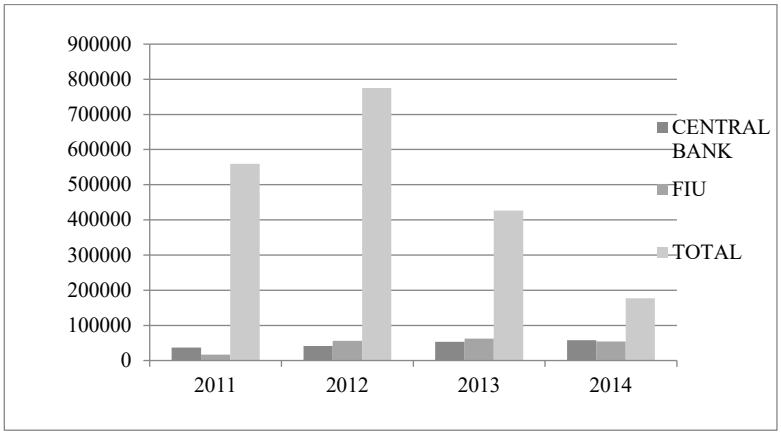
When analyzing the FIUs' reports, it remains clear that the abovementioned informational standards end up creating further issues, besides money laundering itself. As an example, it has been already demonstrated that, in the last decade, the Brazilian FIU suffered a decrease in its efficiency along the years, contrasting with the expansion of informational duties. At the beginning of its activities, there was an increase in the number of

[19] In Brazil, for example, according to the Federal Decree n. 9.003/17, the FIU has 31 employees, 15 of them responsible for analyzing and supervising the SARs. Considering the volume of almost 1.5 million SARs, it is evident that, without automation, the FIUs would be unable to perform their duties. See: José Carlos de Oliveira, Leonardo Simões Agapito and Matheus de Alencar, 'O Modelo de "Autorregulação Regulada" e a Teoria da Captura: Obstáculos à Efetividade no Combate à Lavagem de Dinheiro no Brasil' (2017) 10 Quaestio Iuris 365, 378-381 <https://doi.org/10.12957/rqi.2017.26847> accessed 14 July 2021.

investigations, charges and convictions for money laundering.[20] Furthermore, the Central Bank of Brazil used to be effective back then in demanding and analyzing accurate information from its regulated entities before transferring it to the FIU. Nevertheless, while the number of SARs increased, the efficiency dropped. According to the FIU's statistics, in 2009, 93,270 SARs were reported by the sectors regulated by the Central Bank of Brazil, with a 57% level of efficacy (useful percentage of the information provided).[21] After the internalization of Basel III Standards in 2010 and the intensification of the FATF's demands for an anti-terrorism agenda, the number of operations reported by the Brazilian Central Bank grew to 1,289,087 in 2011; 1,587,427 in 2012; 1,286,233 in 2013; and 1,144,542 in 2014, but the efficacy level was below 30%.[22]

With regard to atypical operations, which have attracted major attention from regulatory authorities, since they present evidence of money laundering, the numbers appear to follow a downward trend: 559,992 in 2011 (37,237 from the Central Bank; 16,684 from the UIF); 775,535 in 2012 (41,819 from the Central Bank; 55,646 from the UIF); 426,153 in 2013 (53,244 from the Central Bank and 62,732 from the UIF) and 177,467 in 2014 (57,455 from the Central Bank and 53,818 from the UIF). In absolute numbers, graphically:[23]

Graph 1 – The SARs reported to the FIU in Brazil[24]



The graph shows the above-mentioned decrease in the total number of suspicious operations over time. If communications plummeted, however, the proportion of suspicious transactions among those notified also decreased, reaching the number of 177,467 atypi-

[20] Remarkably in the period from 2003 to 2006, when a massive computerization and major integration between authorities occurred (especially between the FIU and the Federal Police). See: Vanessa Alessi Manzi, *Compliance no Brasil: consolidação e perspectivas* (Saint Paul 2008) 57-59.
[21] Marcelo de Aguiar Coimbra and Vanessa Alessi Manzi, *Manual de Compliance* (Atlas 2010) 72.
[22] Oliveira, Agapito and Alencar (n 19) 378-381.
[23] ibid. 379.
[24] ibid. 379.

cal transactions compared to 1,144,542 total communications in 2014. This amount to almost one million communications not used for elaborated investigations, only stored as data.

Thus, we can observe some consequences caused by the tightening of duties: i) an initial exponential increase of communications; ii) a reduction of effectiveness of the inspections; iii) changes in procedure of regulated entities, which started to financially behave below 'atypical' standards.

It is crucial to highlight that while this reactive movement is itself capable of creating different problems for the analysis by autonomous systems (especially in terms of bias, false positives and false negatives), the high number of useless operations reveals an even major problem: the data have not been properly used for effective action.

3.2 Unreliable data

The global model of money laundering prevention is based on trust in the communications made by obligated agents. From them, a database of operations will be created, which will be the basis for the analysis and detection of suspicious operations. Therefore, it is important to verify whether these communications are providing a reliable database.

However, there are different mechanisms and opportunities for obligated entities to manipulate and to mislead the authorities. In other words, a regulated agent can take advantage of the regulatory entity's dependence and the minimum possibility of being discovered to falsify or hide essential information. At the same time, he or she can purposefully communicate various supposedly suspicious operations, which are known to be lawful, in order to overload the regulators with information and, thus, increase their dependence.[25]

In such situations, it is evident that the autonomous system of money laundering prevention is hampered by the low quality of the database. A solution to this issue is complex, since the regulatory body does not even have enough means to check all the communications provided. The verification of those that were maliciously provided is even less feasible.

On the other hand, in cases where suspicious activities are not reported, the discovery of these crimes is also unlikely – the discovery being generally dependent on criminal proceedings involving other offenses. And even when they are discovered, administrative punishment for incorrect communication from the obligated agents rarely occurs, given

[25] An exemplary case occurred in Brazil, within the scope of Criminal Action 470/MG ('Mensalão'). One of the defendants was convicted of money laundering by the Supreme Court and the decision was based, among other reasons, on the falsification of data and omission of communications to the FIU. It is important to highlight that this conduct was not detected by the regulatory agency and was only discovered in the course of the criminal action, through documents and witnesses that contradicted the content of the false communications. For more details, see: Oliveira, Agapito and Alencar (n 19) 381.

the lack of alignment between different control bodies, especially between the public prosecutors and enforcement agencies.

Finally, regulated institutions have also developed their own autonomous systems to manage risks and guide enterprises through new opportunities. However, these systems are hardly comprehended even by their developers. In Brazil, for example, the new regulation on stock markets demands a complex report on risk assessment programs, including on numbers of operations detected and notifications.[26] These numbers are consistently useless to guarantee effectiveness but will be utterly understood under a qualitative verification. It is necessary to create a new framework to validate the employment of AI in money laundering risk management programs.

3.3 Limitations of AI

In addition to the aforementioned difficulties encountered in the prevention and prosecution of money laundering, we cannot disregard the fact that autonomous systems and AI have also serious limitations, which certainly reverberate when applied in this field.

Firstly, since these technologies depend on mass processing of data, the concern about the security, reliability and lawfulness of these data is crucial.[27] More specifically, it is important to ensure that they were not obtained by violating rights of their holders. In any case, there is an undeniable risk that these data may be biased, as they may end up reflecting their developers' prejudices and discriminations.[28]

Furthermore, there are undeniable difficulties in understanding, controlling and, consequently, refuting the conclusions reached by the AI and its algorithms. For this reason, AI is considered opaque, since there are no concrete conditions for measuring 'how' and 'why' the outputs are produced, and even the input is often unknown. That is why these algorithms are commonly equated to 'black boxes'.[29]

---

[26] Comissão de Valores Mobiliários, 'Instrução CVM n.617, de 5 de dezembro de 2019' <http://conteudo.cvm.gov.br/legislacao/instrucoes/inst617.html> accessed 13 July 2021.

[27] Caitlin Mulholland and Isabella Z. Frajhof, 'Inteligência Artificial e a Lei Geral de Proteção de Dados Pessoais: Breves Anotações Sobre o Direito à Explicação Perante a Tomada de Decisões por Meio de Machine Learning' in Ana Frazão and Caitlin Mulholland (eds), *Inteligência Artificial e Direito: Ética, Regulação e Responsabilidade* (Thomson Reuters Brasil 2019).

[28] See: Adrienne Yapo and Joseph Weiss, 'Ethical Implications of Bias in Machine Learning' [2018] *Proceedings of the 51st Hawaii International Conference on System Sciences* 5365, 5366; Peixoto and Silva (n 5) 34-35; Túlio Felippe Xavier Januário, 'Considerações Preambulares Acerca das Reverberações da Inteligência Artificial no Direito Penal' in Murilo Siqueira Comério and Tainá Aguiar Junquilho (eds), *Direito e Tecnologia: um debate multidisciplinar* (Lumen Juris 2021).

[29] See: Jenna Burrell, 'How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms' (2016) 3(1) Big Data & Society 1, 1 <https://doi.org/10.1177/2053951715622512> accessed 23 January 2020; William Nicholson Price II, 'Artificial Intelligence in Health Care: Applications and Legal Issues' (2017) 599 U of Michigan Public Law Research Paper 1, 2 <https://ssrn.com/abstract=3078704> accessed 23 January 2020; Miriam Wimmer, 'Inteligência Artificial, Algoritmos e o Direito: Um Panorama dos Principais Desafios' in Ana Paula M. Canto de Lima, Carmina Bezerra Hissa and Paloma Mendes Saldanha (eds), *Direito Digital: Debates Contemporâneos* (Thomson Reuters Brasil 2019); Anabela Miranda

Without disregarding the undeniable benefits of AI, it is certain that its limitations imply difficulties to be faced in the most diverse sectors in which this technology is applied.[30] When it refers to usage that directly or indirectly impacts on the criminal justice system, these difficulties are even more accentuated, given the importance of the interests in question.

Far beyond the relevant reverberations in evidentiary matters and the countless controversies that they raise,[31] the progressive usage of autonomous systems and AI in decision-making in several phases of intelligence, investigation and judicial instruction procedures sparks endless debates regarding AI's feasibility and limits.[32]

In the scope of money laundering prevention, similar questions must be considered. What kind of public and private data can be used by autonomous systems? How people directly affected by these systems could understand the reasons and contest eventual

Rodrigues, 'Inteligência Artificial no Direito Penal – A Justiça Preditiva entre a Americanização e a Europeização' in Anabela Miranda Rodrigues (ed), *A Inteligência Artificial no Direito Penal* (Almedina 2020) 25.

[30] For an exemplary study of these potentialities and difficulties in the sectors of autonomous vehicles, medicine and stock market, see: Túlio Xavier Januário, 'Veículos Autónomos e Imputação de Responsabilidades Criminais por Acidentes' in Anabela Miranda Rodrigues (ed), *A Inteligência Artificial no Direito Penal* (Almedina 2020) 95ff; Túlio Felippe Xavier Januário, 'Inteligência Artificial e Responsabilidade Penal no Setor da Medicina' (2021) 17(34) Lex Medicinae: Revista Portuguesa de Direito da Saúde 37 <https://www.centrodedireitobiomedico.org/publica%C3%A7%C3%B5es/revistas> accessed 15 July 2021; Túlio Felippe Xavier Januário, 'Inteligência Artificial e Manipulação do Mercado de Capitais: uma Análise das Negociações Algorítmicas de Alta Frequência (High-Frequency Trading – HFT) à Luz do Ordenamento Jurídico Brasileiro' (2021) 29(186) Revista Brasileira de Ciências Criminais (forthcoming).

[31] For a broad analysis of possible evidentiary issues arising from artificial intelligence, see: Serena Quattrocolo, *Artificial Intelligence, Computational Modelling and Criminal Proceedings: A Framework for a - European Legal Discussion* (Springer 2020) 37ff; Sabine Gless, 'AI in the Courtroom: A Comparative Analysis of Machine Evidence in Criminal Trials' (2020) 51(2) Georgetown Journal of International Law 195, 202ff <https://ssrn.com/abstract=3602038> accessed 15 July 2021; Sónia Fidalgo, 'A Utilização de Inteligência Artificial no Âmbito da Prova Digital – Direitos Fundamentais (Ainda Mais) Desprotegidos' in Anabela Miranda Rodrigues (ed), *A Inteligência Artificial no Direito Penal* (Almedina 2020) 129ff. For a specific analysis on the digital chain of custody, see: Túlio Felippe Xavier Januário, 'Cadeia de Custódia da Prova e Investigações Internas Empresariais: Possibilidades, Exigibilidade e Consequências Processuais Penais de sua Violação' (2021) 7(2) Revista Brasileira de Direito Processual Penal 1453 <https://doi.org/10.22197/rbdpp.v7i2.453> accessed 12 October 2021.

[32] See: Danielle Kehl, Priscilla Guo, and Samuel Kessler, 'Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing' (2017) *Responsive Communities Initiative* <http://nrs.harvard.edu/urn-3:HUL.InstRepos:33746041> accessed 15 July 2021; Vicent Chiao, 'Fairness, Accountability and Transparency: Notes on Algorithmic Decision-Making in Criminal Justice' (2019) 15 International Journal of Law in Context 126 <https://doi.org/10.1017/S1744552319000077> accessed 15 July 2021; Anabela Miranda Rodrigues, 'A Questão da Pena e a Decisão do Juiz – entre a Dogmática e o Algoritmo' in Anabela Miranda Rodrigues (ed), *A Inteligência Artificial no Direito Penal* (Almedina 2020) 230ff; Luís Greco, *Poder de Julgar sem Responsabilidade de Julgador: a Impossibilidade Jurídica do Juiz-Robô* (Marcial Pons 2020) 17ff.

outputs that may be prejudicial to them? And most importantly, is the output from these technologies really trustworthy?

4 Building a New Framework: Surpassing Data Bias and AI Ambiguities

4.1 Controlling information overload and false positives by autonomous decisions

Even in a country such as Brazil, where the banking sector is underexplored, there is an immense quantity of SARs to be audited. An operational example of this problem is presented by Jun Tang and Lishan Ai in China,[33] where a bank failed to comply with information duties until it was punished for the lack of reports.[34] On the very next 30 days, 1,700 SARs were reported. It seems that compliance programs end up being designed by banks to transfer or avoid responsibility, but not to effectively collaborate. The first mistake is to consider that the ineffectiveness of compliance programs in bank sectors is only their fault, even when they are complying with regulation.

To avoid an overly large number of false positives, information patterns must mature data before the report, which means that banks should check those transactions in a more complex system of conditions and characteristics. A great example was proposed by Zengan Gao and Mao Ye, who indicate that regulators should explore the decision tree and Bayesian inference systems, mixing different criteria to demonstrate how unusual, abnormal, or illegal a specific suspicious transaction might be.[35] Those data would be easily cross-checked by AI programs, which are already used to prevent credit frauds.

As previously presented in another paper, money laundering cannot be recognized by an isolated transaction.[36] It is important to take a step back and look at the big picture, just as in any other organized crime investigation. On banking reports, it is important to assess not only transactions but also people involved, economic activities informed, different groups linked and public profiles. To 'follow the money' is to investigate a complex chain of exchanges, not a simple line of transfers. In this sense, Zengan Gao and Mao Ye propose: a) to identify central members, subgroups and 'money laundering networks'; b) a case-based system of information (which can be elaborated with machine learning

[33] For a comprehensive study on money laundering control in China, see: Jing Lin, *Compliance and Money Laundering Control in China: Self Control, Administrative Control and Penal Control* (Duncker & Humblot 2016) 18ff.

[34] Jun Tang and Lishan Ai, 'The System Integration of Anti-Money Laundering Data Reporting and Customer Relationship Management in Commercial Banks' (2013) 16(3) Journal of Money Laundering Control 231, 232 <https://doi.org/10.1108/JMLC-04-2013-0010> accessed 13 July 2021.

[35] Zengan Gao and Mao Ye, 'A Framework for Data Mining-Based Anti-Money Laundering Research' (2007) 10(2) Journal of Money Laundering Control 170, 171 <http://www.emeraldinsight.com/1368-5201.htm> accessed 13 July 2021.

[36] Matheus de Alencar e Miranda and Leonardo Simões Agapito, 'Critérios de Validade e Eficiência de Compliance e Impactos na Interpretação da Lavagem de Dinheiro' in Eduardo Saad-Diniz, Luís Augusto Brodt, Henrique Abi-Ackel Torres and Luciano Santos Lopes (eds), *Direito Penal Econômico nas Ciências Criminais* (Vorto 2019) 241ff.

programs); c) a data mining technique that could sum 'customer, account, product, geography, and time' information by vectors analysis.[37]

As reported by Jun Tang and Lishan Ai, different mechanisms of data mining have been already applied by financial institutions to comprehend their clients, which are classified and evaluated for commercial and risk assessment proposes. Even home banking behaviors and smartphone apps and cookies are collected as market strategy. Banks knows their clients much more than what has been asked and profiles created should be better explored by the FIUs.[38]

However, that also means that national and international authorities of personal data protection would play a central role in the banking sector, whose institutions must be obliged to present their data mining programs without anonymization. At this point, a more collaborative framework between different authorities of personal data protection and companies (regulator-regulator and regulator-bank) becomes as important as the FIUs' reports.

4.2 Information bias: improving data analyses by human intervention

Changing informational standards might be very ineffective if the reports are not reliable. As demonstrated before, informational standards have been enforced and redesigned, creating new duties that were only able to change information volumes without impact on administrative or penal procedures. To improve data analysis, at least four measures are required, considering the need of a relationship of trust between *gatekeepers* and public auditors.

Taking Brazil as an example once again, the absence of instruments for whistleblowing protection is an important issue for a more collaborative regulatory framework. In this scenario, even the Personal Data Protection Law (13.709/18) failed to define a Data Protection Officer, whose duties accumulated in the same agents (controllers) responsible for creating and controlling those systems. It became the best scheme for private auditors and commerce of certifications by big companies. There is no need of an external auditor if a legal protection for *gatekeepers* exists and if developers and operators of data mining programs demonstrate good performance.

Besides that, international regulatory standards on money laundering prevention rely on an agency model, which has its function compromised by big companies' complexity and a lack of attention from consumers. Public interest is also captured by market's interests. To build a new regulatory framework, third parties' representatives, unions and NGOs should be better listened to. Popular participation is essential for accountability, a balance between regulators and *gatekeepers* and for a plural perspective of data efficiency. It also strengthens the informal social control, which can be aligned with formal

[37] Gao and Ye (n 35) 171.
[38] Tang and Ai (n 34) 232.

social control to counter undesired behavior more effectively. In this case, the undesired behavior is AI manipulation.

Database bias may also be intentionally designed in a way that things that are not informed (false negatives) might be audited by a more proactive performance of regulators. The simplest mechanism of verification is the inspection *in loco*, observed when a public agent has open access to corporate computers, physical files, and workers. The inspection might create some positive effects, such as the institutional 'materialization' and employees' collaboration. When a public agent visits a company, it is an opportunity to solve many questions regarding legal standards and official reports. Agent's reports may also reveal companies' innovations and red flags, creating a better perspective of companies changes through the years. Inspections *in loco* may also create a safe space for employees that intent to collaborate but feel insecure about official channels.

However, institutional 'materialization' also creates opportunities for illegal favors and can also be easily deflected by a reactive attitude and trained behaviors. In this scenario, those inspections could become expensive, with low effectiveness. A remote inspection (by digital platforms) could be cheaper and faster but might also be easily deflected by cosmetic compliance programs.

A second model of verification might occur through a sandbox experiment. Regulatory sandboxes are already used by monetary authorities and reserve banks to develop new ideas and to test business models.[39] A sandbox experiment allows companies to implement an idea for a limited period with special normative conditions. The project must be well demonstrated before its implementation and all the data produced are collected by agencies to understand its potentials, vulnerabilities and opportunities. Thus, it would be possible for FIUs to create sandboxes to validate (or not) institutional systems of surveillance, data mining, and even autonomous reports. This option is much cheaper than inspections and it provides more reliable information, since corporations would have a lot of interest in collaborating and receiving a FIU's certification.

If the present enforcement model works well, it is also possible that the data bias problem is eased, creating the best scenario for using AI. With good data (or avoiding data bias), many of the AI problems (as pointed in 3.3) may be solved. Other problems are usually addressed through upgrades in transparency, by making the AI's objectives public, by development documentation (with making business rules transparent and clear to users) and, eventually, the coding itself.

[39]As an example: Banco Central do Brasil, 'Sandbox Regulatório' <https://www.bcb.gov.br/estabilidade financeira/sandbox> accessed 13 July 2021. On the topic of regulatory sandboxes and their importance in the scope of new technologies, see: Susana Aires de Sousa, '"Não Fui Eu, Foi a Máquina": Teoria do Crime, Responsabilidade e Inteligência Artificial' in Anabela Miranda Rodrigues (ed), *A Inteligência Artificial no Direito Penal* (Almedina 2020) 86ff.

5 Conclusion

As demonstrated, the main issue of autonomous decisions in money-laundering surveillance is data bias created by an empty and insufficient regulatory framework. In addition to that, financial integration promoted by Basel Accords and FATF (top-down regulation) is responsible for similar regulatory struggles in different local economic realities, which means that those regulatory standards might have to be reviewed from FIUs' experiences (bottom-up regulation). These developments might emerge from institutional changes on FIUs, but also from new regulatory experiments.

However, the greatest challenge on autonomous decisions is still related to the question about how to promote the disclosure of autonomous decisions steps. The overcoming of data bias might guarantee a more reliable AI system, but not a more legitimate one. At this point, understanding that autonomous decisions have limitations and might demand human verification in this scope may be necessary. Satisfactory investigations and valid sanctions may never be conducted exclusively by autonomous systems. However, complex algorithms are able to assist human surveillance and to ensure security and anonymity of the data (both useful and useless). That being said, a well-developed system might legitimate itself through its efficiency results during previous tests and permanent monitoring.

References

Abel Souto M, 'Blanqueo, Innovaciones Tecnológicas, Amnistía Fiscal de 2012 y Reforma Penal' (2012) 14 Revista Electrónica de Ciencia Penal y Criminología 1 <http://criminet.ugr.es/recpc/14/recpc14-14.pdf> accessed 14 July 2021

Banco Central do Brasil, 'Sandbox Regulatório' <https://www.bcb.gov.br/estabilidadefinanceira/sandbox> accessed 13 July 2021

Basel Committee on Banking Supervision, 'International Convergence of Capital Measurement and Capital Standards (Basel Capital Accord I)' (1988)

Blanco Cordero I, El Delito de Blanqueo de Capitales (2nd edn, Aranzadi 2002)

Bottini PC, 'Aspectos Conceituais da Lavagem de Dinheiro' in Gustavo Henrique Badaró and Pierpaolo Cruz Bottini (eds), Lavagem de Dinheiro: Aspectos Penais e Processuais Penais: Comentários à Lei 9.613/98, com alterações da Lei 12.683/12 (4th edn, Thomson Reuters Brasil 2019)

Brandão N, Branqueamento de Capitais: O Sistema Comunitário de Prevenção (Coimbra Editora 2002)

Burrell J, 'How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms' (2016) 3(1) Big Data & Society 1 <https://doi.org/10.1177/2053951715622512> accessed 23 January 2020

Calo R, 'Artificial Intelligence Policy: A Primer and Roadmap' (2017) 51(2) UC Davis Law Review 399 <https://lawreview.law.ucdavis.edu/issues/archive.html> accessed 13 July 2021

Canestraro AC, 'Compartilhamento de Dados e Persecução do Crime de Branqueamento de Capitais no Âmbito dos Paraísos Financeiros' (2018) 22(35) Revista de Estudos Jurídicos Unesp 135 <https://doi.org/10.22171/rej.v22i35.2197> accessed 12 July 2021

—— 'Cooperação Internacional em Matéria de Lavagem de Dinheiro: da Importância do Auxílio Direto, dos Tratados Internacionais e os Mecanismos de Prevenção' (2019) 5(2) Revista Brasileira de Direito Processual Penal 623 <https://doi.org/10.22197/rbdpp.v5i2.234> accessed 12 July 2021

Chiao V, 'Fairness, Accountability and Transparency: Notes on Algorithmic Decision-Making in Criminal Justice' (2019) 15 International Journal of Law in Context 126 <https://doi.org/10.1017/S1744552319000077> accessed 15 July 2021

COAF, *Casos & Casos: I Coletânea de Casos Brasileiros de Lavagem de Dinheiro: Edição Comemorativa pelos 10 Anos do Conselho de Controle de Atividades Financeiras* (COAF 2011)

Comissão de Valores Mobiliários, 'Instrução CVM n.617, de 5 de dezembro de 2019' <http://conteudo.cvm.gov.br/legislacao/instrucoes/inst617.html> accessed 13 July 2021

Coffee Jr JC., 'Understanding Enron: It's About the Gatekeepers, Stupid' (2002) 207 Columbia Law School Working Paper 1 <https://dx.doi.org/10.2139/ssrn.325240> accessed 14 July 2021

—— 'The Attorney as Gatekeeper: An Agenda for the SEC' (2003) 103(5) Columbia Law Review 1293 <https://doi.org/10.2307/1123838> accessed 14 July 2021

Coimbra MA and Manzi VA, *Manual de Compliance* (Atlas 2010)

European Commission, 'Communication From the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions (Artificial Intelligence for Europe)' COM(2018) 237 final

Fidalgo S, 'A Utilização de Inteligência Artificial no Âmbito da Prova Digital – Direitos Fundamentais (Ainda Mais) Desprotegidos' in Anabela Miranda Rodrigues (ed), *A Inteligência Artificial no Direito Penal* (Almedina 2020)

Gao Z and Ye M, 'A Framework For Data Mining-Based Anti-Money Laudering Research' (2007) 10(2) Journal of Money Laundering Control 170 <http://www.emerald insight.com/1368-5201.htm> accessed 13 July 2021

Gless S, 'AI in the Courtroom: A Comparative Analysis of Machine Evidence in Criminal Trials' (2020) 51(2) Georgetown Journal of International Law 195 <https://ssrn.com/abstract=3602038> accessed 15 July 2021

Greco L, *Poder de Julgar sem Responsabilidade de Julgador: a Impossibilidade Jurídica do Juiz-Robô* (Marcial Pons 2020)

Hilgendorf E, 'Recht und autonome Maschinen – ein Problemaufriβ' in Eric Hilgendorf and Sven Hötitzsch (eds), *Das Recht vor den Herausforderungen der modernen Technik* (Nomos 2015)

Januário TFX, 'Veículos Autónomos e Imputação de Responsabilidades Criminais por Acidentes' in Anabela Miranda Rodrigues (ed*), A Inteligência Artificial no Direito Penal* (Almedina 2020)

—— 'Inteligência Artificial e Responsabilidade Penal no Setor da Medicina' (2021) 17(34) Lex Medicinae: Revista Portuguesa de Direito da Saúde 37 <https://www.centrodedireitobiomedico.org/publica%C3%A7%C3%B5es/revistas> accessed 15 July 2021

—— 'Cadeia de Custódia da Prova e Investigações Internas Empresariais: Possibilidades, Exigibilidade e Consequências Processuais Penais de sua Violação' (2021) 7(2) Revista Brasileira de Direito Processual Penal 1453 <https://doi.org/10.22197/rbdpp.v7i2.453> accessed 12 October 2021

—— 'Considerações Preambulares Acerca das Reverberações da Inteligência Artificial no Direito Penal' in Murilo Siqueira Comério and Tainá Aguiar Junquilho (eds), *Direito e Tecnologia: um debate multidisciplinar* (Lumen Juris 2021)

—— 'Inteligência Artificial e Manipulação do Mercado de Capitais: uma Análise das Negociações Algorítmicas de Alta Frequência (High-Frequency Trading – HFT) à Luz do Ordenamento Jurídico Brasileiro' (2021) 29(186) Revista Brasileira de Ciências Criminais (forthcoming)

Kehl D, Guo P and Kessler S, 'Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing' (2017) *Responsive Communities Initiative* <http://nrs.harvard.edu/urn-3:HUL.InstRepos:33746041> accessed 15 July 2021

Lin J, *Compliance and Money Laundering Control in China: Self Control, Administrative Control and Penal Control* (Duncker & Humblot 2016)

Maillart JB, 'Anti-Money Laundering Architectures: Between Structural Homogeneity and Functional Diversity' in Benjamin Vogel and Jean-Baptiste Maillart (eds), *National and International Anti-Money Laundering Law: Developing the Architecture of Criminal Justice, Regulation and Data Protection* (Intersentia 2020)

Manzi VA, *Compliance no Brasil: consolidação e perspectivas* (Saint Paul 2008)

Miranda MA and Agapito LS, 'Critérios de Validade e Eficiência de Compliance e Impactos na Interpretação da Lavagem de Dinheiro' in Eduardo Saad-Diniz, Luís Augusto Brodt, Henrique Abi-Ackel Torres and Luciano Santos Lopes (eds), *Direito Penal Econômico nas Ciências Criminais* (Vorto 2019)

Mulholland C and Frajhof IZ, 'Inteligência Artificial e a Lei Geral de Proteção de Dados Pessoais: Breves Anotações Sobre o Direito à Explicação Perante a Tomada de Decisões por Meio de Machine Learning' in Ana Frazão and Caitlin Mulholland (eds), *Inteligência Artificial e Direito: Ética, Regulação e Responsabilidade* (Thomson Reuters Brasil 2019)

Oliveira ACC, *Lavagem de dinheiro: responsabilidade pela omissão de informações* (Tirant lo Blanch 2019)

Oliveira JC, Agapito LS and Alencar M, 'O Modelo de "Autorregulação Regulada" e a Teoria da Captura: Obstáculos à Efetividade no Combate à Lavagem de Dinheiro no Brasil' (2017) 10 Quaestio Iuris 365 <https://doi.org/10.12957/rqi.2017.26847> accessed 14 July 2021

Peixoto FH and Silva RZM, *Inteligência Artificial e Direito* (Alteridade Editora 2019)

Price II WN, 'Artificial Intelligence in Health Care: Applications and Legal Issues' (2017) 599 U of Michigan Public Law Research Paper 1 <https://ssrn.com/abstract=3078704> accessed 23 January 2020

Quattrocolo S, *Artificial Intelligence, Computational Modelling and Criminal Proceedings: A Framework for a European Legal Discussion* (Springer 2020)

Rodrigues AM, 'A Questão da Pena e a Decisão do Juiz – entre a Dogmática e o Algoritmo' in Anabela Miranda Rodrigues (ed), *A Inteligência Artificial no Direito Penal* (Almedina 2020)

—— 'Inteligência Artificial no Direito Penal – A Justiça Preditiva entre a Americanização e a Europeização' in Anabela Miranda Rodrigues (ed), *A Inteligência Artificial no Direito Penal* (Almedina 2020)

Saad-Diniz E, 'Fronteras del Normativismo: a Ejemplo de las Funciones de la Información en los Programas de Criminal Compliance' (2013) 108 Revista da Faculdade de Direito da Universidade de São Paulo 415

Santosuosso A and Bottalico B, 'Autonomous Systems and the Law: Why Intelligence Matters' in Eric Hilgendorf and Uwe Seidel (eds) *Robotics and the Law: Legal Issues Arising from Industry 4.0 Technology Programme of the German Federal Ministry for Economic Affairs and Energy* (Nomos 2017)

Sousa SA, '"Não Fui Eu, Foi a Máquina": Teoria do Crime, Responsabilidade e Inteligência Artificial' in Anabela Miranda Rodrigues (ed), *A Inteligência Artificial no Direito Penal* (Almedina 2020)

Tang J and Ai L, 'The System Integration of Anti-Money Laundering Data Reporting and Customer Relationship Management in Commercial Banks' (2013) 16(3) Journal of Money Laundering Control 231 <https://doi.org/10.1108/JMLC-04-2013-0010> accessed 13 July 2021

Vogel B, 'Introduction' in Benjamin Vogel and Jean-Baptiste Maillart (eds), National and International Anti-Money Laundering Law: Developing the Architecture of Criminal Justice, Regulation and Data Protection (Intersentia 2020)

Went P, 'Basel III Accord: Where Do We Go From Here?' (2010) <https://dx.doi.org/10.2139/ssrn.1693622> accessed 13 July 2021

Wimmer M, 'Inteligência Artificial, Algoritmos e o Direito: Um Panorama dos Principais Desafios' in Ana Paula M. Canto de Lima, Carmina Bezerra Hissa and Paloma Mendes Saldanha (eds), *Direito Digital: Debates Contemporâneos* (Thomson Reuters Brasil 2019)

Yapo A and Weiss J, 'Ethical Implications of Bias in Machine Learning' [2018] Proceedings of the 51st Hawaii International Conference on System Sciences 5365

CRIMES INVOLVING AI:
LIABILITY ISSUES AND JURISDICTIONAL CHALLENGES

AI CRIMES AND MISDEMEANORS: DEBATING THE BOUNDARIES OF CRIMINAL LIABILITY AND IMPUTATION

By Anna Moraiti[*]

Abstract

Increasingly pervasive discussions on the future of artificial intelligence (AI) have sparked up questions under criminal law that would probably seem far-fetched or implausible only a few years ago. This paper focuses on the implications that robots and artificial intelligences (RAIs) create under the scope of the general part of substantive criminal law and attempts to delineate a structure and framework with regard to the ascription of criminal liability, in particular negligent criminal liability, the doctrine of causation and rules of imputation. These basic notions inevitably become linked with one another when we attempt to approach the regulation of AI crimes. It is argued that while negligent criminal liability of programmers, producers and/or users of RAIs may be effectively addressed by EU Member States in the following years, even though it may pose a significant challenge as RAIs become more and more autonomous, the proposition of regarding RAIs as satisfying the requirements for criminal liability remains problematic as well as speculative. Nevertheless, it should not be dismissed as entirely inconsequential. On the contrary, it is of great importance for philosophy of criminal law, because it forces us to reconsider many anthropocentric legal presumptions, reflect on the rights of nonhuman agents as well as on the value of non-retributive approaches to crime and punishment.

1 Introduction

Nowadays, the idea that intelligent agents will become an indispensable element in our societies is gradually reaching a wider audience. Some already established areas of artificial intelligence (AI) application in robotics include transport (semi-autonomous or fully autonomous vehicles)[1], healthcare (diagnoses by robotic assistants)[2] and warfare (Autonomous Weapons Systems, AWSs).[3] While certain risks will inevitably be created,

[*] PhD candidate in Criminal Law, University of Luxembourg; EU qualified lawyer. For correspondence: <anna.moraiti@uni.lu>.

[*] Anna Moraiti is a PhD candidate in Criminal Law at the University of Luxembourg and an EU qualified lawyer.

[1] For a thorough analysis on the implications for Criminal Law, see Ivó Coca-Vila, 'Self-Driving Cars in Dilemmatic Situations: An Approach Based on the Theory of Justification in Criminal Law' (2018) 12 Criminal Law and Philosophy 59.

[2] For instance, Hanson Robotics has announced the mass production of intelligent robotic assistants as of late 2021. They will be based on a prototype called Grace, serving *inter alia* as a direct response to the needs in medical staff created by the COVID-19 pandemic; Rebecca Cairns, 'Meet Grace, the ultra-lifelike nurse robot' (CNN, 19 August 2021) <https://edition.cnn.com/2021/08/19/asia/grace-hanson-robotics-android-nurse-hnk-spc-intl/index.html> accessed 20 August 2021.

[3] For a detailed analysis see Daniele Amoroso, *Autonomous Weapons Systems and International Law: A Study on Human-Machine Interactions in Ethically and Legally Sensitive Domains* (Nomos 2020).

their use cannot be avoided as immense social benefits are expected at the same time, through increased work efficiency and reduced human error.[4]

Our imagination for the days ahead and the range of potential new AI applications runs riot while observing humanoid robots perform manipulation and complex physical activities, such as dancing and parkour.[5] The increasing autonomy of intelligent agents means that while performing any given task they are driven by visual perception, are able to interact with their environment, as well as make decisions and learn through trial and error.[6] Early positive results are being challenged through the experimental combination of state-of-the-art algorithmic methods, such as deep reinforcement learning, imitation learning and transfer learning.[7] In other terms, these are 'bottom-up or stochastic' algorithms, which allow a robot to learn from experience and revise its algorithm over time.[8] While it remains doubtful whether intelligent agents will escape the simulation environment any time soon, regulatory prospects should be debated presently so that legal concepts are not caught 'off guard' by any sudden advances in the field. While only 'narrow or weak' AI corresponds to contemporary technological standards, the notion of 'general or strong AI' (ie, intelligent agents that are able to replicate a wide range of human intellectual capacities) merits further research, albeit on a purely theoretical level.[9]

In the following, I will briefly discuss the extent to which the general part of substantive criminal law appears to be affected by the anticipated deployment of robots and artificial intelligences (RAIs) for the execution of tasks that were previously delegated to humans and that could potentially affect fundamental human rights and interests.[10] The distinction that is presupposed throughout this contribution is that between AI-related crimes, whereby RAIs are involved but only human agents appear as perpetrators (analysed un-

[4] Michael Cheng-Tek Tai, 'The Impact of Artificial Intelligence on Human Society and Bioethics' (2020) 32 Tzu Chi Medical Journal 339.

[5] See, eg, footage of the latest skills displayed by Atlas, a humanoid robot developed by Boston Dynamics; Calvin Hennick, 'Leaps, Bounds, and Backflips' (Boston Dynamics blog, 17 August 2021) <http://blog.bostondynamics.com/atlas-leaps-bounds-and-backflips> accessed 20 August 2021.

[6] Tengteng Zhang and Hongwei Mo, 'Reinforcement Learning for Robot Research: A Comprehensive Review and Open Issues' (2021) 18 International Journal of Advanced Robotic Systems 172988142110073.

[7] Jiang Hua and others, 'Learning for a Robot: Deep Reinforcement Learning, Imitation Learning, Transfer Learning' (2021) 21 Sensors 1278.

[8] John Tasioulas, 'First Steps Towards an Ethics of Robots and Artificial Intelligence' (2019) 7 Journal of Practical Ethics 61.

[9] This contribution will not delve into accounts of future agents that might possess intelligence far surpassing human abilities; for this discussion, see Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press 2014) 22, who defines superintelligence as 'any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest.'

[10] This abbreviation, which here serves as a spectrum term, is introduced by Tasioulas (n 8).

der Part 2), and AI-generated crimes, whereby it seems necessary to examine the hypothesis of holding RAIs criminally liable because there is apparently no culpable human agent involved (analysed under Part 3).[11]

Firstly, I will discuss what the case of RAIs entails for the ascription of negligent criminal liability under Greek/German criminal law against programmers, producers and/or users, as well as for the doctrine of causation, with a particular reference to the theory of objective imputation (*objektive Zurechnung*). Next, I will briefly examine the concept of 'electronic legal personhood', which serves as a possible solution for the diffusion of liability amongst multiple actors and as a prerequisite for the potential criminal liability of RAIs. Then, I will discuss the arguments for and against punishing RAIs and assess whether corporate criminal liability could serve as a paradigm. Finally, I will summarise some general regulatory prospects for AI crimes.

2 The Ascription of Negligent Criminal Liability for Programmers, Producers and Users of RAIs

2.1 Objective-subjective negligence and causation issues

To begin with, I assume that cases in which RAIs are used intentionally as a mere instrument for the perpetration of a crime are of no particular doctrinal interest for the study of AI-related crimes.[12] On the contrary, negligence and the standard of the duty of care are difficult to define *especially* when intelligent agents are involved, because the combined use of deep reinforcement learning methods with neural networks means not only that RAIs may function independently of human supervision and control, but also that their responses become oftentimes unpredictable even for their programmers.[13] Likewise, after a harmful outcome has occurred (which could be an involuntary manslaughter or bodily harm) it may be extremely difficult to explain why the agent acted in a particular way instead of another.[14] Therefore, liability gaps are very likely to emerge from the programming, operation or use of RAIs and it remains to be determined how exactly they will be addressed.

In fact, the ascription of negligent criminal liability could be questioned at a level that Greek criminal law assesses prior to that of the subjective element, namely at the level of the causal link between a human act or omission and the harm caused. This problem

[11] This distinction is roughly made by Ryan Abbott, *The Reasonable Robot: Artificial Intelligence and the Law* (Cambridge University Press 2020) 112.
[12] Dafni Lima, 'Could AI Agents Be Held Criminally Liable? Artificial Intelligence and the Challenges for Criminal Law' (2018) 69 South Carolina Law Review 677, 690.
[13] Tasioulas (n 8).
[14] Susanne Beck, 'Intelligent Agents and Criminal Law—Negligence, Diffusion of Liability and Electronic Personhood' (2016) 86 Robotics and Autonomous Systems 138.

becomes exacerbated by the fact that several persons could be regarded as potential perpetrators (programmers, producers and/or users).[15] Users could be regarded as criminally liable especially in cases of open-source software.[16] From an empirical point of view, it becomes apparent that expert evidence will be decisive for the outcome of the proceedings, as judges will struggle to grasp the technological intricacies behind every individual case. From a normative point of view, the theory of objective imputation in particular can be of assistance to legal practitioners, as it allows them to make use of certain 'evaluative criteria' that would inform their judgement in each specific case, because the risk created by any alleged perpetrator may not necessarily be correlated to the outcome actually produced (*Risikozusammenhang*).[17]

More importantly, with regard to crimes of negligence under Greek criminal law, it is widely held that objective negligence constitutes a structural element of the negligent offence that can be evaluated as an objectively dangerous act.[18] Thus the first step would be to ascertain who amongst the involved individuals violated a duty of care by activating risk factors; as mentioned above, the possibility of 'cumulative flaws' by more than one persons would raise issues of joint liability.[19] And, naturally, if no such flaw can be detected in each individual conduct, the ascription of criminal liability has to be excluded.[20]

According to the view presented here objective negligence is inextricably linked to the theory of objective imputation.[21] Furthermore, the outcome produced must be objectively foreseeable; and so, the question arises as to *how specific* foreseeability has to be in the case of robotics.[22] For all intents and purposes, the answer cannot be uniform and will be tailored to each specific application and the intricacies of the case. For instance, if it is

---

[15] In moral philosophy, this phenomenon is sometimes referred to as 'the problem of many hands'; see Mark Coeckelbergh, 'Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability' (2020) 26 Science and Engineering Ethics 2051, 2056.

[16] Abbott (n 11) 131.

[17] The theory of objective imputation (which is *de facto* applied by Greek criminal courts in crimes of negligence) is regarded as supplementary to the *conditio sine qua non* theory ('but for' legal test), whereas I take no position on the question whether it does in fact meet the criteria of a systematic theory; for such an understanding of the theory, see Konstantinos Tsinas, *Law and Logic - The Case of Criminal Law Theory* (Eurasia Publications 2011) 129–130. For a defence of the theory in Greek criminal law doctrine, see Christos Mylonopoulos, *Criminal Law - The General Part I* (PN Sakkoulas 2007) 199–204.

[18] Mylonopoulos (n 17) 302; Maria Kaiafa-Gbandi, 'Artificial Intelligence as a Challenge for Criminal Law: In Search of a New Model of Criminal Liability?', *Digitalisierung, Automatisierung, KI und Recht - Festgabe zum 10-jährigen Bestehen der Forschungsstelle RobotRecht*, vol 20 (Nomos 2020) 313, who dismisses the theory of objective imputation in favour of an approach on causation mainly based on empirical data. The author supports that the latter view is more restrictive in the affirmation of negligent criminal liability and thus more likely to create liability gaps with respect to the case of RAIs.

[19] Kaiafa-Gbandi (n 18) 314.

[20] According to the principle of the division of labour (Arbeitsteilungsprinzip), in the case of convergent activity of several persons, the breach of the duty of care by one of them – which led to the harmful outcome – cannot be borne by the one who complied with the duty of care.

[21] Mylonopoulos (n 17) 305.

[22] Beck (n 14) 139.

demonstrated that the duty of care was breached but the harmful outcome would have occurred even if the individual in question (eg, one of the programmers) had been reasonably cautious, the ascription of negligent criminal responsibility is not possible (*rechtmäßiges Alternativverhalten*, lawful alternative behaviour). The affirmation of subjective negligence (based on the level of skills or knowledge of the person in question) may come to the protection of programmers, producers and/or users of RAIs, unless the latter have refrained from intervening in order to prevent foreseeable harm.[23] In particular, if they actually foresaw the outcome and decided to disregard it, recklessness would be confirmed.[24]

2.2 Practical issues and the defence of permissible dangerous activity

One of the main difficulties for the examination of compliance with the duty of care also stems from the fact that AI production and operation standards are not yet fully developed.[25] Yet this is bound to change very soon at the EU level, especially since in April 2021 the Commission published a fully-fledged regulatory Proposal for trustworthy AI, which classifies AI applications in accordance with a risk-based approach.[26] The latter is not entirely new, as it had already been introduced with the Commission's White Paper on Artificial Intelligence, published approximately a year before.[27] The Proposal sets out obligations for providers of AI systems with respect to post-market monitoring and reporting/investigating on AI-related incidents and provides for the creation of codes of conduct.[28]

In another respect, the doctrinal argument that criminal law should exceptionally not interfere with a permissible dangerous activity, based on the social acceptance of a subsequent 'admissible risk' (*erlaubtes Risiko*), has been repeatedly put forth in academia with reference to the case of RAIs.[29] And rightly so, as the vast majority of EU Member States is welcoming the expected social benefits stemming from AI applications. Nevertheless, it is a fact that as innovation in the field is currently evolving more rapidly than we would have expected some years ago, we cannot settle on any momentary consensus and need to constantly monitor all applications so as to maintain a sensible level of risk

[23] Mylonopoulos (n 17) 322.
[24] Lima (n 12) 692.
[25] Kaiafa-Gbandi (n 18) 316.
[26] Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts COM [2021] 206 final. The Proposal distinguishes between uses of AI that create an unacceptable risk (Title II), a high risk/AI that gives rise to transparency obligations (Titles III and IV respectively) and a low or minimal risk.
[27] White Paper on Artificial Intelligence – A European approach to excellence and trust COM [2020] 65 final.
[28] See titles VIII and IX of the Proposal.
[29] Sabine Gless, Emily Silverman and Thomas Weigend, 'If Robots Cause Harm, Who Is to Blame? Self-Driving Cars and Criminal Liability' (2016) 19 New Criminal Law Review 412, 432.; Kaiafa-Gbandi (n 18) 316, 319ff; Beck (n 14) 141.

at all times. After all, a dangerous activity is rendered permissible precisely through the observance of a set of protective rules.

From a criminal policy point of view, the establishment of a concrete causal link would not be necessary, had crimes of enhanced danger been prescribed for the regulation of such cases. However, social reality has not affirmed that such a development is absolutely necessary at present, and the ancillary nature of criminal law should prevent EU Member States from immediately considering such a solution.[30] In any case, it is very likely that such a prospect would hinder innovation by discouraging programmers and producers of RAIs from taking some arguably necessary risks when developing their products. But even apparently unjustified risks might not be grave enough or in any case best addressed by means of criminal law.[31]

In conclusion, as AI applications have not been tested enough by social reality, the ascription of negligent criminal liability to any of the persons involved remains questionable but wholly plausible and is highly dependent on circumstance. It does not seem that we should immediately consider stricter standards for what it means to be grossly negligent in the case of RAIs. The extension of the question posed turns to the possibility of ascribing criminal liability against the intelligent agent itself, a prospect which is *inter alia* correlated to the granting of 'electronic personhood' to RAIs.[32] Such a legal status would serve as a symbol for the joint efforts of all parties involved in their production; the issue becomes more problematic if we were to accept that it would also be accompanied by the recognition of a certain moral status on the part of RAIs.[33]

3 Criminal Liability of RAIs

In cases where a harmful outcome could only be attributed to an intelligent agent's decision-making and action, it would be tempting to consider whether criminal liability may be affirmed for the agent in question (AI-generated crimes). Such a prospect is, subject to conditions, regarded as possible by many authors.[34] Before delving into this issue, it is important to reflect more closely on the conceptually preceding notion of legal personality and personhood as a source of rights and obligations.

---

[30] Kaiafa-Gbandi (n 18) 319.

[31] Abbott (n 11) 114.

[32] This issue merits extensive research and will only be addressed concisely here. Indicatively, see Denis Franco Silva, 'From Human to Person: Detaching Personhood from Human Nature' in Visa AJ Kurki and Tomasz Pietrzykowski (eds), *Legal Personhood: Animals, Artificial Intelligence and the Unborn*, vol 119 (Springer International Publishing 2017) 113.

[33] Beck (n 14) 141.

[34] See eg, Gabriel Hallevy, *Liability for Crimes Involving Artificial Intelligence Systems* (Springer International Publishing 2015); Ying Hu, 'Robot Criminals' (2019) 52 U MICH J L REFORM 487; Karsten Gaede, *Künstliche Intelligenz-Rechte und Strafen für Roboter?: Plädoyer für eine Regulierung künstlicher Intelligenz jenseits ihrer reinen Anwendung* (Nomos 2019).

3.1 Electronic legal personhood and rights for RAIs

A lively discussion is currently taking place on the question whether RAIs may possess a moral status that would confer upon them rights and responsibilities ('artificial moral agents'), albeit certainly not the full spectrum that humans are entitled to.[35] From the perspective of civil law, the recognition of legal personality would be invaluable for programmers and producers of RAIs, as it would allow intelligent agents to cover any damages caused to third parties by means of their own property and assets.[36] However, expected repercussions for our understanding of autonomy, personal identity and personhood lead us to examine this prospect as a whole very carefully.[37]

The concept of personhood has been historically linked to human ability for reflection and self-consciousness.[38] It is also associated with the notion of agency, ie, the ability to select and pursue proper goals based on conscious beliefs. RAIs as we know them fail with respect to both criteria, as they are incapable of self-awareness and the objectives they currently pursue depend entirely on their programming.[39] Human decision-making is unique in the sense that it may engage in what Hodgson calls 'plausible reasoning', which he perceives as a faculty that provides a role for consciousness and free will.[40] Presently, many experts in the field anticipate that RAIs' decision-making abilities will at best be oriented towards rationality and efficiency and less towards empathy and emotionality, which creates the possibility of them making typically right but morally questionable choices.[41] However, and at least when discussing moral rights, there are some convincing arguments as to why RAIs should not be excluded, most of them based on the supposition that they will evolve in the years to come and will eventually reach a status similar to that of animals.[42]

With respect to moral rights, the most widely discussed approach is the 'meritocratic approach', centered around the question of *direct* moral standing and the properties of RAIs. The decisive element in this respect is sentience, ie, only sentient entities are entitled to moral rights. Provided that RAIs will attain sentience and are – or at least *seem* to

---

[35] Tasioulas (n 8).

[36] Beck (n 14) 142. See eg, in this respect the European Parliament resolution of 20 October 2020 with recommendations to the Commission on a civil liability regime for artificial intelligence 2020/2014 (INL), following the 2017 report on the same issue. The latter asserted that 'the civil liability for damage caused by robots is a crucial issue which also needs to be analysed and addressed at Union level.'

[37] Beck (n 14) 142.

[38] Gless, Silverman and Weigend (n 29) 415.

[39] See Ulfrid Neumann, 'Structure and Regulatory Content of the Principle of Guilt in Criminal Law' [2020] Penal Chronicles 481 ff. The author straightforwardly rejects the idea of rights and sanctions for robots based on their being utterly dependent on programming.

[40] David Hodgson, *Rationality + Consciousness = Free Will* (Oxford University Press 2012).

[41] Susanne Beck, 'Robotics and Criminal Law - Negligence, Diffusion of Liability and Electronic Personhood' in Eric Hilgendorf and Jochen Feldle (eds), *Digitization and the Law* (Nomos 2018).

[42] See for instance Cass R Sunstein and Martha C Nussbaum (eds), *Animal Rights: Current Debates and New Directions* (Oxford University Press 2004).

be – prone to pain and harm, their status as moral patients or receivers could be confirmed.[43] Based on this account, a decision to give moral standing to robots today would be wholly unjustified.[44]

An alternative view, presented by Coeckelbergh and Gunkel, advocates for the recognition of '*indirect* moral standing' for social robots.[45] This is a critical and relational approach to moral standing, based on 'how we relate to entities rather than on their intrinsic properties'.[46] For instance, moral standing should be confirmed in cases where the user has developed feelings of attachment towards the robot or in cases where the latter is part of a human-robot joint action.[47] The basis of this approach is Kantian, and more specifically, it traces back to Kant's argument that humans have indirect duties to animals such as dogs because cruelty towards them is opposed to humans' duty to *themselves*. Such cruelty dulls their shared feeling of suffering and gradually 'uproots a natural disposition that is very serviceable to morality in one's relations with other people'.[48] Therefore, Coeckelbergh suggests that this argument be extended to social robots for reasons of consistency. The precautionary argument that he also puts forth is indeed compelling: since our views on the moral standing of certain entities have evolved in the course of history, we should be cautious when we evaluate the moral standing of other entities.

Even if we were to agree that social robots could be regarded as moral patients, this does not mean that RAIs would be entitled to legal personhood and *legal* rights in a strict sense – granting them might just as well amount to unjustified anthropomorphism at this stage.[49] But if we do accept that they are of ultimate value, they could hold claim-rights and be passive legal persons.[50] This question is paramount for their criminal liability because if monetary sanctions were to be enforced, the RAIs in question would have to be able to own property.[51] Therefore, the most realistic prospect for the near future is that of endowing RAIs with legal personhood for purely commercial and/or compensatory reasons.

---

[43] Steve Torrance, 'Machine Ethics and the Idea of a More-Than-Human Moral World' in Michael Anderson and Susan Leigh Anderson (eds), *Machine Ethics* (Cambridge University Press 2011) 117.
[44] Deborah G Johnson, 'Computer Systems: Moral Entities but Not Moral Agents' (2006) 8 Ethics and Information Technology 195.
[45] Mark Coeckelbergh, 'Should We Treat Teddy Bear 2.0 as a Kantian Dog? Four Arguments for the Indirect Moral Standing of Personal Social Robots, with Implications for Thinking About Animals and Humans' [2020] Minds and Machines. Coeckelbergh clarifies though that his arguments do not oppose the reasoning based on direct moral standing, but rather complement it.
[46] ibid.
[47] ibid.
[48] Immanuel Kant, *The Metaphysics of Morals* (Cambridge University Press 1999) 564.
[49] Abbott (n 10) 127-128.
[50] Visa AJ Kurki, 'The Legal Personhood of Artificial Intelligences', *A Theory of Legal Personhood* (Oxford University Press 2019) 178.
[51] ibid 181.

3.2 Punishment for RAIs

The question whether RAIs could ever be considered as directly satisfying the requirements for criminal liability is highly controversial. It is a difficult proposition to establish, as at a starting level it seems incompatible with basic notions of criminal law, most notably the principle of culpability which is based on the notion of free will.[52] More specifically, personhood in criminal law is linked to the ability to distinguish right from wrong and freely choose the former, as only someone who could choose otherwise and can be blamed for committing a crime.[53]

And this is precisely the point of departure from the comparison that was previously made with animals: when it comes to criminal liability, the analogy that some authors attempt is that to legal persons (corporate criminal liability), as it is the sole example of criminal liability being extended to artificial agents.[54] Potential sanctions that have been considered are confiscation or destruction of harmful RAIs or the imposition of fines on assets owned by them. Ultimately, though, RAIs differ from both animals and legal persons in one fundamental respect: not only do they challenge some of our basic legal concepts, but they do so while leaving all possibilities open as to their future legal status. We find ourselves in lack of crucial factual data.

It might well be the case that a criminal act can solely be traced back to the action of an intelligent agent; perhaps the most prominent example is a case involving an online shopping bot, the Random Darknet Shopper.[55] In the context of an art exhibition in Zurich, the Random Darknet Shopper was programmed to make random purchases on the Deep Web every week; all of the selected items were put on display by the artists involved, with ecstasy pills being one of them. Had the perpetrator been a human, prosecution would very likely ensue, but the public prosecutor decided to drop charges stating that 'the overweighing interest in the questions raised by the artwork…justify the exhibition of the drugs as artefacts, even if the exhibition does hold a small risk of endangerment of third parties through the drugs exhibited.'[56] The artists themselves escaped being prosecuted only because of the high level of protection afforded to artistic creations by Swiss legislature.

This example illustrates that RAIs can functionally commit crimes. But for the moment this means that their actions can *causally* lead to the commission of crimes – and nothing more.[57] Even if we were to accept that RAIs may be responsible for acts that convey social meaning, it is even more questionable whether they could fulfil the subjective element

---

[52] Francesca Lagioia and Giovanni Sartor, 'AI Systems Under Criminal Law: A Legal Analysis and a Regulatory Perspective' (2020) 33 Philosophy & Technology 433.
[53] Gless, Silverman and Weigend (n 29) 419.
[54] Lima (n 12) 687.
[55] Lagioia and Sartor (n 52) 452.
[56] ibid.
[57] John Danaher, 'Robots, Law and the Retribution Gap' (2016) 18 Ethics and Information Technology 299, 301.

(mens rea).[58] This would require a high level of sophistication, meaning that the agent should be able to discern the moral and legal implications of different choices and exercise judgement accordingly, and this would be a prerequisite for all modes of criminal liability (direct perpetration, joint criminal enterprise, indirect perpetration, command responsibility).[59]

Hallevy, one of the most prominent advocates of punishment for RAIs, presents corporate criminal liability as a paradigm. He argues that RAIs, like corporations, participate increasingly in a great spectrum of human activities and ultimately there is no convincing legal argument for the two cases to be treated differently.[60] Hu and Gaede are also arguing in favour of imposing sanctions on strong AI systems.[61] Hu in particular highlights that an advanced agent of the future may be equipped with 'moral algorithms' that would continuously evolve based on its interactions. Therefore, 'each person who is authorized to influence the robot's moral algorithms should also be required to contribute financially.'[62] If the law already punishes artificial persons that cannot strictly speaking act or possess mental states, why should such a development be precluded, especially since it does not directly imply any moral rights on the part of RAIs?

The answer cannot but be influenced by the ancillary nature of criminal law.[63] While we are not convinced that criminal law is strictly necessary for the case at hand, we cannot convincingly argue in this direction. The next pertinent question is: would any of the functions of sentencing (theories of punishment) be fulfilled by punishing RAIs? To begin with, would limiting an intelligent agent's functional autonomy properly count as 'incapacitation'? Would punishment lead to reform? Deterrence seems also unlikely, as RAIs cannot discern the similarities between their choices and those of other agents that have been punished.[64] Deterrence would work only for their programmers, producers and/or users, but as argued above, this might not be necessary as of yet.[65] As far as the compensatory function for victims is concerned, Mulligan has argued that punishing RAIs may be needed in order to 'create psychological satisfaction in those whom robots harm'.[66] But then again, this argument presupposes that RAIs are being perceived as

[58] For an analysis of the objective and subjective elements of criminal liability in German criminal law, see Markus Dirk Dubber and Tatjana Hörnle, *Criminal Law: A Comparative Approach* (First edition, Oxford University Press 2014) 194ff and 241ff.
[59] In this direction see eg Lima (n 12) 690.
[60] Gabriel Hallevy, 'The Criminal Liability of Artificial Intelligence Entities - from Science Fiction to Legal Social Control' 4 Akron Intellectual Property Journal 171, 191.
[61] Hu (n 34); Gaede (n 34) 66.
[62] Hu (n 34) 530.
[63] In this direction, see Abbott (n 10) 115ff.
[64] Peter Asaro, 'A Body to Kick, but Still No Soul to Damn: Legal Perspectives on Robotics' in Patrick Lin, Keith Abney and George A Bekey (eds), *Robot Ethics: The Ethical and Social Implications of Robotics* (MIT Press 2012) 181.
[65] See Section 2.2.
[66] Christina Mulligan, 'Revenge Against Robots' (2018) 69 South Carolina Law Review 579, 580.

blameworthy.[67] Finally, the retributivist account of punishment would not be fitting considering RAIs' inability to engage in culpable wrongdoing.

Drawing from corporate criminal liability, the doctrinal tool that merits consideration is *respondeat superior*, which allows mental states of an agent to be imputed to the corporation provided they were acting within the scope of their employment.[68] But under closer examination, it becomes apparent that it would be more difficult to identify culpable human actors in the case of RAIs than in corporate structures.[69] And when this is in fact the case the latter could be held accountable under the conditions delineated in the first part of this contribution. Ultimately, it seems we cannot avoid attaching punishment for AI-generated crimes to what Abbott calls 'a functional analogue of a standard mens rea'.[70] This means that we should be able to discern that an intelligent agent – quite literally here – is purposefully directing their conduct towards a certain criminal outcome.[71] As he puts it, '[i]f an AI is monitoring conditions around it to identify ways to make the outcome (harm a bystander) more likely, and it is then disposed to make behavioral adjustments to make the outcome more likely relative to its background probability of occurring […] the AI could be said to have the purpose of causing that outcome.'[72]

In conclusion, unless we can confirm that AI-generated crimes pose a serious social threat so as to raise the standards of the duty of care or that criminal sanctions would be effective when dealing with artificial agents that could demonstrate some form of intent, which is similar and relatable to the human condition, it may be that criminal law would not be the appropriate means to address such crimes. In the years to come, we should reflect on *whether* and, if so, *how* we can refine existent doctrinal constructions so as to address deviant forms of conduct stemming from sentient and intelligent beings that, even though they do not quite look at the world from our eyes, remain relatable and useful for human experience. And this would be for our benefit, as this hypothesis could put to the test the very notions of punishment and blame.

4 Some Thoughts on the Regulatory Prospects of AI Crimes

The first issue to resolve would be whether to enact new negligence AI-related crimes against programmers, producers and/or users of RAIs who could activate risk factors for serious harm. As stated above, this is a wholly plausible possibility for the future, and it could be that the establishment of clear rules would help all individuals involved to become fully aware of their duties and exercise caution. But the same prospect could have the reverse effect and it should be balanced to the well-known danger of hindering technological progress. I would argue that it should not be considered any time soon. EU

[67] Abbott (n 11) 117.
[68] Abigail H Lipman, 'Corporate Criminal Liability' (2009) 46 American Criminal Law Review 359, 360.
[69] Abbott (n 11) 119.
[70] ibid 122.
[71] Michael Bratman, *Intention, Plans, and Practical Reason* (Center for the Study of Language and Information 1999) 141.
[72] Abbott (n 11) 122.

Member States should refrain from such an initiative *unless* it becomes factually evident that AI-related crimes constitute a real threat and heightened duties of care should be established.[73] With respect to present-day RAIs, criminal law as it stands could still address many crimes of negligence and in any case, the possibility of some admissible risk – in areas where great benefits are expected, such as with the deployment of autonomous vehicles – should also be taken into account.[74]

As for AI-generated crimes, punishing RAIs and the prospect of having recourse to a second legal fiction should also be out of the picture for the future, due to the difficulties that have already been outlined. However, the example of the Random Darknet Shopper indicates that there has to be a legal response to the possibility of serious harm occurring even though no particular individual can be held accountable.[75] Therefore, the prospect of expanding civil liability for persons involved in the deployment of RAIs or the creation of an insurance fund maintained by programmers, producers or users would be sensible measures so as to ensure compensation for victims of such crimes.[76]

Admittedly though, Danaher rightly argues that such 'compensation gaps' can be easily addressed and that instead, our focus should shift towards the 'retribution gap' that emerges as RAIs become more and more autonomous and common in social activities.[77] Harmful outcomes are bound to occur with frequency and placing blame on programmers, producers or RAIs themselves may seem simply wrong. The principle of culpability is being questioned with regard to both scales of AI crimes. He argues that one of the consequences could be an erosion of rule of law principles, but I choose to place emphasis on a different possibility that he foresees: this development could spark a wider debate on the value of our current criminal justice system and allow non-retributive accounts to crime and punishment to be brought to the fore.

5 Concluding Remarks

Swift advances in the development of AI applications and extensive research on the intersection of AI and law are bound to reshape many fundamental legal concepts, not least in the area of criminal law, but for the moment the delineation of any definitive legal measures seems impossible. The imposition of new negligence crimes for programmers, producers and/or users of RAIs does not seem necessary as of yet. In any case, legal responses based on criminal law should remain an exception and a civil liability regime should be tested before any measures that could seriously hinder innovation are considered.

Moreover, the prospect of punishing RAIs – even though it remains simply a possibility for the remote future – should be discussed more widely, at least as a source of insight

[73] ibid 130.
[74] Gless, Silverman and Weigend (n 29) 432.
[75] Abbott (n 11) 129.
[76] ibid 132.
[77] Danaher (n 57) 308.

for the structure and purposes served by our criminal justice system. Comparisons made to the model of corporate criminal liability should be neither disregarded nor overestimated but be further elaborated and examined in light of new factual data.

As for intelligent agents of the future, their conceptualisation may contribute to the establishment of principles for a harmonious coexistence between human and nonhuman (natural or artificial) moral agents. The disruptive nature of RAIs poses a number of fascinating questions on the nature of personhood and criminal agency. In the end, would any truly intelligent but emotionally immature artificial agent essentially differ from a child, slowly building upon its perception of the world of adults, or from an individual born with a condition related to reduced affect display?[78] It is very unlikely that we will live to witness the comparison spring to life. But this intuitive thought should be further explored as a possible claim for the sake of inclusion in the following years.

References

Abbott R, *The Reasonable Robot: Artificial Intelligence and the Law* (1st edn, Cambridge University Press 2020)

Amoroso D, *Autonomous Weapons Systems and International Law: A Study on Human-Machine Interactions in Ethically and Legally Sensitive Domains* (Nomos 2020)

Asaro P, 'A Body to Kick, but Still No Soul to Damn: Legal Perspectives on Robotics' in Patrick Lin, Keith Abney and George A Bekey (eds), *Robot Ethics: the Ethical and Social Implications of Robotics* (MIT Press 2012)

Beck S, 'Intelligent Agents and Criminal Law—Negligence, Diffusion of Liability and Electronic Personhood' (2016) 86 Robotics and Autonomous Systems 138

—— 'Robotics and Criminal Law - Negligence, Diffusion of Liability and Electronic Personhood' in Eric Hilgendorf and Jochen Feldle (eds), *Digitization and the Law* (Nomos 2018)

[78] A literary reference for such a reflection appears in Philip K. Dick's 1968 novel *Do Androids Dream of Electric Sheep?*, which in turn inspired the 1982 sci-fi film, *Blade Runner*. As the plot unfolds the protagonist, bounty hunter Rick Deckard, gradually begins to question the strict distinction maintained between androids and humans in a dystopian, future account of American society. The contrast of his opinions at the beginning and towards the end of the book is revelatory of a moral review: 'He had wondered, as had most people at one time or another, precisely why an android bounced helplessly about when confronted by an empathy measuring test … For one thing, the empathic faculty probably required an unimpaired group instinct; a solitary organism, such as a spider, would have no use for it … It would make him conscious of the desire to live on the part of his prey… ultimately, the empathic gift blurred the boundaries between hunter and victim, between the successful and the defeated … Evidently the humanoid robot constituted a solitary predator. Rick liked to think of them that way; it made his job palatable.' (pp 28-29); cf: 'Empathy toward an artificial construct? he asked himself. Something that only pretends to be alive? But Luba Luft had seemed *genuinely* alive; it had not worn the aspect of a simulation.' (p 118).

Bostrom N, *Superintelligence: Paths, Dangers, Strategies* (First edition, Oxford University Press 2014)

Bratman M, *Intention, Plans, and Practical Reason* (Center for the Study of Language and Information 1999)

Coca-Vila I, 'Self-Driving Cars in Dilemmatic Situations: An Approach Based on the Theory of Justification in Criminal Law' (2018) 12 Criminal Law and Philosophy 59

Coeckelbergh M, 'Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability' (2020) 26 Science and Engineering Ethics 2051

—— 'Should We Treat Teddy Bear 2.0 as a Kantian Dog? Four Arguments for the Indirect Moral Standing of Personal Social Robots, with Implications for Thinking About Animals and Humans' [2020] Minds and Machines <http://link.springer.com/10.1007/s11023-020-09554-3>

Danaher J, 'Robots, Law and the Retribution Gap' (2016) 18 Ethics and Information Technology 299

Dick P K, *Do Androids Dream of Electric Sheep?* (Ballantine Books 2008)

Dubber MD and Hörnle T, *Criminal Law: A Comparative Approach* (First edition, Oxford University Press 2014)

Gaede K, *Künstliche Intelligenz-Rechte und Strafen für Roboter?: Plädoyer für eine Regulierung künstlicher Intelligenz jenseits ihrer reinen Anwendung* (Nomos 2019)

Gless S, Silverman E and Weigend T, 'If Robots Cause Harm, Who Is to Blame? Self-Driving Cars and Criminal Liability' (2016) 19 New Criminal Law Review 412

Hallevy G, *Liability for Crimes Involving Artificial Intelligence Systems* (Springer International Publishing 2015) <http://link.springer.com/10.1007/978-3-319-10124-8>

—— 'The Criminal Liability of Artificial Intelligence Entities - from Science Fiction to Legal Social Control' 4 Akron Intellectual Property Journal 171

Hodgson D, *Rationality + Consciousness = Free Will* (Oxford University Press 2012)

Hu Y, 'Robot Criminals' (2019) 52 U MICH J L REFORM 487

Hua J and others, 'Learning for a Robot: Deep Reinforcement Learning, Imitation Learning, Transfer Learning' (2021) 21 Sensors 1278

Johnson DG, 'Computer Systems: Moral Entities but Not Moral Agents' (2006) 8 Ethics and Information Technology 195

Kaiafa-Gbandi M, 'Artificial Intelligence as a Challenge for Criminal Law: In Search of a New Model of Criminal Liability?', *Digitalisierung, Automatisierung, KI und Recht - Festgabe zum 10-jährigen Bestehen der Forschungsstelle RobotRecht*, vol 20 (Nomos 2020)

Kant I, *The Metaphysics of Morals* (Cambridge University Press 1999)

Kurki VAJ, 'The Legal Personhood of Artificial Intelligences', *A Theory of Legal Personhood* (Oxford University Press 2019)

Lagioia F and Sartor G, 'AI Systems Under Criminal Law: A Legal Analysis and a Regulatory Perspective' (2020) 33 Philosophy & Technology 433

Lima D, 'Could AI Agents Be Held Criminally Liable? Artificial Intelligence and the Challenges for Criminal Law' (2018) 69 South Carolina Law Review 677

Lipman AH, 'Corporate Criminal Liability' (2009) 46 American Criminal Law Review 359

Mulligan C, 'Revenge Against Robots' (2018) 69 South Carolina Law Review 579

Mylonopoulos C, *Criminal Law - The General Part I* (PN Sakkoulas 2007)

Neumann U, 'Structure and Regulatory Content of the Principle of Guilt in Criminal Law' [2020] Penal Chronicles 481

Silva DF, 'From Human to Person: Detaching Personhood from Human Nature' in Visa AJ Kurki and Tomasz Pietrzykowski (eds), *Legal Personhood: Animals, Artificial Intelligence and the Unborn*, vol 119 (Springer International Publishing 2017) <http://link.springer.com/10.1007/978-3-319-53462-6_8>

Sunstein CR and Nussbaum MC (eds), *Animal Rights: Current Debates and New Directions* (Oxford University Press 2004)

Tai M-T, 'The Impact of Artificial Intelligence on Human Society and Bioethics' (2020) 32 Tzu Chi Medical Journal 339

Tasioulas J, 'First Steps Towards an Ethics of Robots and Artificial Intelligence' (2019) 7 61

Torrance S, 'Machine Ethics and the Idea of a More-Than-Human Moral World' in Michael Anderson and Susan Leigh Anderson (eds), *Machine Ethics* (Cambridge University Press 2011) <https://www.cambridge.org/core/product/identifier/CBO9780511978036A019/type/book_part>

Tsinas K, *Law and Logic - The Case of Criminal Law Theory* (Eurasia Publications 2011)

Zhang T and Mo H, 'Reinforcement Learning for Robot Research: A Comprehensive Review and Open Issues' (2021) 18 International Journal of Advanced Robotic Systems 172988142110073

AI AND CRIMINAL LAW:
THE MYTH OF 'CONTROL' IN A DATA-DRIVEN SOCIETY

By Beatrice Panattoni[*]

Abstract

Artificial intelligence (AI) systems, like all other technologies, are not mere instruments but processes to be developed. However, unlike all other technologies, their autonomous functioning allows them to have a stronger active contribution in the interaction processes with the users, raising new challenges for the law.
Since AI systems may be involved in different ways in the commission of a crime, in the future, it is likely that existing offences will have to be adapted or new AI-based crimes created. One of the main issues that the autonomy of AI systems poses to criminal law is who is to be held criminally responsible in case of harmful events caused by their emergent behaviors. In these cases, a responsibility gap could follow. The artificial agent cannot be held directly responsible and the human agent, who has no full control over the system's autonomous functioning, cannot always be criminally reprovable for not exercising the duty to act required to him. According to the new regulatory framework recently proposed by the European Institutions (the so-called Artificial Intelligence Act), the paper aims to describe the possible and future criminal policies that will allow avoiding a responsibility gap.

1 The Passage from Technological Automation to Artificial Autonomy

The impact of digital technology on perception and understanding of the realities surrounding us started already with Information and Communication Technology (ICT), environmental forces, and not just tools, that changed radically and rapidly our social interactions. When facing the processes related to the digital revolution, a brief premise must be made. Since we understand our reality with concepts, when it changes quickly, we found ourselves conceptually wrong-footed, so that «our current conceptual toolbox is no longer fitted to address new ICT-related challenges».[1] Therefore, in describing and analyzing some of the impacts of the innovations related to technological progress and discoveries, we often resort known, but not well-suited concepts, that could not be able to capture the nature and implications of our *onlife* experiences[2]. Two features of the process of transformation that started with ICT are the progressive loss of control and complexity. According to their level of sophistication, artifacts can, to a greater or lesser extent, autonomously modify their states, ceasing to be mere instruments that execute human instructions and escaping human control. Further, not only they are created through a reality of network, but in turn also create, through their functioning, networks of socio-technical interactions. One of the effects of this action of mediation exercised by digital technologies, which increases with the development of the automation of technological

[*] PhD Student in Criminal Law, University of Verona. For correspondence: <beatrice.panattoni@univr.it>.
[1] Luciano Floridi (ed), *The onlife manifesto. Being human in a hyperconnected era* (Springer 2014).
[2] ibid.

artifacts, is that it widens the *distance* between human actions and their consequences.[3] It is not only a distance related to the 'dematerialization' of our activities, made possible by the cyberspace, where we can operate remotely, but it is also a 'material'[4] distance, filled by the autonomation, and now the autonomy, that defines new digital technologies.

Artificial intelligence (AI) applications are part of this prolonged transformation, and they amplify its features and effects since the processes related to digital technologies are shifting from being automated to being autonomous. With artificial agents, the distance is not only between the human action and its consequences, but also between the state of mind of the individual, who intends or knows what is the object of a particular behavior, and the concrete execution of the tasks that allow to materially reach the intended purpose: one may refer to an «alienation of responsibility».[5] In other words, AI systems «have a unique capacity to splinter a criminal act, where a human manifest (sic) the *mens rea* and the robot commits the *actus reus*».[6] Criminal law is not without conceptual tools in facing these kinds of problems: similar issues (although not the same ones) can be found in cases of harms related to private-business activities of corporate organizations, as within complex organizations as well, the decision-making process is fragmented (more than alienated), and a specific harmful result cannot be traced back to the responsibility of a single individual. Legal systems found a solution to this scenario in corporate criminal liability laws, which can be – as it will be analyzed further in the next paragraphs – a strategic model in the contrast and prevention of AI Crimes.

However, the study of legal implications of AI systems brings forward peculiar and new issues that need to be addressed. Specifically, what distinguishes the debate around AI systems lies in their attribute of 'autonomy.' This concept has a general meaning and must be distinguished from related concepts such as adaptability and interactivity. 'Autonomy' is the ability of the system to perform the function for which it was programmed by modifying its own state or internal properties and by controlling its 'actions' without any human intervention.

One of the most significant implications of AI applications that we must keep in mind is then 'emergence.' 'Emergence'[7] is a stated goal of robotics and AI, and it can be understood as a different way of referring to AI autonomy. This feature can lead the system, thanks to its post-design experiences, to elaborate solutions to the goal assigned to it that the operator wouldn't have thought about. On the one hand, AI systems can collect and interpret data in an unpredictable way and, on the other hand, they are characterized by

[3] Mirielle Hildebrandt, 'Criminal Law and Technology in a Data-Driven society', in Markus Dubber, Tatjana Hörnle (eds), *The Oxford Handbook of Criminal Law* (Oxford 2014).
[4] Emphasis added.
[5] Carla Bagnoli, *Teoria della responsabilità* (Il Mulino 2019) 77.
[6] Amanda Mcallister, 'Stranger than science fiction: The rise of AI interrogation in the dawn of autonomous robots and the need for an additional protocol to the UN convention against torture' (2017) 101 Minnesota Law Review 2572.
[7] Ryan Calo, 'Robotics and the Lessons of Cyberlaw' (2016) 103 California Law Review 513, 538.

an epistemology problem: they can incur measurement errors due to their lack of semantic knowledge.[8] The process of 'translation' of our multiform reality into electrical signals can only be 'approximative', leaving an inevitable margin of error, and leading the artificial agent to make significant mistakes that a human agent would find very easy to avoid (for instance in image recognition and classification). Therefore, since AI systems are built to be 'unpredictable by design',[9] they create a predictability gap that needs to be addressed.

In questioning the traditional relationship between agent and instrument, the role of the operator changes in the processes in which these new artifacts are used. The human agent is gradually removed from the decision-making processes involving an artificial agent. The different degrees of autonomy that characterize AI systems are a fundamental element in the development of the different models of responsibility of the operators; the more the level of autonomy increases, the more difficult it becomes to allocate forms of responsibility, which have to face what has been defined as a 'control dilemma'[10]. The different degrees of autonomy of the artificial agent and the different positions that the operators may have (who can be in command or in, on, out of the loop[11]) have a twofold consequence on the configurable models of criminal responsibility. In the first place, corresponding to a more or less-marked delegation to AI systems in the execution of the various tasks, charges of responsibility for omissive rather than commissive conduct will likely be more relevant. Already with semi-autonomous systems (taking self-driving cars as a model, this would be levels up to 2 or 3), it will be less frequent to fall within the category of commissive crimes, in which the user directly contributes to the decision-making process that leads to the harmful outcome. Secondly, the more autonomous an AI system is, the more relevant will become forms of responsibility of the human agents involved in the processes not so much of use, but of design, development, and production, to avoid any form of excessive responsibility of the user, who could risk becoming in such cases a scapegoat to solve the difficulties of reconstruction of the causal chain. However, responsibility for commission by omission implies a duty to act usually based on a position of control. And here lies one of the main challenges that AI practices pose to criminal law categories. One of the main consequences of the emergence of AI is that it challenges responsibility models based on the human agent's control role. Furthermore, in case of harms related to emergent behaviors of AI systems, the criminal offense

[8] William D. Smart, Cindy M. Grimm, Woodrow Hartzog, 'An education theory of fault for autonomous systems' (2021) 2 Notre Dame Journal on Emerging Technologies 33.
[9] Ryan Calo, 'Robotics and the Lessons of Cyberlaw' (n 6) 513.
[10] Eric Hilgendorf, 'Automated Driving and the Law', in Hilgendorf E., Seidel U. (eds), *Robotics, Autonomics and the Law* (Nomos 2017).
[11] Paul Scharre, Michael Horowitz, 'An Introduction to Autonomy in Weapon Systems', Center for a New American Security Working Paper 2015.

can be characterized, as suggested by Alex Sarch and Ryan Abbott, by its potential *irreducibility*. In their words: 'it may be difficult to reduce Al crime to an individual due to Al autonomy, complexity, or lack of explainability'.[12]

The challenges outlined above create what has been named as a 'responsibility gap', a situation where the unpredictability of the artificial agent and the distance between the outputs of its functioning and the operator determine that «the traditional ways of attributing responsibility are no longer compatible with our feeling of justice and the moral preconditions of society, since no-one has sufficient control over the actions of the machine, to be able to take responsibility».[13]

2 AI and Criminal Law: AI Crimes

Despite its novelty, the study of the challenges posed to criminal law by the new era of the digital revolution set in motion by AI technologies has for some time now been on the attention of scholars[14] and is destined to become increasingly important given the rapid development of these technologies. Since AI applications have already and will make it possible to perform new activities and to reach new objectives, it is likely that new ways of committing existing criminal offenses will begin to occur and that also new forms of crimes will require the intervention of the legislator in criminalizing them.

There have already been several relevant cases. Some of them are the result of experiments organized by researchers,[15] others are real cases where harmful events are linked to the first applications of such systems (for instance, road accidents occurred in the United States involving a driverless car;[16] or the spreading of hate messages operated by social bots).[17] Many scholars have started to suggest that we will soon need to identify a specific group of crimes, that we may call AI crimes.[18] In trying to organize this subject, we may start thinking of a classification. In the light of the recent character of this field

[12] Ryan Abbott, Alex Sarch, 'Punishing Artificial Intelligence: Legal Fiction or Science Fiction' (2019) 53 UC Davis Law Review 325, 334.

[13] Andreas Matthias, *Automaten als Träger von Rechten. Plädoyer für eine Gesetzänderung* (dissertation, Humboldt Universität 2007) 22.

[14] Thomas C. King, Nikita Aggarwal, Mariarosaria Taddeo, Luciano Floridi, 'Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions' (2020) 26 Science and Engineering Ethics 89; Thomas King, 'Projecting AI-Crime: A Review of Plausible Threats', in Carl Öhman and David Watson (eds), *The 2018 Yearbook of the Digital Ethics Lab* (Springer 2019).

[15] John Seymour, Philip Tully, 'Weaponizing Data Science for Social Engineering: Automated E2E Spear Phising on Twitter' (2016) <https://www.blackhat.com/docs/us-16/materials/us-16-Seymour-Tully-Weaponizing-Data-Science-For-Social-Engineering-Automated-E2E-Spear-Phishing-On-Twitter-wp.pdf> accessed 2 July 2021.

[16] Bryan Pietsch, '2 Killed in Driverless Tesla Car Crash, Officials Say' *New York Times* (New York, 18 April 2021).

[17] Gina Neff, Peter Nagy, 'Talking to Bots: Symbiotic Agency and the Case of Tay' (2016) International Journal of Communication 10; Madeline Lamo, Ryan Calo, 'Regulating Bot Speech' (2019) U.C.L.A. Law Review 988.

[18] Thomas C. King, Nikita Aggarwal, Mariarosaria Taddeo, Luciano Floridi, 'Artificial Intelligence Crime' (n 13); Ryan Abbott, Alex Sarch, 'Punishing Artificial Intelligence' (n 11).

of research, we could resort to a technically oriented classification. This choice was made also in the first studies of computer crimes, when new criminal behaviors related to computer uses were beginning to rise.[19] Given that AI crime will be, as happened with computer and cybercrimes, a broad phenomenon that will brace different kinds of criminal behaviors and harms, it is suggestable to use not a substantial, but a functional definition, keeping it, therefore, a fluid concept. We can suggest dividing these new future crimes into three different groups: (i) 'criminal AI', cases where the AI system is used or created by a criminal agent as the means to commit the criminal offense; (ii) abuses against AI, cases where the AI system is the 'object' against which the criminal offense is committed; (iii) crimes committed 'directly' by AI systems, without the criminal intent of any operator that stands 'behind' the system for the specific harm caused.

After a brief description of the first two groups, the paper will focus on the analysis of the last one, where the challenging effects that AI systems have on criminal responsibility models emerge more clearly.

2.1 Malicious uses of AI

In their capacity to be tools for human agents, AI systems make it possible to perform new conducts and pursue new objectives. Thus, the new potentialities unleashed using artificial agents can facilitate the development of new ways for committing existing offenses or the appearance of new criminal behaviors. Among scholars,[20] some proposals are already aimed at criminalizing the creation of 'criminal AI.' These cases could include, for example, the programming of AI systems to carry out cyber-attacks[21] or to create unlawful deepfakes, such as those with sexually explicit content created without the consent of the person portrayed, which are often used by criminals for extortionary purposes.[22] To date, we have few reports on the subject, namely the United Nations Interregional Crime and Justice Research Institute (UNICRI) and Europol's report, 'Malicious Uses (means) and Abuses (object) of Artificial Intelligence.'[23] According to the latter report, the present and future threats of malicious uses of AI include, besides AI as a weapon for cyberattack at large scales, 'human impersonation,' that uses social bot or speech synthesis systems that learn to imitate individuals' voices, and 'criminal robot[s],' which are drones used to deliver illegal drugs or to carry out terroristic attacks.

[19] Donn Parker, *Crime by Computer* (1st edn, Cengage GALE 1976) 12.
[20] Ryan Abbott, Alex Sarch, 'Punishing Artificial Intelligence' (n 11).
[21] Sadie Creese, 'The threat from AI', in Dennis J. Baker, Paul H. Robinson (eds), *Artificial Intelligence and the Law. Cybercrime and Criminal Liability* (Routledge 2021) 211.
[22] Keith J. Hayward, Matthijs M Maas, 'Artificial intelligence and crime: A primer for criminologists' (2020) Crime Media Culture 1.
[23] United Nations Interregional Crime and Justice Research Institute and Europol, *Malicious Uses (means) and Abuses (object) of Artificial Intelligence* (Trend Micro Research 2020).

2.2 Abuses against AI

AI applications, which are spreading in every sector of daily life (eg Internet of Things and smart homes), can also be considered as new attack surfaces. Thus, criminal behaviors against such systems, aimed at affecting, damaging, or manipulating their functioning, will become increasingly frequent and dangerous. The harms in this new field can include, for instance, personal data breaches, sabotage of digital systems, and loss of access to digital services.[24] Moreover, new threats will have to be dealt with, such as 'tricking' artificial agents by polluting the system datasets. Researchers have shown, for instance, that by placing stickers on a traffic signal it is possible to make a self-driving car ignore a speed limit; that by sending hidden voice commands to voice assistants such as Alexa or Siri it is possible to make them compile certain phone numbers or open certain websites; that by applying makeup in a certain way, it is possible to escape facial recognition cameras. Abuses against Image Recognition Systems can also be found as present and future threats in the Report of UNICRI.

Faced with these possibilities, it will be necessary to verify whether and to what extent existing cybercrimes apply to these cases, and which gaps in the legislation should be filled. Indeed, some US authors[25] have already begun to wonder whether tricking a robot, a question that could also be extended to algorithmic agents not implemented in cyber-physical systems, can be considered a case of hacking under the Computer Fraud and Abuse Act of 1986.[26] It seems that we are at a similar time as the one that characterized the elaboration and introduction of computer (and then cyber) crimes, probably following a similar course.[27]

2.3 Harms committed directly by AI systems

Given their capacity to perform the tasks assigned with a certain degree of autonomy, some AI systems can also be the direct 'executors' of criminal behaviors. The most evident case regards the application of AI technologies in the field of robotics, for instance, self-driving cars and surgical robots, where physical harms may be the consequence of the interaction with users. However, the range of this third group could include many other hypotheses that already occurred, such as cases of market manipulation by high-frequency traders or of diffusion of illegal contents online by social bots.

---

[24] Sadie Creese, 'The threat from AI' (n 20).

[25] Ryan Calo, Ivan Evtimov, Earlence Fernandes, Tadayoshi Kohno, David O'Hair, 'Is Tricking a Robot Hacking?' (2019) University of Washington School of Law Research 1, 5; Sara Sun Beale, Peter Berris, 'Hacking the Internet of Things: Vulnerabilities, Dangers, and Legal Responses' (2018) 16 Duke Law & Technology Review 161.

[26] Computer Fraud and Abuse Act (CFAA), 18 U.S.C.A. § 1030 (2008).

[27] Francesca Lagioia, Giovanni Sartor, 'AI systems under Criminal Law: a Legal Analysis and a Regulatory Perspective' (2019) Philosophy & Technology 1, 33; Ugo Pagallo, Serena Quattrocolo, 'The impact of AI on criminal law, and its twofold procedures', in Woodrow Barfield and Ugo Pagallo (eds), *Research Handbook on the Law of Artificial Intelligence* (Edward Elgar Pub Cheltenham 2018) 408.

In facing this scenario, some scholars suggested drafting forms of direct criminal responsibility of AI systems.[28] However, it is a thesis that finds place only in the theoretical elaborations of some authors and does not seem to be considered as a valid alternative, being excluded by most scholars.[29] Considering that artificial agents cannot be qualified as 'free' moral agents capable of choosing whether to act against the law or not, it does not seem possible to attribute individual criminal responsibility to these entities, which, although may engage in material legally relevant conducts, cannot, in any case, be considered guilty of them.[30] It thus seems preferable to avoid dogmatic elaborations that allow considering artificial entities as subjects of criminal law. This approach seems sound even when we interpret criminal law's legal categories on the basis only of the manifestations of the social reality (and therefore not on free will, but the attributions of autonomy in the context of social interactions), [31] thereby recognizing the limits of criminal law in being able to access ontological realities that are 'behind' society. Punishing the machine bears the risk of obscuring a fundamental function of criminal law, which is to guarantee not only social security but also the freedom of the (potential) transgressor. This function can be pursued only if policy choices do not focus primarily on the protection of society, but also on the offender, on his dignity and autonomy, holding him responsible only for reprehensible conducts[32].

Finally, not even the construction of parallelism with corporate criminal liability is a valid argument supporting the idea that AI systems can be qualified as subjects of criminal law. Even if it seems that we are facing new persons 'without a soul to damn,' there are fundamental differences between AI systems and corporations. In particular, the punishment of companies, which consists primarily of fines, has a deterrent and preventive function, affecting the essential objective of the existence of corporations, ie their profit. Conversely, framing the essential purpose of AI systems can be very complex. Therefore, even if a «body to kick» exists, especially for AI applications in robotics, it is not clear if «kicking [it] would achieve the traditional goals of punishment»;[33] any penalty we think for the artificial agent, even technical in nature, such as reprogramming or

[28] Gabriel Hallevy, *Liability for crimes involving Artificial Intelligence Systems* (Springer 2015); Ying Hu, 'Robot Criminals' (2019) 52 University of Michigan Journal of Law Reform 488.

[29] Susanne Beck, 'Google Cars, Software Agents, Autonomous Weapons Systems – New Challenges for Criminal Law?', in Eric Hilgendorf and Uwe Seidel (eds), *Robotics, Autonomics, and the Law* (Nomos 2017); Dafni Lima, 'Could AI agents be held criminally liable: Artificial intelligence and the challenges for criminal law' (2018) 69 South Carolina Law Review 677; Sabine Gless, Emily Silverman, Thomas Weigend, 'If robots cause harm, who is to blame: Self-driving cars and criminal liability' (2016) 19 New Criminal Law Review 412.

[30] Not only would the artificial agents lack guilt, but also the punishment that would be imposed on them could not pursue any of the classic functions required by the theories of punishment. See Peter M. Asaro, 'A body to kick, but still no soul to damn: legal perspectives on Robotics', in Patrick Lin, Keith Abney and George A. Bekey (eds), *Robot Ethics: The Ethical and Social Implications of Robotics* (MIT Press 2012) 181.

[31] Monika Simmler and Nora Markwalder, 'Guilty Robots? – Rethinking the Nature of Culpability and Legal Personhood in an Age of Artificial Intelligence' (2019) 30 Criminal Law Forum 1.

[32] Susanne Beck, 'Google Cars, Software Agents' (n 28).

[33] Peter M. Asaro, 'A body to kick' (n 29).

destruction, would always affect the owner or the user. Therefore, the responsibility for harms related to the functioning of artificial agents, however technologically sophisticated, must always and in any case fall on the 'operators'[34] holding a certain position or role in relation to the expert system.

Among the most relevant legal disciplines that should offer legal coverage in regulating such cases, product liability laws provide for a set of rules and standards in this context, given the possible qualification of AI artifacts as 'products.'[35] Therefore, the first risk that operators will have to manage is the probability that the functioning of AI systems causes harm to the users who interact with them because of a defect in their production process.[36] However, there may be new situations where the harm is linked to the emergence that characterizes their functioning, and, therefore, it is not (yet) easily framed within the category of genetic or functional 'defect' (related to the quality of the product or the way it was produced). Moreover, given the complexity of AI systems, which are made of different components created by the cooperation of many organizations and often operating as black boxes, tracing back the harm that occurred to a specific defect can be very complicated (if not impossible).

3 AI Act: A Risk-Based Approach

Since forms of *post facto* control are not a suitable solution in the context of AI, the European Institutions have suggested adopting a risk-based approach, considered as the most suitable tool to guarantee the protection of human rights in the context of AI and, consequently, the creation of trustworthy AI systems. This choice has been recently confirmed by the Proposal for a Regulation on Artificial Intelligence,[37] the Artificial Intelligence Act (AIA), that aims at setting a robust and flexible legal framework for AI. The AIA puts in place a proportionate regulatory system based on the graduation of the risk associated with AI applications.

Given the preliminary stage of the drafting of the Regulation, the text will likely be reviewed and amended. In any event, the articulated content of the proposal cannot be examined here in detail. It will be just outlined, for the purposes of this paper, that the AIA also acknowledges that the regulation of harms related to autonomous artificial

[34] 'Operators' is an umbrella concept that include users and all value chain participants. According to art. 3 § 1 n. 8 of the European Proposal for a Regulation on Artificial Intelligence, 'operators' means the provider, the user, the authorized representative, the importer and the distributor.

[35] European Commission group of experts on 'Responsibility and new technologies', *Liability for artificial intelligence and other emerging digital technologies*, (Bruxelles 2019); among the scholars see Alberto Bertolini, 'Robot as Products: The case for a realistic analysis of robotic applications and liability rules' (2013) 5 Law Innovation and Technology 214.

[36] Carlo Piergallini, 'Intelligenza Artificiale: da 'mezzo' ad 'autore' del reato?' (2020) 4 Rivista italiana diritto e procedura penale 1745.

[37] Commission, 'Proposal for a Regulation of the European Parliament and of the Council, Laying Down Harmonized Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts', COM (2021) 206 final.

agents cannot be based on positions of control of the operators. According to its art. 14, the AIA requires high-risk AI systems to be designed and developed in such a way as to guarantee effective *oversight* by natural persons. The 'oversight' by natural persons is likely to be an overly tight obligation, since it requires the human agent to fully understand the function of the AI system, to be able to correctly interpret its outcomes, and to overcome the so-called 'automation bias' just by being aware of it (namely the bias of over-relying on the functioning of the artificial agent, compromising the effectiveness of monitoring and lowering the level of attention of the operator). Notwithstanding this, the AIA is also focused on the legal obligations of private actors who stand 'behind' the design, development, production, and sale of an artificial agent. What follows is that the core of the proposed Regulation is the compliance requirements, which include a risk assessment and management also through obligations regarding the design phase of high-risk AI systems.

The legislative policy chosen by the European institutions will likely influence the relevant future criminal policies that aim to address the issues raised by crimes related to emergent behaviors, suggesting that also in the field of criminal law a risk-based approach based on the position of private actors that design, develop, produce, and sell AI systems is the most suitable solution.

4 AI Systems in the Risk Society

As the AIA shows us, given their disruptive nature and their increasing diffusive use in many sectors of activities, AI technologies can cause significant and grave violations of human rights and legal goods. To regulate this subject and to offer an answer to the problem of the distance between human behavior and the concrete causation of the harm by an artificial agent, the legal framework cannot be limited to interventions in reaction to these harms. On the contrary, the need for preventive measures, safeguarded by forms of responsibility based on their violation, clearly emerges. Therefore, as new manifestations of our 'risk society,'[38] AI applications must deal with the regulatory logic of anticipation of legal protection that characterizes activities with significant levels of risk. This consideration, together with the progressive loss of control that we witness in dealing with automated and autonomous technological applications, forces the models of criminal law to be based not only on what has been already done but also on what should be done, positioning its center of gravity on cooperation and prevention rather than coercion and retribution. Although the effects that the risk society has had on the categories of criminal law and on the choices of criminal policy have long been the focus of scholars and legislators [39], it remains, to some extent, a fragile and unstable ground for criminal

[38] Ulrich Beck, *Risk Society: Towards a New Modernity* (Sage 1992).

[39] Among countless contributions, see Winfried Hassemer, 'La prevenzione nel diritto penale', (1986) Dei delitti e delle pene 438; Winfried Hassemer, 'Kennzeichen und Krisen des modernen Strafrechts', (1992) Zeitschrift für Rechtspolitik 378; Jesús Maria Silva Sánchez, *La expansión del Derecho penal. Aspectos de política criminal en las sociedades postindustriales* (3rd edn, Edisofer 2011); Andrew Ashworth and Lucia

responsibility models. As known, resorting to preventive criminal law may lead to dangerous criminal policies that give voice to security instances and expressions of feelings of fear dominate from which it is necessary to take clear and safe distances. Nevertheless, when policy choices are based on empirical elements of actual danger to human rights and legal goods, as might be those related to the autonomy and emergence of artificial agents (a new generation of man-made risks systematically produced), the legal system must offer an answer to protect its citizens from possible violations.

Following this logic, the operators could be held criminally responsible for not complying with the standards and legal obligations provided for by the relevant legal sources, such as the AI Act, product liability laws, and other specific international standards (for example, the AI-based medical devises regulation), when their behavior can be qualified as 'guilty,' due to intent, recklessness or negligence, and when the chain of causation between the conduct and the harm can be established and proven. However, these legal obligations should be able also to manage the risk of emergent behaviors of AI. The limits of these policy choices, as is well known, concern the state of mind requirement and the chain of causation.[40] Narrowing the analysis (for brevity reasons) to the first element, expanding the scope of criminal law to the harms related to emergent behaviors of AI systems challenges the possibility to hold the human operator criminally responsible in compliance with the principle of culpability, which is an indispensable stronghold, especially when criminal law is applied in the context of risk management, since it becomes a fundamental safeguard to ensure that criminal law also addresses and protects the dignity and re-socialization of the offender, avoiding focusing only on crime control. These limits will have to be addressed in the context of AI, especially to not leave their analysis and solution only to the decisions of the Courts.

Leaving a responsibility gap is not an acceptable solution both for ethical and legal reasons. 'Emergence' is a specific technical feature of the most sophisticated AI systems. Since this property consents them to perform better, they are willingly designed and produced through techniques that allow them, to an extent, to elaborate outcomes that the operator had not thought in advance. Therefore, 'emergence' and the risks related to the 'unpredictability' that comes with it can be (and should be) managed by the same operators that create and actualize them. In this perspective, policymakers should distinguish, as suggested by some American scholars,[41] between the 'known unknowns,' for instance, possible system's failure[42] for which human operator, if behaved with intent, reckless, or negligence, can be held responsible, and the 'unknown unknowns', truly unforeseeable outcomes that remain possible even though human operators have complied with the duty of care provided for by the relevant legislation and international standards.

Zedner, *Preventive Justice* (Oxford University Press 2014); Massimo Donini, Pavarini (eds.), *Sicurezza e diritto penale* (Bologna 2011).

[40] Carlo Piergallini, 'Intelligenza Artificiale' (n 35).

[41] William D. Smart, Cindy M. Grimm, Woodrow Hartzog, 'An education theory' (n 7).

[42] ibid.

Therefore, if only the human operator can be responsible for harms related to the autonomous functioning of an AI system, the nature of this responsibility remains to be established.

5 The Responsibility Gap: A Possible Roadmap

Given that this field of investigation is in its early stages, we can only draft the alternative scenarios that the responsibility gap can determine. Aware of the limit of the still very theoretical nature of this analysis, it is possible to outline two possible approaches.

The first one is a 'techno-skeptical' approach. It relies on the precautionary principle, which imposes general abstentionism in case of situations of scientific uncertainty, and it is articulated on the proposal to limit the autonomy of AI systems under certain circumstances or in the performance of specific tasks.

The second approach aims instead at developing a solution to the responsibility gap. As we found ourselves in uncharted territories, the consequent question would be whether criminal law, following its polar star, the principle of *ultima ratio*, can be included among the legal instruments that shape the regulatory framework in the context of AI or whether, in case of harms related to the emergence of AI systems, only tort laws or administrative sanctions can be applied, as expressions of a broader conception of 'punitive law.' Thus, the road of this second scenario forks, leading to two different policies: (a) limiting the scope of criminal law to the case of harms related to emergent behaviors of AI systems, due to the progressive acceptance of the risk of harm they create, which becomes allowed given the benefits that the use of these technologies offers to the society as a whole; (b) resorting to criminal law and, if necessary, adapting the modes of criminal responsibility in cases involving autonomous machines, for instance through an adaptation of the concept of negligence.

Whether or not the national legislators choose to use the criminal instrument or to limit it in the context of AI, it is desirable, as suggested by the European Committee on Crime Problems of the Council of Europe,[43] that national regulations develop within an international and collaborative framework: in this preliminary phase, a Council of Europe instrument on AI and criminal law would indeed have a powerful and essential impact to guarantee a sufficient level of legal certainty and international cooperation.

The first path is based on a scientific-skeptical approach that uses the precautionary principle[44] either as a directive of legislative policy or as the basis to elaborate models of responsibility, with the inevitable result of limiting the autonomy of AI systems. Because

[43] Council of Europe, European Committee on Crime Problems, 'Feasibility Study on a future Council of Europe instrument on Artificial Intelligence and Criminal Law' CDPC(2020)3Rev.

[44] Das Vorsorgeprinzip in the German legal system, see Cornelius Prittwitz, *Strafrecht und Risiko. Untersuchungen zur Krise von Strafrecht und Kriminalpolitik in der Risikogesellschaft, (*Klostermann 1993); Donato Catronuovo, *Principio di precauzione e diritto penale. Paradigmi dell'incertezza nella struttura del reato,* (Aracne 2012).

of the uncertainty that could characterize the outputs of AI systems used in applications capable to harm fundamental human rights and legal goods (such as the right of protection of personal data, the right of non-discrimination, or the right to the integrity of the person), the legislator can decide to limit the tasks that can be assigned to an AI system or the autonomy that can characterize its functioning. The proposed AIA follows this approach in prohibiting certain AI practices, deemed too dangerous for the protection of human rights. However, prohibitions of certain uses of AI technologies, especially if defined with broad and uncertain terms, cannot be a general approach in the context of new technologies regulation, because they could be very limiting, given the potential of the development and use of AI.

Therefore, even though it seems preferable that legislators avoid assuming a general 'precautionary approach' in their criminal policies on AI, at the same time, a 'moderate' conceptualization of the precautionary principle may be useful in a phase of creation and development of new scientific and technological innovations such as high-risk AI systems. This approach has also been adopted by the proposed AIA, article 53 of which provides for AI regulatory sandboxes: controlled environments that facilitate the development, testing, and validation of innovative AI systems for a limited time before their placement on the market or putting into service, to ensure compliance with the requirements of the Regulation. Therefore, even though certain limits to technological autonomy will have to be established (and their violation can perhaps even be criminalized), at the same time, the path of limiting the autonomy of AI systems cannot be a general approach in regulating harms that can be caused by artificial agents' functioning.

6 Emergence and Guilt: An Irreconcilable Dualism for Criminal Law?

In the passage from control and responsibility to design and accountability, the question is whether criminal law is the right tool not only to contrast and prevent, but also to regulate situations where the harm is related to emergent behaviors of an artificial agent.

A first proposal may be to limit the criminal liability of operators in case of harms related to the autonomous and unpredictable functioning of the AI system «reducing of their duty of care». According to this proposal, operators who comply with the strict standards provided for by the specific legislation (eg product liability laws and AIA proposal) «have fulfilled their duty of care, even if they remain aware (together with all society) of the permanence of certain risks».[45] This approach is based on the prospect that the use of AI applications will become increasingly widespread and generalized, leading the legislator to consider the risk that may arise from artificial agents as a risk objectively and subjectively tolerated, accepted for the great benefits that the use of such technologies brings to the society itself, which would therefore bear the consequences of harmful results (considered as 'side effects') caused by the emergent behavior of the artificial system. According to this first solution, in case of harm caused by the emergent functioning

[45] Sabine Gless, Emily Silverman and Thomas Weigend, 'If robots cause harm' (n 28).

of the AI system, if the operator complies with the standard of care required, they cannot be held criminally liable, and only forms of civil liability remain.

However, this hypothesis presents two limits. On the one hand, the qualification of the risk as accepted by society due to the benefits brought by the implementation of AI systems can only be articulated in a single sector (the literature on the point concerns self-driving cars); it may lead to different considerations according to each AI application and its level of risk. On the other hand, resorting only to remedies of an administrative or civil nature may not be sufficient to guarantee effective results in terms of deterrence and prevention, that must be proportionate with the gravity of the harm. Moreover, cases in which a specific victim cannot be identified would remain without compensation.

A second path may be to include criminal law among the regulatory legal sources and to consider the possibility of elaborating a new legal framework for AI-related crimes (in compliance with the *ne bis in idem* principle regarding the administrative sanctions provided for by the proposed AIA), which will have to look at the definition of the standard of care required in this context, since it might need some adjustments. The content of 'negligence' should indeed be able to face the challenges posed by: (i) the distance from the actions of the operators and the execution of the task and its results; (ii) the control gap; (iii) the unpredictability gap. In this last scenario, one of the first questions will be whether to use crimes of negligence directed to single persons or crimes involving corporate liability.[46]

Resorting to personal criminal responsibility in the context of AI seems particularly problematic because it presents two issues. The first concerns, as mentioned before, the culpability principle. The commission of the crime in these cases is 'contaminated' by the autonomous functioning of the AI system and therefore the set of preventive measures that the operator must implement *ex ante* are very distant from the realization of the possibly unforeseeable harm. This will inevitably lead to forms of strict liability in criminal law, that are not constitutionally legitimate in all legal systems and therefore cannot be considered as an answer to our question. The second problem is the reality of networks that stands behind the AI system, the problem of many hands. Another transformation brought already by the ICT and outlined by Luciano Floridi is the «shift from the primacy of stand-alone things, properties, and binary relations to the primacy of interactions, processes and networks».[47] AI technologies amplify this process, as it is very difficult (if not impossible) to trace back the harmful output of the artificial agent to a single human agent.

These two elements could be the cause but also the solution to the problem. They indeed suggest that forms of distributed legal responsibility, elaborated on a risk-based approach that keeps the principle of accountability on its core, represent the most suitable

[46] Woodrow Barfield, Ugo Pagallo, *Advanced Introduction to Law and Artificial Intelligence* (E Elgar 2020) 108.

[47] Luciano Floridi (ed), *The onlife manifesto* (n 1).

option in the context of AI. Within those legal systems that adopt the organizational approach in their corporate liability laws,[48] in case of AI applications that present a high risk for the protection of human rights and legal goods, this possible form of corporate liability could be structured either on the failure, realized by not complying with the system of legal obligations relevant for the case, to prevent the risk of occurrence of criminal offenses related to the AI application programmed, developed, produced or sold; or on the creation of new corporate crimes, following the model, for instance, of the English legal system.[49] In either case, the due diligence defense would allow the organizations to avoid liability.

Regardless of which policy choice is to be adopted, the legal obligations placed on corporations, as the AIA suggests, should primarily relate to the design phase of AI systems (concept that encompasses the different phases of programming, development, testing, and production), subject to additional *ex post* monitoring obligations. Adopting this approach, the activity of private actors that create, actualize, and manage the risk related to AI applications would be informed not only by the principle of virtuous self-organization but also by a new general duty shaped by the principle of Legal Protection by Design,[50] which entails not only organizational measures and choices but also technical ones. The combination of these two principles, whether the criminal policy chosen requires a form of fault of corporations for them to be held criminally liable, seems to lead to the possible conceptualization and definition of a specific form of *mens rea*, that could be more suitable in this context: what we might call a '*guilt in design*.'[51] The principal aim of this perspective is not to identify some form of control over AI systems that, if not exercised correctly, may lay the basis to hold liable the corporation. It is instead to shift our focus on forms of accountability (that could then lead to criminal liability) based on the (legal) design of the decision-making process that we decide to delegate (completely or partially) to an artificial agent, to make it explainable and contestable, and also, built on the necessary interaction between AI systems and operators or users involved in the different phases of the life cycle of the artificial agent.

---

[48] Johannes Keiler and David Roef (eds), *Comparative Concepts of Criminal Law* (3rd edn, Intersentia 2019) 337.

[49] See Corporate Manslaughter and Corporate Homicide Act 2007; failure to prevent bribery according to sec. 7 Bribery Act 2010; and failure to prevent facilitation of tax evasion offenses according to sec. 45 and 46 Criminal Finances Act.

[50] Mireille Hildebrandt, 'Legal Protection by Design: Objections and Refutations' (2011) 5 Legisprudence 223.

[51] From a technical point of view, 'design' means the process of developing and engineering specific technologies, and the process of introducing and employing those technologies in human society (design as a verb), including also the outcome of that process (design as a noun). Design is not only about the form of the interface, but also about the back end of technological systems and how they frame interactions in the front end. See Mireille Hildebrandt, 'Saved by Design? The Case of Legal Protection by Design', (2017) 11 Nanoethics 307. However, this concept is also being colored by a more properly sociological and anthropological dimension, as a tool in the hands of the poietic disciplines that create and manage the digital technologies that are, in fact, 'designing' our reality and society. See Luciano Floridi, *The 4th Revolution. How the Infosphere is Reshaping Human Reality* (OUP Oxford 2014).

7 Concluding Remarks: The Power of Architecture

The paper aimed to outline some of the possible and future criminal policies regarding the regulation of harms related to the use and functioning of AI systems. Even though each policy choice presents downsides, legislators might need to use every one of them depending on the specific AI practice and system, its area of application and its level of risk. While AI regulation is undoubtedly a new field that create certain conceptual gaps, the novelty is not absolute and useful references can be found not only in the general issues posed by preventive criminal law, but also in the studies related to previous generations of digital technologies, such as the Internet. Thus, Internet governance, cyberlaw and ICT's criminal law can offer a base line of legal categories and conceptual tools from which to start.

These disciplines, along with Science and Technology Studies, highlight that the choices of architecture have a powerful effect in shaping the human affordances related to new technological artifacts. Using the words of the scholar Viktor Mayer-Schönberger, the 'mutuality of influences between technology and society' generates the need of coloring with normativity the design, seen in its sociological meaning as the series of decisions that affect individual and group behavior. The alternative to this approach is to be regulated by architecture, as Lawrence Lessig explained with his famous expression 'code is law'. Given the early stages of the field of AI and criminal law what can be outlined are the options that lay ahead. However, in picturing a possible roadmap, the supranational context must be taken into first consideration, and, if we look at the European Union and the packages of regulations and proposals aimed to gain and establish digital sovereignty over digital technologies applications and services (starting from the GDPR until the AI Act) the choice to attribute legal significance to the 'design' phase seems to acquire a progressively clear identity.

References

Abbott R., Sarch A., 'Punishing Artificial Intelligence: Legal Fiction or Science Fiction' (2019) 53 UC Davis Law Review 325

Asaro P. M., 'A body to kick, but still no soul to damn: legal perspectives on Robotics', in Lin P., Abney K., Bekey G. A. (eds), *Robot Ethics: The Ethical and Social Implications of Robotics* (MIT Press, Boston 2012)

Bagnoli C., *Teoria della responsabilità* (Il Mulino 2019)

Beck S., 'Dealing with the diffusion of legal responsibility: the case of robotics', in Battaglia F., Mukerji N., Nida-Rümelin J. (eds), *Rethinking Responsibility in Science and Technology*, (Pisa University Press 2014)

—— 'Intelligent agents and criminal law-Negligence, diffusion of liability and electronic personhood' (2016) 86 Robotics and Autonomous systems 138

—— 'Google Cars, Software Agents, Autonomous Weapons Systems – New Challenges for Criminal Law?', in Hilgendorf E. and Seidel U. (eds), *Robotics, Autonomics, and the Law* (Nomos 2017)

Beck U., *Risk Society: Towards a New Modernity* (Sage 1992)

Calo R., 'Robotics and the Lessons of Cyberlaw' (2016) 103 California Law Review 513

—— and Evtimov I., Fernandes E., Kohno T., O'Hair D., 'Is Tricking a Robot Hacking?' (2019) 5 University of Washington School of Law Research 1

——*Law and technology as method* (Oxford University Press forthcoming)

Cappellini A., 'Machina delinquere non potest? Brevi appunti su intelligenza artificiale e responsabilità penale' (2019) Discrimen 1

Castronuovo D., *Principio di precauzione e diritto penale. Paradigmi dell'incertezza nella struttura del reato*, (Aracne 2012)

Creese S., 'The threat from AI', in Dennis J. Baker and Paul H. Robinson (eds), *Artificial Intelligence and the Law. Cybercrime and Criminal Liability* (Routledge 2021)

De Francesco G. and Morgante G. (eds), *Il diritto penale di fronte alle sfide della «società del rischio». Un difficile rapporto tra nuove esigenze di tutela e classici equilibri di sistema* (Milano 2017)

Donini M., *Il volto attuale dell'illecito penale. La democrazia penale tra differenziazione e sussidiarietà*, (Giuffrè 2004)

Donini M. and Pavarini M. (eds), *Sicurezza e diritto penale* (Bononia 2011)

Floridi L. (ed), *The onlife manifesto. Being human in a hyperconnected era* (Springer 2014)

—— *The 4th Revolution. How the Infosphere is Reshaping Human Reality* (OUP Oxford 2014)

Gless S., Silverman E., Weigend T., 'If robots cause harm, who is to blame: Self-driving cars and criminal liability' (2016) 19 New Criminal Law Review 412

Grote T., Di Nucci E., 'Algorithmic Decision-Making and the Problem of Control', in Beck B. and Kühler M. (eds), *Technology, Anthropology, and Dimensions of Responsibility* (Springer 2020)

Hallevy G., *Liability for crimes involving Artificial Intelligence Systems* (Springer 2015)

Hayward K. J., Maas M. M., 'Artificial intelligence and crime: A primer for criminologists' (2020) Crime Media Culture 1

Hildebrandt M., 'Legal Protection by Design: Objections and Refutations' (2011) Legisprudence 223

—— 'Criminal Law and Technology in a Data-Driven society', in Dubber M. and Hörnle T. (eds), *The Oxford Handbook of Criminal Law* (Oxford 2014)

—— 'Saved by Design? The Case of Legal Protection by Design' (2017) 11 Nanoethics 307

Hilgendorf E., 'Automated Driving and the Law', in Hilgendorf E. and Seidel U. (eds), *Robotics, Autonomics and the Law* (Nomos 2017)

Hu Y., 'Robot Criminals' (2019) 52 University of Michigan Journal of Law Reform 488

Keiler J. and Roef D. (eds), *Comparative Concepts of Criminal Law* (3rd edn, Intersentia 2019)

King T., 'Projecting AI-Crime: A Review of Plausible Threats', in Öhman C. and Watson D. (eds), *The 2018 Yearbook of the Digital Ethics Lab* (Springer 2019)

King T., Aggarwal N., Taddeo M., Floridi L., 'Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions' (2020) 26 Science and Engineering Ethics 89

Lagioia F. and Sartor G., 'AI systems under Criminal Law: a Legal Analysis and a Regulatory Perspective' (2019) Philosophy & Technology 1

Lamo M. and Calo R., 'Regulating Bot Speech' (2019) U.C.L.A. Law Review 988

Lima D., 'Could AI agents be held criminally liable: Artificial intelligence and the challenges for criminal law' (2018) 69 South Carolina Law Review 677

Matthias A., 'The responsibility gap: Ascribing responsibility for the actions of learning automata' (2004) 6 Ethics and Information Technology 175

——*Automaten als Träger von Rechten. Plädoyer für eine Gesetzänderung* (dissertation, Humboldt Universität 2007)

Mcallister A., 'Stranger than science fiction: The rise of AI interrogation in the dawn of autonomous robots and the need for an additional protocol to the UN convention against torture' (2017) 101 Minnesota Law Review 2572

Neff G., Nagy P., 'Talking to Bots: Symbiotic Agency and the Case of Tay' (2016) 10 International Journal of Communication 4915

Pagallo U. Quattrocolo S., 'The impact of AI on criminal law, and its twofold procedures', in Barfield W and Pagallo U. (eds), *Research Handbook on the Law of Artificial Intelligence* (Cheltenham 2018)

Pagallo U., 'From automation to autonomous systems: a legal phenomenology with problems of accountability', in *IJCAI International Joint Conference on Artificial Intelligence* (International Joint Conferences on Artificial Intelligence 2017)

Parker D., *Crime by Computer* (1st edn, Cengage GALE 1976)

Piergallini C., 'Intelligenza Artificiale: da 'mezzo' ad 'autore' del reato?' (2020) 4 Rivista italiana diritto e procedura penale 1745

Prittwitz C., *Strafrecht und Risiko* (Klostermann 1993)

Salvadori I., 'Agenti artificiali, opacità tecnologica e distribuzione della responsabilità penale', (2021) 1 Rivista italiana diritto e procedura penale 83

Scharre P. and Michael Horowitz, 'An Introduction to Autonomy in Weapon Systems', Center for a New American Security Working Paper 2015

Simmler M., Markwalder N., 'Guilty Robots? – Rethinking the Nature of Culpability and Legal Personhood in an Age of Artificial Intelligence' (2019) 30 Criminal Law Forum 1

Smart W., Grimm C., Hartzog W., 'An education theory of fault for autonomous systems' (2021) 2 Notre Dame Journal on Emerging Technologies 33

Stortoni L. and Foffani L. (eds), *Critica e giustificazione del diritto penale nel cambio di secolo. L'analisi critica della scuola di Francoforte,* (Giuffrè 2004)

THE IMPACT OF AI ON CORPORATE CRIMINAL LIABILITY: ALGORITHMIC MISCONDUCT IN THE PRISM OF DERIVATIVE AND HOLISTIC THEORIES

By Federico Mazzacuva[*]

Abstract

While the issues related to artificial intelligence (AI) accountability are largely discussed by scholars, algorithmic corporate liability may represent a problem equally important and urgent to deal with, since the main risks come from the use of new technologies by corporations. This article precisely reviews the interplay between AI and the main systems of corporate criminal liability: the focus is on purely algorithmic corporate misconduct, rather than the cases where employees purposely, knowingly, or recklessly design AI systems to break the law. To this end, the most common regimes of corporate criminal liability will be considered: strict and vicarious liability; the principle of identification and, finally, the responsibility based on organizational fault or corporate culture. Some concluding remarks follow on the opportunity of an ad hoc regulation in order to incentivize corporations to accelerate their embrace of automation and, at the same time, to promote compliance.

1 Introduction

It is well known that technological and economic developments that influence society eventually result in a change to the existing legal framework; this is also true in the field of criminal law, even if its interaction with technology is discussed by scholars.[1]

From this viewpoint, the two foci of this article, ie artificial intelligence (AI) and corporate criminal liability, are both formidable examples of this process.

On the one hand, during the last decades corporate criminal liability has faced several challenges posed by globalisation as well as by the recent market crisis;[2] on the other hand, digitalisation and the advent AI have shaped all aspects of human life:[3] not only

[*] PhD in Criminal Law at the University of Milan-Bicocca, Department of Business and Law, Milan. For correspondence: <federico.mazzacuva@unimib.it>.

[1] For a more extensive discussion on the interplay between criminal law and technology, see for instance Serena Quattrocolo, *Artificial Intelligence, Computational Modelling and Criminal Proceedings. A Framework for A European Legal Discussion* (Springer 2020) 3ff. On the impact of AI on criminal law, see Christoph Burchard, 'Is Artificial Intelligence putting an End to Criminal Law? On the Algorithmic Transformation of Society'(2019) Riv it dir proc pen 1909ff.

[2] In the recent literature, see for instance Kataline Ligeti and Stanisław Tosza, 'Challenges and Trends in Enforcing Economic and Financial Crime. Criminal Law and Alternatives in Europe and the US', in Katalin Ligeti and Stanisław Tosza (eds.), *White Collar Crime. A Comparative Perspective* (Hart Publishing 2019) 1ff.

[3] Indeed, the digital turn led to a cultural, mental, physic and even 'postural' revolution, since the combination screen-keyboard-human being almost defines the present era: in these terms, see Alessandro Baricco, *The Game* (Einaudi 2018) 42; more widely on these topics, see also George Dyson, *Turing's Cathedral.*

are the new technologies influencing modern business, but they are also radically changing the criminal justice system, in a 'semiotic' as well as a 'spatial' fashion.[4]

As far as corporations are concerned, AI has already taken over human functions throughout the organization hierarchy, from the lowest-level operations to the highest. If at present the market faces an increasing prevalence of AI throughout a variety of industries and fields, experts predict that corporate reliance on digital autonomation will increase exponentially over the coming years.[5]

While algorithms and AI promise to make corporations more efficient, they may also soon replace employees as the leading cause of corporate wrongdoing; thus, in the case in which a crime is committed by AI, is the legal system able to allocate responsibility? In other words: under what conditions should corporations be liable when AI engages in misconduct?[6]

Indeed, 'in cases of algorithmic misconduct, it is particularly important that the path hold open the possibility of corporate liability. As corporations replace employees with algorithms, corporate liability becomes the *only* means of redress'.[7] For instance, the lack of a liability theory recently led prosecutors to decline charges against Uber when one of its self-driving cars struck and killed a pedestrian in Arizona.[8]

It is worth emphasizing that the issue herein discussed does not refer to cases where employees purposely, knowingly, or recklessly design AI systems to break the law; rather to the case where the employee's misconduct is 'removed' from the picture, so that law has to handle a purely algorithmic corporate misconduct.

In such a perspective, the article aims to discuss some initial considerations about the interplay between AI and the main systems of corporate criminal liability, and to analyse how they fit together. Although several critical differences persist between the common law and the civil law traditions, and among the legal systems of continental Europe, the most common regimes of corporate criminal liability will be considered, starting from

---

The Origins of the Digital Universe (Penguin Books 2012) and, in the Italian literature, Roberto Cingolani, *L'altra specie* (Il Mulino 2019).

[4] According to the analysis conducted by Antoine Garapon, *Justice digitale* (Presses Universitaires de France 2018); see also Antoine Garapon, *La despazializzazione della giustizia* (Mimesis edizioni 2021).

[5] See Mihailis E Diamantis, 'The Extended Corporate Mind: When Corporations Use AI to Break the Law' (2020) 98 N C L Rev, 895ff.

[6] The issue herein discussed has been defined the 'dark side' of the use of AI and data analytics for corporate compliance: see Rossella Sabia, *Artificial Intelligence and Environmental Criminal Compliance* (2020) 91(1) RIDP 179. From a broader perspective, with regards to the risk that 'AI may produce a novel generation of loopholes in the criminal law field, forcing lawmakers to intervene at both national and international level', see Ugo Pagallo and Serena Quattrocolo, 'The impact of AI on criminal law, and its twofold procedures', in Woodrow Barfield and Ugo Pagallo (eds.), *Research Handbook on the Law of Artificial Intelligence* (Edward Elgar Publishing 2018) 386.

[7] Diamantis (n 5) 898.

[8] ibid. 898.

the more objective one (strict liability), followed by 'derivative theories' (*respondeat superior* and the principle of identification) and 'holistic theories'[9].

Before proceeding any further, it is worth precising the notion of AI to which the article refers, even if this is not an easy task. Indeed, since AI covers a broad range of evolving technologies and disciplines (not only computer science, but also philosophy, psychology, linguistics etc.), its definition varies from the context.

While the concept of 'artificial intelligence' was firstly used in the '50s,[10] in 2020 the Joint Research Centre (JCR), the EU Commission's Science and Knowledge Service, delivered a comprehensive investigation on the several definitions of AI developed so far,[11] proposing 'an operational definition of AI formed by a concise taxonomy and a set of keywords that characteries the core and transversal domains of AI'.[12] The report identifies four common features, within the manifold range of AI definitions, namely the fact that a system is based on: consideration of the real-world complexity; information processing

[9] For a more extensive discussion on the different models of corporate fault, see William S Laufer, *Corporate Bodies and Guilty Minds. The Failure of Corporate Criminal Liability* (The University of Chicago Press 2006).

[10] More extensively, see Gabriel Hallevy, *Liability for Crimes Involving Artifical Intelligence Systems* (Springer 2015) 3; Fabio Basile, 'Artificial intelligence and criminal law: four possible research leads' (2019) DPU <https://archiviodpc.dirittopenaleuomo.org/upload/3089-basile2019.pdf> accessed 28 August 2021.

[11] There are indeed other definitions of AI: see Hallevy (note 10) 6ff. See also, for instance, the definition provided by the 'European ethical Charter on the use of Artificial Intelligence in judicial systems and their environment' (2018), adopted by the European Commission for the Efficiency of Justice (CEPEJ) 69 <https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c> accessed 28 August 2021, according to which AI is 'a set of scientific methods, theories and techniques whose aim is to reproduce, by a machine, the cognitive abilities of human beings'; or by the European Commission in the 'Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative Acts' (2021) <https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF > accessed 28 August 2021, whose article 3 defines an 'artificial intelligence system' as a 'software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with' (Annex I refers to the following techniques and approaches: machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning; logic- and knowledge-based approaches, including knowledge representation, inducive logic programming, knowledge bases, inference and deductive engines, symbolic reasoning and expert systems; statistical approaches, Bayesian estimation, search and optimization methods').

[12] JCR Technical Reports, 'AI Watch: Defining Artificial Intelligence. Towards an operational definition and taxonomy of artificial intelligence' (2020) <https://publications.jrc.ec.europa.eu/repository/handle/JRC118163> accessed 28 August 2021. Moreover, according to the JCR Flagship report on Artificial Intelligence (2018), AI is 'a generic term that refers to any machine or algorithm that is capable of observing its environment, learning, and based on the knowledge and experience gained, take intelligent actions or propose decisions' (European Commission, Joint Research Centre, 'Artificial Intelligence. A European Perspective') <https://publications.jrc.ec.europa.eu/repository/handle/JRC113826> accessed 28 August 2021.

(collecting and elaborating inputs); decision making and, finally, achievement of specific goals.

The broad definition above-mentioned fits the purposes of the present article since the interactions between AI and corporate liability are precisely considered from a functional viewpoint.[13]

2 Algorithmic Misconducts and Derivative Theories of Corporate Criminal Liability

2.1 Strict liability and vicarious liability

Starting from the impact of new technologies on strict liability and *respondeat superior* regimes, it is useful to remember that the former does not require proof of a state of mind, but only the material element of the wrongdoing (*actus reus*),[14] while the latter defines corporate *mens rea* in terms of employee mental states.[15]

More specifically, in the US and in the English case law, *respondeat superior* is an example of vicarious liability borrowed from tort-law. According to this principle, the whole organization shall be held liable when an offence has been committed by one of its employees while performing collective activities to obtain a general, rather than an individual, benefit.[16]

Having said this, if a corporation uses AI to perform its activity, and if wrongdoing occurs in these circumstances, the only way to avoid legal loophole is resorting to extensive interpretation. In this regard, scholars have already theorized some 'surgical interventions' to the legal framework, precisely aimed at proposing an extension, rather than a rewriting, of current law.

---

[13] More widely on the functional definition of AI, see Quattrocolo (note 1) 7ff. For the stake of brevity, the terms 'algorithm', 'AI', 'big data analytics', 'robots' etc. are used interchangeably throughout this article, knowing that this entails a severe simplification of the differentiation among these categories and diverse subcategories.

[14] The strict liability typically applies to a vast catalogue of regulatory offences in the UK legal system, ie statutory offences that regulate sectors such as health and safety, labour and trading standards: such kind of provisions are usually defined not in terms of result (eg causing death), but in terms of a failure to comply with risk-assessed standards. More generally on the different models of corporate criminal liability, see James Gobert, 'Corporate criminality: four models of fault' (1994) 14 Legal Stud 393ff.

[15] There are, indeed, other models of liability, such as the so-called collective knowledge and collective intent, which allow courts to aggregate employee knowledge: in this regard, however, it is easy to observe that corporate mental state is once again derived from employees' mental state. For more details on these models, see for instance Nicola Selvaggi, 'Criminal Liability of Corporations and Compliance Programs in the U.S. System' in Antonio Fiorella (ed.), *Corporate Criminal Liability and Compliance Programs*, vol. I, *Liability 'Ex Crimine' of Legal Entities in Members States* (Jovene editore 2012), 601ff.

[16] *New York Central and Hudson River Railroad Company v United States* [1909] 212 U.S. 481. For related comments see, *inter alia*, Leonard Orland, 'The Transformation of Corporate Criminal Law' (2006) 1 Brook J Cor Finn & Comm L 45ff. As far as the English case law is concerned, see Gobert (note 14).

The basic assumption is that corporate activities and minds can be extended beyond their traditional limits (ie the activities and minds of individual employees) to include other functionally integrated corporate systems. Functionalism is indeed neutral about the persons/systems that operate or realize mental states. In other words, algorithms do not act or think on their own; rather, corporations act or think through their AI systems,[17] so that through an extended interpretation of strict liability or *respondeat superior* regimes, law can properly hold corporations accountable for the wrongdoing of their algorithms.[18]

More specifically, while strict liability seems to raise fewer problems, since on the one hand the new '*homo technologicus*' is capable of performing 'acts'[19] and, on the other hand, AI's *actus reus* can be extensively considered a corporate conduct; from the vicarious liability perspective, on the contrary, a collective entity can be held accountable for AI misconduct if: i) the crime has been committed through some information 'known' by the algorithm and, thus, by the corporation; ii) the algorithm knows the information within the scope of its use inside the organization; iii) the algorithm must use the information in a way that accrues some benefit to the company itself.[20]

2.2 Identification principle

The situation is partially different when the corporate criminal liability regime is based on the identification principle (or *alter ego* principle).

As it is well known, in the UK legal system, corporate criminal liability did not develop in a consistent or linear fashion, as it is rather the result of a 'pragmatic adaptation rather than formulated policy'.[21] Thus, a part from the above-mentioned regulatory offences and the special statutory provisions considered below, the principle of identification is a criterion to charge a corporation with a direct liability with reference to *mens rea* offences, ie those crimes which require proof of a blameworthy mental state such as knowledge or recklessness. According to this criterion, some individuals, especially qualified in the organisation of the company, identify themselves as the company's 'brains' or 'minds' (the so-called 'senior' or 'controlling officers'). As a result, if a wrongdoing is committed by such kind of persons while performing their own functions, it must also be considered as coming directly from the company itself.[22]

---

[17] From this standpoint, scholars have already argued that AI has 'the capacity of fulfilling the awareness requirements in criminal law': Hallevy (note 10) 91.

[18] For a more extensive discussion on the extended mind thesis, see Diamantis (note 5).

[19] As has been observed, 'this is true not only for strong artificial intelligence technology…even sub-artificial intelligence technology machines have the factual capability of performing acts': see Hallevy (note 10) 61.

[20] According to the so-called 'extended mind thesis' proposed by Diamantis (note 5).

[21] In these terms, see Celia Wells, 'Economic Crime in the UK. Corporate an Individual Liability' in Katalin Ligeti and Stanisław Tosza (eds.), *White Collar Crime. A Comparative Perspective* (Hart Publishing 2019) 256.

[22] The leading case is the well-known judgment *Tesco Supermarkets Ltd v. Nattrass* [1972] AC 153, 170 and 173. See recently, among other authors, Wells (note 21) 256ff.

Since identification principle is a peculiar derivative theory, the adaptive interpretation discussed in the precedent paragraph may operate, even though some further 'adjustments' are required. More specifically if, according to the functionalist understanding of mental states, any system carrying out the same functional role as employees can also form part of the corporate mind, the only relevant algorithms – in the light of the *alter ego* doctrine – are those that perform the highest-level operations.

The issue under discussion is less theoretical than one might expect. As it has been noted, even if 'it seems even more unlikely that a task erroneously performed by AI can be linked in some way to top management…since…in the corporate context such tools usually assist or replace compliance staff., ie intermediate or low level employees',[23] there are several companies that have already appointed an AI software to their board of directors as a fully voting member of their management team, instead of an individual.[24]

3 Robots and Holistic Theories of Corporate Criminal Liability

3.1 Organizational fault

Holistic theories represent a different approach for ascribing liability, which prevailingly attaches importance to the idea of a fully 'subjective' corporate blameworthiness: knowing that this entails a severe simplification, these theories generally refer to the notions of organizational failure or corporate culture.[25]

The defective organization model is the basic paradigm of the Italian corporate criminal liability system introduced by Legislative Decree No. 231/2001 (Decree 231), implementing the Law No. 300/2000.[26] It is well known, indeed, that even if articles 5-7 of the Decree 231 provide several imputation criteria, compliance programs play a pivotal role within the discipline of the *ex crimine* corporate liability, since their adoption could definitely

[23] Sabia (note 6) 189.

[24] To take just few examples, in 2014 the Hong Kong based venture capital firm, Deep Market Ventures, appointed an AI software entity, Vital, to its board of directors; although extant law prohibited Vital from enjoying the formal legal status of a board member, the other human directors afforded Vital the 'observer' status at each board meeting and allowed Vital to vote on all financial investment decisions. Moreover, although no AI entity like Vital currently occupies a formal seat on a corporate board, at least one European company, Tieto, has already appointed a similar autonomous AI entity, Alicia T, as a fully voting member of its management team. The mentioned cases are discussed by Pagallo and Quattrocolo, (note 6) 386 and by Michael R Siebecker, 'Making Corporations More Humane through Artificial Intelligence' (2019) 45 J Corp L 96-97.

[25] For a comparative study, see for instance Enrica Villani, 'Compliance Programs and "Organisational fault" in Europe' in Antonio Fiorella (ed.), *Corporate Criminal Liability and Compliance Programs*, vol. II, *Towards a Common Model In The European Union*, (Jovene editore 2012), 249ff.

[26] It is worth remembering that, even if the responsibility is formally defined by law as 'administrative', it can be considered at least as para-criminal liability, due to the punitive weight of the sanctions provided by the Decree 231.

exclude the possibility of holding the *societas* responsible for the misconduct of its members.[27]

However, as stated above, the organizational fault is only a *conditio sine qua non* rather than a *conditio per quam*: indeed, according to the Italian legal system, in order to hold a corporation liable it is necessary, first of all, that one of the crimes listed by the Decree 231 has been committed 'in the interest or for the benefit' of the corporation itself by a 'person' placed in the corporate hierarchy, from the lowest-level operations to the highest (the objective requirement provided by article 5, paragraph 1 of the Decree 231)[28]. In this regard, some questions raise immediately: indeed, even assuming that AI can be considered as a 'person' (which requires at least the recognition of its legal personhood), the existence of such a requirement is not automatic with reference to algorithmic misconduct, unless we demonstrate that the use of AI to perform business activity either is intentionally aimed at obtaining illegal profits or represents itself a saving for the business.[29]

As far as the UK legal system is concerned, a similar model characterizes the offences of 'corporate manslaughter' and 'failure of commercial organisations to prevent bribery'. According to the Corporate Manslaughter and Corporate Homicide Act (2007), the collective entity shall be held liable when the fatal event derives from the violation of a duty imposed on the entity and from the way in which activities were conducted by the senior management. As for the Bribery Act (2010), instead, under section 7 a commercial organization shall be held liable for bribery committed by individuals operating in the interest of the company itself; however, the corporate defendant may prove that 'adequate procedures to prevent' bribery have been adopted before the commission of the crime.[30]

---

[27] This is also the interpretation given by the Italian Supreme Court: see Cass., Sez. Un., sent. 24 April 2014, n. 38343; for this position see, among scholars, Vincenzo Mongillo, *La responsabilità penale tra individuo ed ente collettivo* (Giappichelli 2018) 193 and 205. However, this is not a pacific standpoint, since according to a different interpretation the basis of Decree 231 regime is the identification principle (see, for instance, Giovanni Cocco, 'L'illecito degli enti dipendente da reato ed il ruolo dei modelli di prevenzione' (2004) Riv it dir proc pen 90.

[28] According to the second paragraph of article 5, the corporation will not be held liable if the individual acts only in its own interest.

[29] In these terms, see Ivan Salvadori, 'Agenti artificiali, opacità tecnologica e distribuzione della responsabilità penale' (2021) Riv it dir proc pen 106. Such an interpretation has been adopted by Italian courts with reference to the cases in which work accidents are caused by inadequate health and security measures adopted by the employer: see for instance: see see Cass., Sez. Un., sent. 24 April 2014, n. 38343 and, more recently, Cass., Sez. IV, sent. 8 January 2021, n. 32899. For comments, see Tomaso Emilio Epidendio, 'Criteri di attribuzione della responsabilità amministrativa' in Angelo Giarda *et al*. (eds.), *Responsabilità "penale" delle persone giuridiche* (Ipsoa 2007) 45.

[30] Both statutes apply to England and Wales, Northern Ireland, and Scotland. In this regard, it is worth noticing that the Bribery Act model has had a clear influence on the new tax evasion offence under section 45 of the Criminal Finances Act (2016), which replicates the same defence mechanism: see Wells (note 21) 273.

That said, in organizational fault regimes, the use of AI generally represents a double-edge sword: on the one hand, AI is a formidable instrument to ensure corporate compliance and, thus, to reduce the risk of wrongdoing;[31] on the other hand, the use of robots in certain circumstances may represent itself a risk and, consequently, lead to an organizational failure. In other words, the new technologies are both a risk and a way to manage it.

With regards to digital criminal compliance, it is easy to observe that big data can be considered not only a governance problem, but also a means to prevent crime; indeed, as companies get larger and more complex, attempting to predict possible risks, misconducts and damages become an extremely challenging task. Thanks to their computational and predictive capabilities, sophisticated AI software can analyse data much more accurately and efficiently than previously imagined[32]: as it has been observed, 'whether with respect to cyber security, organizational weaknesses, personnel inefficiencies, financial irregularities, or any area of corporate performance, AI software can develop targeted assessments of risk exposure that might impose costs or liabilities to the firm'.[33]

In the field of bribery prevention, for instance, big data analytics techniques, already used in public and private anti-corruption compliance, are progressively transforming the current features of risk assessment and risk management activities.[34] More specifically, through algorithms and AI software, it is possible to collect and compare internal and external data concerning the identification of anomaly and bribery risk index as well as the so-called red-flags (for instance: purchase prices, professional fees, conflicts of interest; etc.); the email traffic monitoring, with special focus on key words symptomatic of wrongdoing; the management reporting on red flags related to third parties,[35] and so on.

From this point of view, as far as the Italian regime is concerned, AI could represent a solution to the critical judicial scrutiny of the compliance program: in the field of anti-

[31] In these terms, see – in the recent Italian literature – Stefano Preziosi, 'Responsabilità da reato degli enti e intelligenza artificiale' (2020) 4 Resp amm soc enti 173; Paola Severino, 'Intelligenza artificiale e diritto penale' in Ugo Ruffolo (ed.), *Intelligenza artificiale. Il diritto, i diritti, l'etica* (Giuffrè Francis Lefebvre 2020) 536.

[32] See Sabia (note 6) 180. On the deployment of new technologies in corporate compliance strategies, see William S Laufer, 'The Missing Account of Progressive Corporate Criminal Law' (2017) 14(1) NYU J L B 87ff.

[33] Siebecker (note 24) 107.

[34] See William P Olsen, Dam Reynolds and Ales Kolston, 'Using Data Analytics to Meet the Government's Anti-Corruption Compliance Expectations' (2016) Anti-Corruption Report <https://www.anti-corruption.com/2568936/using-data-analytics-to-meet-the-government-s-anti-corruption-compliance-expectations.thtml> accessed 28 August 2021; Donna Daniels *et al.*, 'Real Risks, Artificial Intelligence: The Next Wave of Anti-Corruption Compliance?' (2018) Anti-Corruption Report <https://www.anti-corruption.com/2567091/real-risks-artificial-intelligence-the-next-wave-of-anti-corruption-compliance-.thtml> accessed 28 August 2021.

[35] Emanuele Birritteri (2019) 2 Dir Pen Cont 290.

corruption compliance, for instance, the adoption of certain technologies could be considered *ex lege* an adequate measure to exclude the organizational fault, thus avoiding judicial uncertainties.[36]

Furthermore, in the field of environmental compliance, public agencies have already started using AI in the fight against the so-called 'green-crime', thus enhancing regulatory effectiveness. Hence, the adoption of AI tools can help corporations developing compliance strategies to cope with environmental regulations and to avoid sanctions 'not only in term of "bureaucratic" requirements – eg reading complex legal documents and processing compliance content such as regulations, permits, policies, etc. – but also by intervening with reference to specific productions…with the effect of reducing the risk of violations of environmental law'.[37] Again, thanks to its extraordinary computational and predictive abilities, AI software can constantly monitor that eg air emissions or water discharges do not exceed the limits established by law.

Scholars have also argued that, from a more general perspective, the increased utilization of new technologies by corporate managers could not only revitalize the fiduciary bond between the latter and the corporation they serve, but also make corporate decision-making more attentive to the interests of corporate shareholders, stakeholders, and the community thus, paradoxically, more 'humane'.[38]

As far as the organizational failure is concerned, corporate fault could depend not only on the lack of internal procedures to assess the algorithmic output and possibly to take decisions departing from it, but also on the very choice to rely on AI systems.

Suffice it to mention – for the stake of brevity – the work-place related law: at present, numerous robots are already in use in workplaces throughout the world; thus, AI is becoming more and more integrated into the human workplace, completing tasks autonomously or even enhancing human performance and safety.[39] Not only do the use of robots enable physical and psychological relief for humans doing work which is dangerous, damaging to their health, or physically very strenuous for them (eg applications such as exoskeletons support employees in the physical performance of their work), but they also help to conduct monitoring activities in order to identify potential accident risks at the early planning stage for machinery and equipment, in order to avoid them.

However, these developments raise several problems.

---

[36] For this proposal, see Birritteri (note 35) 289.

[37] Sabia (note 6) 192.

[38] In these terms Siebecker (note 24) 95; the Author observes that 'the very proliferation of AI technologies facilitates easy adoption of a robust sense of "encapsulated trust" to refine and strengthen corporate fiduciary duties' (ibid. 114).

[39] See Isabelle Wildhaber, 'Artificial intelligence and robotics, the workplace, and workplace-related law' in Woodrow Barfield & Ugo Pagallo (eds.), *Research Handbook on the Law of Artificial Intelligence* (Edward Elgar Publishing 2018).

First of all, taking into account the Italian legislation, the employer must take all necessary measures which are able to avoid accidents at work and occupational illnesses and are applicable in conformity with experience and technological progress (article 2087 of the Italian Civil code). If these occupational safety and health measures are not taken, the employer will be liable due to a breach of the duty of care.

Thus, the first consequence is that AI is incapable of fulfilling such obligations since they are directed uniquely to the 'employer' individual person[40].

Furthermore, if the use of AI could be considered – in some cases – a necessary measure required by the technological progress (eg the use of wearable technology to detect risks), it has also to be considered that a vast body of occupational health and safety standards (such as those required for OSHA or ISO certifications) is designed to ensure that workers and machines operate separately.[41]

It is therefore essential to develop technical norms and standards, as well as a legal framework, that address the above-mentioned issues, in order to incentivize corporation to accelerate their embrace of automation and, at the same time, to promote compliance.[42] The reference is to so-called RegTech (from the merging of the words 'regulatory' and 'technology'), ie the use of technology in the context of regulatory monitoring, reporting and compliance. As the technological and regulatory fields continue to evolve, it will be of utmost importance to maintain cooperation between public agencies and firms employing AI in order to enable the latter to cut cost and meet their compliance obligations.[43]

3.2 Corporate culture

Finally, another holistic model that has to be mentioned is the liability for 'corporate culture': this formula makes reference to 'an attitude, policy, rule, course of conduct or practice within the corporate body generally or in the part of the body corporate where the offence occurred'.[44] Such a model has been adopted by the Australia's Criminal Code Act (1995), whose section 12 provides that, for offences of intention, knowledge, or recklessness, the 'fault element must be attributed to a body corporate that expressly, tacitly or

[40] See – in the Italian literature – Preziosi (note 31) 173.

[41] Dave Perkon, 'Educating OSHA Compliance Officers on Robot Safety' (2016) Control Design <https://www.controldesign.com/articles/2016/educating-osha-compliance-officers-on-robot-safety/> accessed 28 August 2021.

[42] For instance, the new generation of industrial robots called 'cobots', which collaborate with human individuals, present a challenge to the safety at work: they are subject to the provisions of ISO 10218 and Directive 2006/42/EC on machines. The ISO technical specification on collaborative robots (ISO/TS 15066:2016), published in 2016, introduces specific, data-driven safety rules to assess and control the risks related to cobots: since then, traditional barriers and safety measures to keep humans and robots apart are no longer necessary for certain industrial robots in conformity with ISO 10218.

[43] On the 'increasingly widespread search for a dialogue between public actors and business enterprises in order to foster an exchange regarding experiences and the dissemination of best practices' see Paola Severino, 'The Importance of Corporate Compliance in the Digital Era' (2020) 91(2) RIDP 435.

[44] Celia Wells, *Corporations and Criminal Responsibility* (Oxford University Press 2001) 137.

impliedly authorised or permitted the commission of the offence'. Authorization or permission can be shown in different ways: the first one echoes the identification liability, while the last ones are based on of corporate culture. Indeed, the corporation can be held liable if: *i)* 'a corporate culture existed within the body corporate that directed, encourage, tolerated or let to non-compliance with the relevant provision'; *ii)* the 'body corporate failed to create and maintain a corporate culture that requires compliance with the relevant provision'.

It is easy to observe that the same considerations expressed about the organizational fault model can be extended, *mutatis mutandis*, to the notion of corporate culture: while the use of AI to perform corporate function can be the result of attitudes, polices, practices or rules inside the collective entity, corporate blameworthiness may depend on the possibility of considering AI, under given circumstances, a risk factor rather than an instrument of managing it.

4 Concluding Remarks

Albeit the impact of AI on the criminal justice system has received significant attention, the several challenges posed by the interplay between the new technologies and corporate criminal liability seem still to be rather under-discussed. At a closer look, however, algorithmic corporate liability could represent an issue equally important and urgent to discuss in addition to AI accountability, since the main risks come from the use of new technologies by corporations.[45]

The issue is evidently too complex to be adequately addressed here; however, it is possible to conclude with some brief remarks, in the light of the considerations expressed so far.

First of all, it bears emphasizing that the question herein discussed is not AI responsibility, but rather the impact of the use of AI on corporate criminal liability: these two topics, although closely related, are different from each other.

Indeed, even if several scholars suggest that it is possible to understand the AI responsibility by using the notion of corporate criminal liability, other authors argue that there are differences between these two models. From this latter perspective, while it is theoretically possible to create an 'electronic personhood' (following the path already paved by the recognition of legal personhood of corporations),[46] and thus assuming the existence of a robotic *actus reus*, the conundrum is to identify an AI *mens rea* as well as the way

[45] For similar considerations see – in the Italian literature – Preziosi (note 31) 176 and Carlo Piergallini, 'Intelligenza artificiale: da "mezzo" ad "autore" del reato? Piergallini, 'Intelligenza artificiale: da "mezzo" ad "autore" del reato?' (2020) Riv it dir proc pen 1755ff.

[46] On this issue, see Robert van den Hoven van Genderen, 'Legal personhood in the age of artificially intelligent robots', in Woodrow Barfield & Ugo Pagallo (eds.), *Research Handbook on the Law of Artificial Intelligence* (Edward Elgar Publishing 2018) 213ff; Ugo Ruffolo, 'La "personalità robotica"' in Ugo Ruffolo (ed.), *Intelligenza artificiale. Il diritto, io diritto, l'etica* (Giuffrè Francis Lefebvre 2020) 213ff. In this regard, although in the field of civil liability, it is noteworthy the reference to the European Parliament resolution

to sanction the *homo technologicus*. According to the common standpoint, it is possible to fine corporations or limit their fundamental rights and freedoms (such as the property right and the freedom of economic activity), but AI lacks body and pocketbook; however, even such an assumption has been radically challenged in recent years.[47]

Focusing on algorithmic corporate liability, technological progress raises several legal issues that not always can be solved through an adaptive interpretation of the present legal framework[48]. An *ad hoc* regulation is therefore often required, which should take into account the multidisciplinary aspects of the field. In this regard, several Countries have already tested different forms of experimentation through lawfully de-regulated special zones, which represent the legal basis on which to collect empirical data and sufficient knowledge to make rational decisions for a number of critical issues.[49]

The principled approach aside, it is conclusively noteworthy that, from a pragmatic point of view, in the North American experience prosecutors may also address the need for criminal liability in cases of algorithmic corporate misconduct on their own by out-of-court deals, according to the guidelines provided by the US Attorneys' Manual. Instead of proving the conditions under which a company can be held responsible, the practical solution is offered by the possibility of concluding a pre-trial agreement with the corporate offender, which generally encompasses restoration for victims.[50] This solution,

---

with recommendations to the Commission on 'Civil Law Rules on Robotics' [2017] 2015/2103(INL), 2018/C 252/25, which calls on the Commission to consider the possibility of 'creating a specific legal status for robots in the long run, so that at least the most sophisticated autonomous robots could be established as having the status of electronic persons responsible for making good any damage they may cause, and possibly applying electronic personality to cases where robots make autonomous decisions or otherwise interact with third parties indipendently' <https://www.europarl.europa.eu/doceo/document/TA-8-2017-0051_EN.html#title1> accessed 28 august 2021.

[47] In this regard see generally Ryan Abbott and Alex Sarch, 'Punishing Artificial Intelligence: Legal Fiction or Science Fiction' (2019) 53 UC Davis L Rev 323; Vikram R Bhargava and Manuel Velasquez, 'Is Corporate Responsibility Relevant to Artificial Intelligence Responsibility?' (2019) 17 Geo JL & Pub Pol'y 829; Monika Simmler and Nora Markwalder, 'Guilty Robots? Rethinking the Nature of Culpability and Legal Personhood in the Age of Artificial Intelligente' (2019) Crim L Forum 9ff. More specifically, some author argues that, since AI technology has the capability of fulfilling the objective as well as the subjective requirements in criminal law, there is no reason why the general purposes of punishment and sentencing, ie retribution and deterrence, rehabilitation and incapacitation, down to capital penalty, cannot be applied to AI machines: in these terms, Hallevy (note 10) 185ff. From the same standpoint, arguing that *machina delinquere et puniri potest*, see Ugo Ruffolo, '*Machina delinquere potest?* Responsabilità ed "illeciti" (anche penali?) della "persona elettronica" e tutele per gli agenti software autonomi', in Ugo Ruffolo (ed.), *XXVI Lezioni di diritto dell'intelligenza artificiale* (Giappichelli 2021) 295-307. As far as Italian literature is concerned, skepticism about a preventive function of criminal law towards AI is expressed by Basile (note 10) 31ff; Salvadori (note 29) 98ff and Piergallini (note 45) 1770.

[48] Without considering that 'adaptive' interpretations raise the fundamental problem of their compatibility with the principle of legality, especially in civil law systems: such an issue is discussed, among other authors, by Pagallo and Quattrocolo (note 6) 390.

[49] See Pagallo and Quattrocolo (note 6) 407.

[50] See 'Principles of Federal Prosecution of Business Organizations' <https://www.justice.gov/jm/jm-9-28000-principles-federal-prosecution-business-organizations> accessed 28 August 2021, paras. 9-28.000ff. For related comments, see Brandon L Garret, 'International Corporate Prosecution' in Darryl K Brown,

which dates back to the '90s, when the first memorandum was published by the Department of Justice,[51] has been adopted over the years by other legal systems, such as England and Wales and France, in the international bribery field.[52]

References

Abbott R and Sarch A, 'Punishing Artificial Intelligence: Legal Fiction or Science Fiction' (2019) 53 UC Davis L Rev 323

Baricco A, *The Game* (Einaudi 2018)

Basile F, 'Artificial intelligence and criminal law: four possible research leads' (2019) DPU <https://archiviodpc.dirittopenaleuomo.org/upload/3089-basile2019.pdf> accessed 28 August 2021

Bhargava V R and Velasquez M, 'Is Corporate Responsibility Relevant to Artificial Intelligence Responsibility?' (2019) 17 Geo JL & Pub Pol'y 829

Birritteri E, 'Big Data Analytics and Anti-corruption Compliance. Critical Issues of Current Practice and Future Scenarios' (2019) 2 Dir Pen Cont 290

Burchard C, 'Is Artificial Intelligence putting an End to Criminal Law? On the Algorithmic Transformation of Society' (2019) Riv it dir proc pen 1909

Cingolani R, *L'altra specie* (Il Mulino 2019)

Cocco G, 'L'illecito degli enti dipendente da reato ed il ruolo dei modelli di prevenzione' (2004) Riv it dir proc pen 90

Daniels D *et al.*, 'Real Risks, Artificial Intelligence: The Next Wave of Anti-Corruption Compliance?' (2018) Anti-Corruption Report <https://www.anti-corruption.com/2567091/real-risks-artificial-intelligence-the-next-wave-of-anti-corruption-compliance-t.html> accessed 28 August 2021

Jenia Iontcheva Turner and Bettina Weisser (eds.), *The Oxford Handbook of Criminal Process* (Oxford University Press 2019); Federico Mazzacuva, *Corporate Liability and Diversion: Towards a New Criminal Law for Collective Entities?* (2018) 89(1) RIDP 159ff.

[51] US Department of Justice, 'Federal Prosecution of Corporations', Memorandum from E. Holder, Deputy Attorney General to Components Heads and US Attorneys, June 6, 1999 <https://www.justice.gov/sites/default/files/criminal-fraud/legacy/2010/04/11/charging-corps.PDF> accessed 28 August 2021.

[52] On the one hand, the Crime and Courts Act (2013), come into force in 2014, introduced deferred prosecution agreements in England and Wales; on the other hand, France's law on corruption (*loi n° 2016-1691 du 9 décembre 2016 relative à la transparence, à la lutte contre la corruption et à la modernisation de la vie économique*), the so-called *Sapin II* after Finance Minister Michel Sapin who presented the legislation, introduced the '*convention judiciaire d'intérêt public*' for collective entities. For a more general comparative perspective, see Tina Søreide, Abiola Makinwa (eds.), *Negotiated Settlements in Bribery Cases. A principled Approach* (Edward Elgar Publishing 2020).

Diamantis M E, 'The Extended Corporate Mind: When Corporations Use AI to Break the Law' (2020) 98 N C L Rev 894

Dyson G, *Turing's Cathedral. The Origins of the Digital Universe* (Penguin Books 2012)

Epidendio T E, 'Criteri di attribuzione della responsabilità amministrativa' in Angelo Giarda *et al*. (eds.), *Responsabilità "penale" delle persone giuridiche* (Ipsoa 2007) 45

Garapon A, *Justice digitale* (Presses Universitaires de France 2018)

—— *La despazializzazione della giustizia* (Mimesis edizioni 2021)

Garret B L, 'International Corporate Prosecution' in Darryl K Brown, Jenia Iontcheva Turner and Bettina Weisser (eds.), *The Oxford Handbook of Criminal Process* (Oxford University Press 2019)

Gobert J, 'Corporate criminality: four models of fault' (1994) 14 Legal Stud 393

Hallevy G, *Liability for Crimes Involving Artifical Intelligence Systems* (Springer 2015)

Laufer W S, *Corporate Bodies and Guilty Minds. The Failure of Corporate Criminal Liability* (The University of Chicago Press 2006)

—— 'The Missing Account of Progressive Corporate Criminal Law' (2017) 14(1) NYU J L B 71

Ligeti K and Tosza S, 'Challenges and Trends in Enforcing Economic and Financial Crime. Criminal Law and Alternatives in Europe and the US', in Katalin Ligeti and Stanisław Tosza (eds.), *White Collar Crime. A Comparative Perspective* (Hart Publishing 2019)

Mazzacuva F, *Corporate Liability and Diversion: Towards a New Criminal Law for Collective Entities?* (2018) 89(1) RIDP 159.

Mongillo V, *La responsabilità penale tra individuo ed ente collettivo* (Giappichelli 2018)

Olsen W P, Reynolds D and Ales Kolston A, 'Using Data Analytics to Meet the Government's Anti-Corruption Compliance Expectations' (2016), Anti-Corruption Report <https://www.anti-corruption.com/2568936/using-data-analytics-to-meet-the-government-s-anti-corruption-compliance-expectations.thtml> accessed 28 August 2021

Orland L, 'The Transformation of Corporate Criminal Law' (2006) 1 Brook J Cor Finn & Comm L 45

Pagallo U and Quattrocolo S, 'The impact of AI on criminal law, and its twofold procedures', in Woodrow Barfield & Ugo Pagallo (eds.), *Research Handbook on the Law of Artificial Intelligence* (Edward Elgar Publishing 2018)

Perkon D, 'Educating OSHA Compliance Officers on Robot Safety' (2016) Control Design <https://www.controldesign.com/articles/2016/educating-osha-compliance-officers-on-robot-safety/>

Piergallini C, 'Intelligenza artificiale: da "mezzo" ad "autore" del reato?' (2020) Riv it dir proc pen 1745

Preziosi S, 'Responsabilità da reato degli enti e intelligenza artificiale' (2020) 4 Resp amm soc enti 173

Quattrocolo S, *Artificial Intelligence, Computational Modelling and Criminal Proceedings. A Framework for A European Legal Discussion* (Springer 2020)

Ruffolo U, 'La "personalità robotica"' in Ugo Ruffolo (ed.), *Intelligenza artificiale. Il diritto, io diritto, l'etica* (Giuffrè Francis Lefebvre 2020)c

—— '*Machina delinquere potest?* Responsabilità ed "illeciti" (anche penali?) della "persona elettronica" e tutele per gli agenti software autonomi', in Ugo Ruffolo (ed.), *XXVI Lezioni di diritto dell'intelligenza artificiale* (Giappichelli 2021)

Sabia R, *Artificial Intelligence and Environmental Criminal Compliance* (2020) 91(1) RIDP 179

Salvadori I, 'Agenti artificiali, opacità tecnologica e distribuzione della responsabilità penale' (2021) Riv it dir proc pen 83

Selvaggi N, 'Criminal Liability of Corporations and Compliance Programs in the U.S. System' in Antonio Fiorella (ed.), *Corporate Criminal Liability and Compliance Programs*, vol. I, *Liability 'Ex Crimine' of Legal Entities in Members States* (Jovene editore 2012)

Severino P, 'Intelligenza artificiale e diritto penale' in Ugo Ruffolo (ed.), *Intelligenza artificiale. Il diritto, i diritti, l'etica* (Giuffrè Francis Lefebvre 2020)

—— 'The Importance of Corporate Compliance in the Digital Era' (2020) 91(2) RIDP 435

Siebecker M R, 'Making Corporations More Humane through Artificial Intelligence' (2019) 45 J Corp L 95

Simmler M and Markwalder N, 'Guilty Robots? Rethinking the Nature of Culpability and Legal Personhood in the Age of Artificial Intelligente' (2019) Crim L Forum 1

Søreide T and Makinwa A (eds.), *Negotiated Settlements in Bribery Cases. A principled Approach* (Edward Elgar Publishing 2020)

van den Hoven van Genderen R, 'Legal personhood in the age of artificially intelligent robots', in Woodrow Barfield & Ugo Pagallo (eds.), *Research Handbook on the Law of Artificial Intelligence* (Edward Elgar Publishing 2018)

Villani E, 'Compliance Programs and "Organisational fault" in Europe' in Antonio Fiorella (ed.), *Corporate Criminal Liability and Compliance Programs*, vol. II, *Towards a Common Model In The European Union*, (Jovene editore 2012)

Wells C, Wells, *Corporations and Criminal Responsibility* (Oxford University Press 2001)

—— 'Economic Crime in the UK. Corporate an Individual Liability' in Katalin Ligeti and Stanisław Tosza (eds.), *White Collar Crime. A Comparative Perspective* (Hart Publishing 2019)

Wildhaber I, 'Artificial intelligence and robotics, the workplace, and workplace-related law' in Woodrow Barfield and Ugo Pagallo (eds.), *Research Handbook on the Law of Artificial Intelligence* (Edward Elgar Publishing 2018)

THE CHALLENGES OF AI FOR TRANSNATIONAL CRIMINAL LAW: JURISDICTION AND COOPERATION

By Miguel João Costa[*] and António Manuel Abrantes[**]

Abstract

This paper addresses the challenges raised by artificial intelligence (AI) to jurisdiction and international cooperation in criminal matters. Regarding jurisdiction, problems lie mainly with positive conflicts. We submit that proper balance between individual rights and public interests depends largely on the adopted model of liability for crimes involving AI systems. As for cooperation, new challenges concern the dual criminality principle and the use of AI tools in criminal proceedings. We submit that fast-moving areas like AI require more flexible legal concepts, and that, in the present moment of undefinition, the political branch should be more active in cases of doubtful compliance with fundamental rights, by refusing cooperation where the judicial branch cannot.

1 Introduction[***]

At the plenary session that took place in Strasbourg at the end of 2019, the European Committee on Crime Problems (CPDC) set out its priorities for the present biennial, two of them being artificial intelligence (AI) and international judicial cooperation in criminal matters.[1] In contrast with AI-related legal challenges, international cooperation is – even if considered in only its modern shape (late 18th century) – an age-old normative area.[2] The reason why it remains so topical as to stand as a priority together with such forward-looking fields as AI law is that it shares a meaningful common denominator with them: technological development. It is not incidental that international cooperation, as we presently know it, was founded against the background of the industrial revolution and of the revolution in the means of transportation, which brought individuals, peoples, and

[*] Advisor, Portuguese Constitutional Court; Guest Lecturer and Integrated Researcher, Faculty of Law, UCILeR, University of Coimbra; Secretary of the Portuguese Group of the AIDP. For correspondence: <uc45139@uc.pt>.

[**] Advisor, Portuguese Constitutional Court; Guest Lecturer and Integrated Researcher, Portuguese Catholic University, Lisbon School of Law (Católica Research Centre for the Future of Law). For correspondence: <antonioabrantess@gmail.com>.

[***] This paper is based on research by the same authors published in Anabela Miranda Rodrigues (ed), *A Inteligência Artificial no Direito Penal* (Almedina 2020) 163-217.

[1] Council of Europe, European Committee on Crime Problems, *List of Decisions of the 77th Plenary Session* (2019) 2. The others were the protection of the environment through criminal law, smuggling of migrants and prison overcrowding.

[2] In fact, the oldest known legal instrument containing provisions on extradition is also the oldest known diplomatic document: the Treaty of Qadesh, from the 13th century BCE: see eg Paul Bernard, *Traité Théorique et Pratique de l'Extradition*, vol II (Arthur Rosseau Éditeur 1883) 31.

States closer together.[3] The latest outburst of innovation comprises the Internet of Things (IoT) and AI, and it is being regarded as the 4th industrial revolution.[4]

It has become common sense that this evolution has shuddered the traditional correspondence between sovereignty and territory, bringing into crisis one concept as well as the other. Not only did national sovereignties erode, but the part of them which did remain has also trouble imposing itself. States are more and more interdependent; they find it more and more difficult to enforce their own criminal laws even with regards to the acts that affect them the most (ie those committed in their territories or against their interests).[5] Challenging already in itself, this problem reaches new levels of difficulty when crossed with developments such as those brought by AI.

Jurisdiction too is inevitably shaken by such developments, notably by the emergence of a new 'space' with no physical existence. Crimes committed in (or rather, *through*)[6] cyberspace raise challenges for prescriptive, adjudicative and executive jurisdiction. Dematerialisation has been an issue for quite a while, but it too reaches new levels when two of the greatest inventions which humanity has ever produced – the Internet and AI – cross. Conflicts of jurisdiction have also been an issue for long, and now the criteria for solving them are becoming even blurrier by the coming into play of non-human actors. Moreover, we move slowly but surely towards a coexistence between humans and AI systems in the physical world itself, with self-driving cars and autonomous lethal weapons, *inter alia*, which also raises interesting jurisdictional issues.

The purpose of this paper is to take a glance at these developments from a transnational angle, to identify some of the main questions that they raise and to put forward some tentative answers. In the certainty that all that may be grasped at this point is only the tip of a huge iceberg.

[3] Although this was not the sole factor, as another relevant transformation was taking place in the same period: a transformation – described by Michel Foucault, *Discipline and Punish: The Birth of the Prison* (Vintage Books 1977) 7ff, 73ff, *passim* – of the very foundations of criminal justice, with the efficacy of punishment being no longer deemed to stem from its brutality but rather from its inevitability: see further Miguel João Costa, *Extradition Law: Reviewing Grounds for Refusal from the Classic Paradigm to Mutual Recognition and Beyond* (Brill | Nijhoff 2019) 319ff, 327ff.

[4] Klaus Schwab, 'The Fourth Industrial Revolution – What It Means and How to Respond' (2015) Foreign Aff. <www.foreignaffairs.com> accessed 25 August 2021.

[5] See Costa (n 3) 352ff; European Parliamentary Research Service (Wouter van Ballegooij), *European Arrest Warrant: European Implementation Assessment* (PE 642.839, June 2020) 63.

[6] See David da Silva Ramalho, *Métodos Ocultos de Investigação Criminal em Ambiente Digital* (Almedina 2017) 47.

2 Jurisdiction in Criminal Matters

The emergence of a new 'space' without physical borders set the concept of territoriality in crisis and raised problems of delimitation of jurisdiction among States.[7] With the development of AI, this digital space is expected to expand,[8] which projects on two different levels.

On the one hand, the crossing of AI and the Internet will likely increase pluri-localisation of offences – which in turn leads to two different problems: (i) the incorporation of the IoT in AI systems may render them vulnerable to hacker attacks that meddle in their programming in such a way as to induce them into committing crimes in the physical space; (ii) some AI systems (eg the so-called 'bots') may exist exclusively in the cyberspace, which may signify one single system, possibly through one single action, committing a crime that causes damages in numerous States at once.

On the other hand, the coexistence between human beings and AI systems magnifies offence pluri-localisation, in consequence of the *scission* between the place where the AI system is produced and that where it is utilised. This increases positive conflicts of territorial jurisdiction, which ought to be tackled both in the name of proper administration of justice and of such fundamental rights as *ne bis in idem*.[9] It is therefore necessary to determine the factors that should be given primacy, a question the answer to which varies depending on the model of criminal liability adopted.[10]

2.1 Jurisdiction in the 'natural probable consequence liability model'

This model is focused on the (negligent) liability of (i) the producer/programmer, or of (ii) the user of the AI system. The first variant relies on the most classical construction of

[7] See Michael A. Geist, 'Is there a there there? Toward Greater Certainty for Internet Jurisdiction' (2001) 16 Berkeley Technol. Law J. 8 f; Cedric Ryngaert, *Jurisdiction in International Law* (2nd edn, Oxford University Press 2015) 80; André Klip, 'International criminal law. Information society and penal law' (2014) 85 RIDP 401ff.

[8] See European Committee on Crime Problems (Sabine Gless), *Artificial Intelligence and its Impact on CPDC Work – The case of automated driving – Thematic session on Artificial Intelligence and Criminal Law* (2018), 2.

[9] See Pedro Caeiro, 'Jurisdiction in criminal matters in the EU: negative and positive conflicts, and beyond' (2010) 4 KritV 366; Klip (n 7); Frank Zimmermann, 'Conflicts of Criminal Jurisdiction in the European Union' (2015) 3 BJCL&CJ; Dominik Brodowski, 'Cybercrime, human rights and digital politics' in Ben Wagner and Matthias C. Kettemann and Kilian Vieth (eds), *Research Handbook on Human Rights and Digital Technology: Global Politics, Law and International Relations* (Edward Elgar Publishing 2019) 101ff.

[10] On these models, see eg Gabriel Hallevy, 'The Criminal Liability of Artificial Intelligence Entities' (2010) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1564096> accessed 25 August 2021, 8 f; Dafni Lima, 'Could AI agents be held criminally liable? Artificial Intelligence and the challenges for criminal law' (2018) 69 S. C. Law Rev. 689ff; Ugo Pagallo and Serena Quattrocolo, 'The impact of AI on criminal law, and its twofold procedures', in Woodrow Barfield and Ugo Pagallo (eds.), *Research Handbook on the Law of Artificial Intelligence* (Edward Elgar Publishing 2018) 309ff; Katalin Ligeti, 'Artificial Intelligence and Criminal Justice' (2019) <http://www.penal.org/en/information> accessed 25 August 2021, 3ff; Susana Aires de Sousa, '"Não fui eu, foi a máquina": Teoria do Crime, Responsabilidade e Inteligência Artificial', in Anabela Miranda Rodrigues (ed), *A Inteligência Artificial no Direito Penal* (Almedina 2020) 59ff.

criminal liability for defective production; the latter shifts liability so as to tend to the situations where the prohibited harm ultimately caused by the AI system results from the intermediate action (or omission) of the person utilising it. The resolution of positive jurisdictional conflicts in these cases will be a matter of preponderance in the responsibility for the result.

If the result is preponderantly attributable to the producer/programmer, the conflict will result from the mentioned scission (where it indeed exists) between place of production/programming and place of utilisation/materialisation of the damage. For instance, an autonomous vehicle produced and programmed in the USA causes the death of a pedestrian whilst circulating in Portugal, in consequence of failure by the producer to install a sensor indispensable for detecting pedestrians or of failure by the programmer to preordain the vehicle to break when detecting a pedestrian. In such cases, it seems that primacy in asserting jurisdiction should be given to the State in whose territory the action took place or should have taken place. Only the standards of precaution (ie the duties of care) established in the law of this State can be reasonably required from the individual. It is fundamentally by reference to them that the individual could have oriented his/her conduct and be blamed for having failed to do so properly. Considering that the product can, in the abstract, be exported to any State in the world, it would be excessive to require the producer/programmer to comply with as many criminal legislations (even assuming that they could have been identified beforehand with some certainty). In fact, the result might even materialise in a State other than that to which the vehicle had been exported: eg if the vehicle that had been exported to Portugal is utilised in Spain and the accident occurs there. The predictability of criminal punishment required by the principle of legality will tend to supersede the punitive interests of the State where the result occurred, even if the latter are rather intense.[11]

By the same token, if the result is mainly attributable to the user, then in principle no conflict of jurisdiction will emerge. Going back to the example, let us assume that the vehicle had an autonomy level of only 3 on the 1 to 5 scale of the Standard SAE J3016, and that it ran over and caused the death of the pedestrian as a result of a failure by the user to address a takeover request issued by the vehicle (eg because he/she had fallen asleep). In this instance, the place where the person acted or (as in this case) should have acted is also the place where the result occurred, and only that State will have territorial jurisdiction. In fact, the situation here has few or no differences with a traditional accident.

However, both scenarios presented above assume that it is instantly perceptible whether the 'fault' lies with the producer/programmer or, rather, with the user. In reality, however, this will often only be possible to ascertain in the course of a criminal procedure,

[11] The importance of predictability in positive conflicts of jurisdiction is emphasised by several authors, such as Uta Kohl, *Jurisdiction and the Internet – Regulatory Competence over Online Activity* (Cambridge University Press 2007) 141ff; Caeiro (n 9) 377; Klip (n 7) 401ff; Zimmermann (n 9) 16; Ryngaert (n 7) 80ff.

which would require that some State had already began to exert its adjudicative jurisdiction. In such situations – without detriment to the possibility that the States at issue have agreed or come to agree on a different solution –,[12] we would submit that primacy should in principle be given to the State in whose territory the *result* occurred, as this State is better placed to carry out the investigation as to whose fault it was, by examining the vehicle, hearing witnesses, etc. If, upon a preliminary investigation, it is concluded that the accident was due to a production/programming problem, then international cooperation may come into play, notably with the procedure being transferred to the latter State.

2.2 Jurisdiction in the 'perpetration by another model'

This model is based on the mediate or indirect responsibility of the person for the acts of the AI system, and it envisages to solve the situations where the latter is intentionally used by the former to commit a crime (eg where someone pre-sets a drone to kill a person). This is entirely similar to what already occurs in situations involving persons only.

In cases engaging this liability model, the fitting starting point seems to be instead that of giving primacy to the law of the State where the result occurs, at least where the utilisation of the AI system was from the outset specifically intended to harm or put at risk a legal interest held by a person or entity located in that given territory. The difference of approach in relation to the previous model is due to the fact that, whereas negligent offences presuppose the breach of a duty of care which is immanent to the conduct, intentional offences presuppose the will (or at least the acquiescence) to cause a given result. Therefore, the level of predictability required by the legality principle will be more easily met in these cases.[13]

However, the solution should differ where the utilisation of the AI system is aimed at damaging or putting at risk a legal interest held by an undetermined set of individuals and/or dispersed by numerous places. For instance: a programmer/producer of robots designed to perform domestic tasks manipulates an algorithm so that the robots kill their owners sometime upon being activated; the robots are exported worldwide and, in the set date, carry out the wave of homicides. In such a case, when acting, the programmer does not have in mind specific human beings or aims for the result to materialise in specific places, such that the legality principle and the predictability it entails will again bear more weight.

But the question will increase in complexity if, in addition to the result occurring in numerous places, the place of the action itself is undetermined. For instance, when a hacker located in an uncertain place accesses through cyberspace the operating system of a set of robots with the same algorithm and, as an act of rebellion, instructs them to cause disturbance in public spaces; or where a programmer creates a bot to fish data on a massive scale. In both cases, applying the law of the place of the conduct will be unviable, as

[12] See the end of § 2.2., below.
[13] See Klip (n 7) 401ff.

it would require prior identification of the territory where the hacker or programmer acted, which tends to be extraordinarily difficult in such cases. The only possibility that appears to remain would be to apply concurrently the laws of all the States where the results occurred (whose punitive interests will tend to be identical), but this will still raise positive conflicts of jurisdiction of difficult resolution, because in the light of the principle of ubiquity each of those States will have territorial jurisdiction, although only over the part of the facts that materialised in their territory. This will likely lead to a multiplication of proceedings, in detriment both of proper administration of justice and of fundamental rights, in the terms already mentioned. So diffused are the consequences of such a conduct that the only way to adequately address a case like this would be through a specific supranational response (such as the creation of a supranational judicial body), or through peculiar trials that are formally national, but which have, in fact, prominent international features (as is, to some extent, the case of the trial conducted in the Netherlands for crimes related to the downing of Malaysia Airlines flight MH17).[14]

2.3 Jurisdiction in the 'direct liability model'

Finally, there is the direct liability model, which conceives of AI systems themselves as criminally liable and proposes the application of criminal penalties to them. At first glance, this model would not seem to raise positive conflicts of jurisdiction different from the traditional ones. Since this model assumes that AI systems have reached an evolution stage that allows for an analogy with human beings, no specificities would seem to exist. Thus, if for instance a completely autonomous vehicle were to negligently run over a pedestrian in Spain, its liability would be ascertained in Spain, as this is where the conduct took place and the harm materialised. As for cases of pluri-localisation, they too would follow the rules already in place for human beings.

However, to think that this model would not raise specific problems would be a hasty conclusion. Even if AI systems do develop to such an extent as to warrant comparison with humans (eg in terms of their capacity to process information and take decisions), that does not mean that they would be perfectly *identical*: on the one hand, there are human capacities that they may never replicate; on the other hand, they may have capacities that humans lack. One issue stems from the interaction between AI and the Internet, which makes it possible for an AI system to exist exclusively in non-corporeal terms dispersed by a plurality of devices connected to a network – an existence which, in this sense, is in itself ubiquitous –, as is already the case with blockchain technology: suppose that one such AI system, which had been conceived for detecting threats to the network, decides – following its learning process – to destroy every information system that it

[14] See Agreement between the Kingdom of the Netherlands and Ukraine on International Legal Cooperation regarding Crimes connected with the Downing of Malaysia Airlines Flight MH17 on 17 July 2014, signed in Tallinn on 7 July 2017.

classifies as a threat, thereby causing harm to users across the globe.[15] Another issue relates to the interconnection between different AI systems in the digital space: imagine a scenario where all transportation services are provided by AI systems capable of taking decisions without human intervention, and that an AI system simultaneously manipulates all other systems in such a way as to cause widespread car collisions, train derails, plane crashes.[16]

Such dystopian scenarios lend strength to the view that, in the same manner as these new forms of criminality, so too adjudicative jurisdiction should be *dispersed*, exerted conjointly in a wide multilateral manner. We are quite aware that this may seem farfetched at present, but at some point in time a supranational jurisdiction may prove necessary to tackle developments associated with AI. In an era where vehicles drive themselves, bots converse and robots take decisions, changes once thought remote may arrive suddenly.

3 International Cooperation in Criminal Matters

Through its different modalities – eg extradition, transfer of proceedings, enforcement of foreign sentences, transfer of prisoners, mutual legal assistance –, international cooperation may permeate the different moments of criminal law, from prevention to investigation, from trial to enforcement. It follows that, within the context of AI, many of the issues that are being identified as emerging challenges for criminal law in a stricter sense constitute also challenges for international cooperation in criminal matters.

3.1 New types of criminality and the principle of dual criminality

The first example of that concerns the emergence of new criminal offences. Indeed, AI has not only great potential to be used in the commission of traditional offences (eg the use of autonomous drones in trafficking activities), but its development and widespread use are also expected to lead to the enactment of truly new criminal offences.[17] For instance, attacking or disturbing the regular functioning of an AI system or its interaction with a human being in such a way as to cause damage;[18] or sabotaging an AI system in such a way as to prevent it from accomplishing the goals for which it was conceived or to carry it to commit a crime[19] (in which case the AI system will simultaneously be the 'victim' and the instrument of the crime).

---

[15] This example is based on a hypothetical case presented by Hallevy, 'The Criminal Liability of Artificial Intelligence Entities' (n 10) 16, in respect of the limitations of the models that rely exclusively on the responsibility of the producer/programmer.

[16] This scenario is also relevant for consideration within the context of the 'perpetration by another model', if the interference with the AI systems is carried out by a human (eg, for a terrorist attack).

[17] See eg Interpol / United Nations Interregional Crime and Justice Research Institute (UNICRI), *Artificial Intelligence and Robotics for Law Enforcement* (2019) 23.

[18] See European Committee on Crime Problems (Sabine Gless), *Assessment of the Answers to the Questionnaire on Artificial Intelligence and Criminal Justice (using the example of Automated Driving)* (2019) 10.

[19] See Ligeti (n 10) 5ff.

More generally, AI systems can be instrumentalised for a number of problematic ends of different sorts, namely digital, physical and political.[20] A noteworthy point is that this includes behaviour that, if taken in an isolated manner, does not generally constitute a crime, but which, if associated with AI, gains a magnitude that may already justify criminalisation.[21] These are cases where the quantitative modification is so great that it translates into a qualitative modification. A clear example is the 'weaponisation of information',[22] eg through the production of fake news and the automation of influence campaigns.[23]

In sum, several types of conduct are being flagged as problematic and many others will plausibly follow, which seems to have already crystalised into a clear-cut criminal policy premise: there is a punitive gap opened by the coming into play of AI.[24] It is therefore likely that a criminalisation movement will ensue. At the transnational level this has direct implications for dual criminality, one of the most classic principles of international cooperation, according to which a State will only agree to cooperate with another State for acts that it also criminalises.[25]

States will likely conclude international treaties on AI-related aspects, which will create some regulatory common-ground (as already occurred with cybercrime, for instance).[26] Yet, if even in respect of the most traditional criminal offences there is significant discrepancy among States in the criminalisation of behaviour – which is but a natural expression of criminal law's eminently localised character, in such a new and dynamic context as AI, it is almost inevitable that such a disparity will be significant.[27] Dual criminality will pose a problem in that it will prevent cooperation in criminal matters. The archetypical situation is that where a person commits a crime in a given State and flees to a State where the acts were not criminalised; or the person acted in the latter State, but the result materialised in the former State, into which the person never entered.

---

[20] Miles Brundage and Sahar Avin *et al*., 'The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation' (2018) <https://doi.org/10.17863/CAM.22520> accessed 25 August 2021, 23ff.

[21] See Ligeti (n 10) 6.

[22] See eg Shannon Raj Singh, 'Move fast and break societies: the weaponisation of social media and options for accountability under international criminal law' (2019) 8 Camb. Int. Law J. 331ff.

[23] Brundage and Avin *et al.* (n 20) 28.

[24] See eg European Committee on Crime Problems (n 18); Matthijs M. Maas, 'International Law does not Compute: Artificial Intelligence and the Development, Displacement or Destruction of the Global Legal Order' (2019) 20 Melb. J. Int. Law 33, 39ff.

[25] The dual criminality principle also manifests itself at the jurisdictional level, as a requirement for the exercise of certain bases of extraterritorial jurisdiction: see eg Christine van den Wyngaert, 'Double criminality as a requirement to jurisdiction', John Dugard and Christine van den Wyngaert (eds), *International Criminal Law and Procedure* (Dartmouth Pub. Co. 1996); Pedro Caeiro, *Fundamento, Conteúdo e Limites da Jurisdição Penal do Estado. O Caso Português* (Coimbra Editora | Wolters Kluwer 2010) 355ff.

[26] See, notably, the 2001 CoE Convention on Cybercrime.

[27] Which is further heightened by the possible disparities in the different models of imputation that come to be adopted in each State: see *supra*, § 2.

One must ask, first of all, whether this should really be regarded as a *problem*. There certainly are plausible reasons to uphold that the 'universalisation of local punitive claims' is not something to be lightly accepted; that a State should not readily assist other States in fulfilling their criminal policy programmes if it does not itself criminalise the acts.[28] It may even be held that, if the conduct is not criminalised in that State, then one cannot even properly speak of cooperation in *criminal* matters.[29] In sum, that these regulatory discrepancies may be accepted as natural.

Nevertheless, we think that the question does constitute a criminal law problem: the fact that the conduct is criminalised in the State in whose territory it was carried out is, in the abstract, sufficient to justify cooperation on the part of other States, because the basis for jurisdiction of undisputed primacy is,[30] in the view of *all States*, territoriality.[31]

Moreover, reluctance to relinquish dual criminality is more justified in relation to traditional types of behaviour. In these cases, penal regulation will tend to reflect more consolidated views of each community as to what is and what is not worthy of criminalisation; such that non-criminalisation may be understood as a space of freedom unequivocally desired by that community, rather than as an unintended punitive gap. In contrast, new types of activity raise other sorts of considerations, especially when such activities are particularly dynamic and impactful,[32] as is evidently the case with AI, as well as with activities that are damaging to environment (which, as noted earlier, is precisely one of the other priorities identified by the CDPC). In fact, this is not the sole common denominator between AI and the environment in the light of criminal law:[33] both involve activities that can be extremely beneficial for humankind, but which, if certain parameters are not respected, can turn out to be extremely detrimental.[34] In contexts like these, the stakes

---

[28] Pedro Caeiro, 'Editorial of the dossier "International Judicial Cooperation in Criminal Matters" – Current problems in a global perspective' (2019) 5 RBDPP 557.

[29] See Bert Swart, 'Human Rights and the Abolition of Traditional Principles', in Albin Eser and Otto Lagodny (eds), *Principles and Procedures for a New Transnational Criminal Law* (Max Planck Institut 1992) 253; Jorge de Figueiredo Dias and Pedro Caeiro, 'Comentário ao Acórdão do Tribunal de Justiça de 3 de Maio de 2007, proc. C-303/ 05, Advocaten voor de Wereld VZW contra Leden van de Ministerraad', in Eduardo Paz Ferreira and Maria Luísa Duarte and Miguel Sousa Ferro (eds), *Jurisprudência Cunha Rodrigues – Comentários* (Associação Académica da Faculdade de Direito de Lisboa Editora 2013) 28.

[30] Leaving aside the specificities of protective jurisdiction, which for the sake of simplicity will not be addressed here.

[31] See Costa (n 3) 481.

[32] See again Brundage and Avin *et al.* (n 20) *passim*; also eg United Nations (UN) High-level Panel on Digital Cooperation (Luke Kemp et al.), *A Proposal for International AI Governance* (2019) 1ff; European Parliament, Committee on Civil Liberties, Justice and Home Affairs (Tudor Ciuhodaru), *Draft Report on artificial intelligence in criminal law and its use by the police and judicial authorities in criminal matters* (2020/2016(INI)) 5-6 (presented and adopted by the Committee on Civil Liberties, Justice and Home Affairs on 29 June 2021), where attention is also drawn to '*the power assymmetry between those who develop and employ AI tecnhologies and those who interact and are subject to them*'.

[33] It is therefore unsurprising that the UN High-level Panel on Digital Cooperation (n 32) 2, suggests adapting certain environmental law principles to the realm of AI.

[34] An example of this within the AI thematic is the use of autonomous lethal weapons (ALWs) in military context. While many propose a complete ban, an analysis of the laws of war (*jus in bello*) suggests that,

are much higher than usual. Both the advantages of allowing for the development of certain activities and the damages possibly arising therefrom are heftier. Hence, even those who in line of principle argue in favour of keeping dual criminality must at least give it a second thought inasmuch as such realms of activity are concerned.

Of course, for substantive criminal law, such a differentiation – between 'deliberate' and 'unintended' non-criminalisation – is immaterial. If the conduct was not criminalised in a given State, such a State cannot punish it. Moreover, if that is the only State with territorial jurisdiction, the conduct should not be punished elsewhere either; otherwise one would be admitting an interference with the sovereignty of that State (for a State is entitled to have other States refrain from prosecuting acts committed in its territory that were not criminalised there) and an intolerable disruption of the principle of legality (for if the acts were not criminalised in the territory where they were committed the person could not expect to be prosecuted elsewhere). This is unequivocal and cannot be stressed enough.

But things are pointedly different when it comes to international cooperation. If the acts did constitute a crime in the State where they were committed, then the individual was or can legitimately be expected to have been aware of that, with the consequence that there is no disruption of the legality principle. And if that State is requesting cooperation, then obviously there is no problem of sovereignty meddling either. This is why, in the abstract, it is not illegitimate to cooperate in criminal matters for acts that one does not criminalise.[35] As noted before, in highly dynamic areas such as AI, the fact that a given conduct is not criminalised might result from such casual reasons as a delay on the part of a State in becoming aware of the harmful character of that conduct or a delay in legislating accordingly. And in these highly dynamic areas, certain types of conduct might be so harmful that it becomes untenable to accept impunity. Impunity could only find justification in the imperative of legal certainty, but legal certainty is not really affected in these instances, because the conduct *was* criminalised in the place where it was committed and the person knew it or should have known it.

If one accepts this idea, then the principle of dual criminality should be mitigated. In our view, it should not be abandoned *entirely*. A pure abolition of dual criminality would not pay sufficient respect to the different conceptions of criminal justice of the different

---

within certain parameters, this normative area not only does not preclude that use, but in fact might even be invoked in support of it. This is because this normative area is fundamentally concerned with minimising the damages arising from war – a phenomenon that it does not regard as acceptable but which it assumes as inevitable –, and ALWs, due to their precision, offer non-negligible perspectives of reduction of civil casualties and of other war-related calamities: see Miguel Lemos and Miguel João Costa, 'Inteligência Artificial e Direito da Guerra: Reflexões sobre as Armas Autónomas Mortíferas', in *Revista Julgar – Special Issue: O Direito na Era da Inteligência Artificial*, forthcoming, concluding that, therefore, if reasons do exist for a complete ban of ALWs, they should be sought not in the laws of war, but elsewhere, such as moral philosophy and philosophy of science, political science and international politics.

[35] As ruled for instance by the CJEU in *Advocaten voor de Wereld VZW v. Leden van de Ministrraad* (C- 303/05), Judgment of 3 May 2007, esp. paras 55ff.

States,[36] which would thereby be cooperating even for acts that they do not criminalise out of principle. The road should instead be that of *reforming* dual criminality, complementing it with a more flexible concept such as that of the 'qualitative relevance of the acts'.[37] Cooperation would no longer depend on the conduct being criminalised in the requested State. If it is, then obviously it may cooperate. But even if it is not, it might still cooperate, unless the acts at issue, in addition to not being criminalised, *could not* possibly be criminalised in view of its Constitution, for instance because they correspond to the exercise of a constitutionally enshrined fundamental individual right.

The alternative, as mentioned, is to accept the punitive gap. This is always the core predicament underlying discussions on interstate cooperation and on how far it should be taken. And both positions can be defended. However, the relative weight of the arguments varies depending on the unpredictability and on the potential harmfulness of the activity in question, both of which are apparently quite elevated in the case AI-related activity.

3.2 Algorithmic justice and fundamental rights

The second constellation of challenges relates to the use of AI in a predictive or prognostic manner. Once again, the diagnosis seems to be that this type of use opens perspectives of great efficiency but also great risks. Nobody doubts the potential advantages of AI. The question is whether they outweigh the potential disadvantages and, at any rate, whether they are really worth the risk.[38]

(i) To begin with, this predictive approach is being applied at a *pre-procedural stage*, with AI tools of 'predictive policing'[39] which process large quantities of data in order to anticipate the very commitment of crimes. An example is 'Connect', used in the UK in the analysis of financial transactions; also 'International Child Sexual Exploitation', managed by Interpol.[40] (ii) Secondly, it can be applied *in the criminal procedure*, notably for evidence gathering[41] or in the application of measures that require some sort of prognostic assessment, such as coercive measures aimed at obviating the escape of the suspect, a risk that AI may help calculating.[42] Which *a fortiori* is relevant also for international cooperation

---

[36] Ie to the different cultural and legal 'biotopes', to borrow an expression from Jürgen Habermas, *The Crisis of the European Union – A Response* (Polity 2012) 42, 53.

[37] This is the approach proposed in Costa (n 3) 449-497.

[38] See eg European Parliament, Panel for the Future of Science and Technology, *The impact of the General Data Protection Regulation (GDPR) on artificial intelligence* (2020), 27, questioning: '*Should we just ask (…) whether these systems provide reliable assessments, or should we rather ask whether they should be built at all[?]*'.

[39] See eg Ligeti (n 10) 7ff; European Parliament, Panel for the Future of Science and Technology (n 38) 39; Council of Europe, European Commission for the Efficiency of Justice, *European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their environment* (2018) 49.

[40] See Council of Europe, European Commission for the Efficiency of Justice (n 39) 50ff.

[41] See Sónia Fidalgo, 'A utilização de inteligência artificial no âmbito da prova digital – direitos fundamentais (ainda mais) em perigo' in Anabela Miranda Rodrigues (ed), *A Inteligência Artificial no Direito Penal* (Almedina 2020) 129-62.

[42] See eg ibid.

proceedings, notably extradition, where a risk of escape is often met.[43] (iii) At the *sentencing* level there is also room for deploying such tools – examples of which are 'COMPAS' (*Correctional Offender Management Profiling for Alternative Sanctions*) and 'HARM' (*Harm Assessment Risk Tool*) –,[44] namely in evaluating the likelihood of a person reoffending or in deciding whether to apply a custodial rather than a non-custodial sanction; or yet in evaluating the dangerousness of a defendant who suffers from a mental illness. More generally, this is also where the whole discussion on 'judge-robots' in criminal proceedings fits.[45] Which again is also relevant for cooperation proceedings, with humans being potentially replaced by intelligent robots in receiving and deciding requests issued by other States.[46] (iv) Finally, AI may be used at the *enforcement* level, for instance to help evaluate whether or not a person should be granted early release or a leave licence from prison, which as well involve prospective assessments such as the risk of escape or the likelihood of the convict behaving correctly upon release.[47] And also in the surveillance of inmates, with a view to preventing an escape or a crime being committed within the facility.[48]

At the High-Level Conference co-organised by the Council of Europe and Finland's Presidency of the Council of the EU, which took place in Helsinki in February 2019, the following proclamation was made:

> Existing landmark international instruments, including the Universal Declaration of Human Rights and the European Convention for the Protection of Human Rights and Fundamental Freedoms, are applicable irrespective of contextual changes brought about by AI.[49]

The statement seems redundant at first glance. Why would they not be applicable? But one characteristic of AI's impact upon fundamental rights is that it is *discrete*.[50] Thus, reiterating the obvious might not be so superfluous after all. Perhaps many of the forms of utilisation of AI mentioned above can indeed provide good service to the administration of criminal justice, but it is very clear that they jeopardise fundamental rights, notably:[51]

---

[43] See Michael Vien, 'Provisional Arrest, Release and the Role of Interpol' (1991) 62 RIDP 63ff.

[44] See eg Council of Europe, European Commission for the Efficiency of Justice (n 39) 51.

[45] See eg Anabela Miranda Rodrigues, 'Medida da pena de prisão – desafios na era da Inteligência Artificial' 149 RLJ (2020) 265ff; Anabela Miranda Rodrigues, 'A questão da pena e a decisão do juiz – entre a dogmática e o algoritmo', in Anabela Miranda Rodrigues (ed), *A Inteligência Artificial no Direito Penal* (Almedina 2020) 219-44.

[46] See Ligeti (n 10) 13; Interpol / United Nations Interregional Crime and Justice Research Institute (UNICRI) (n 17) *vi*.

[47] See eg European Parliament, Panel for the Future of Science and Technology (n 38) 15, 21.

[48] See Aleš Završnik, 'Criminal justice, artificial intelligence systems, and human rights' (2020) 20 ERA Forum 572ff.

[49] Council of Europe, *Governing the Game Changer – Impacts of artificial intelligence development on human rights, democracy and the rule of law: Conclusions from the Conference*» (Helsinki, 26-27 February 2019).

[50] On this characteristic, see Maas (n 24) 51ff.

[51] See UN High-level Panel on Digital Cooperation (n 32) 1ff; Council of Europe, European Commission for the Efficiency of Justice (n 39) 49ff; European Committee on Crime Problems (Sabine Gless*), Working*

(i) equality and prohibition of discrimination (eg because those predictions are based on samples and do not, therefore, take sufficient stock of the singularising elements of the subject); (ii) privacy (as AI enables an extreme monitoring of nearly all aspects of life); (iii) defence rights and fair trial (for numerous reasons, but for instance because of the opacity of AI systems, a problem which is further magnified by the fact that some systems are owned by private companies which are reluctant to disclose details about their programming; and even if the system as such is not unfair, because of the disproportionality it introduces between the capacity of the prosecution and that of the defence); (iv) right to appeal (for obviously one cannot appeal against a decision the rationale of which is not really clear); (v) and, more transversally, as transversal are these principles themselves, presumption of innocence[52] and judicial independence.[53] Some of these problems, such as the opacity of the algorithm, might be possible to mitigate.[54] Others, such as the use of non-individualised risk assessments for sentencing and early release purposes, are not even entirely new.[55] However, it seems clear that the present is marked by uncertainty and undefinition. At the legislative level the first steps are still being taken; case law is still scant.

From the transnational criminal law standpoint, the scenario described above unfolds into two main sets of problems. The first one concerns the use of AI in the *decision-making process* of cooperation proceedings. Such a use is already quite questionable in the context of actual criminal proceedings,[56] and in our view it is even more so in that of cooperation proceedings. Indeed, although criminal proceedings carry more profound limitations of fundamental individual rights, they operate within a normative framework which, for that very reason, is bound to offer very high levels of legal certainty. In consequence of the legality principle, the norms applied in criminal proceedings must necessarily rely on concepts that are fairly unambiguous. In contrast, in the context of transnational pro-

Group of Experts on Artificial Intelligence and Criminal Law – Working Paper II, 1st meeting (2019); European Parliament, Panel for the Future of Science and Technology (n 38) 5, 21, 48ff; Interpol / United Nations Interregional Crime and Justice Research Institute (UNICRI), (n 17) 36ff. See also Brundage and Avin *et al.* (n 20); Ligeti (n 10); Rodrigues, 'Medida da pena de prisão – desafios na era da Inteligência Artificial' (n 45) 265ff; Završnik (n 48) 574ff; Sabine Gless, 'AI in the Courtroom: A Comparative Analysis of Machine Evidence in Criminal Trials' (2020) 51 Georget. J. Int. Law 195ff.

[52] See Mireille Hildebrandt, 'Criminal Law and Technology in a Data-Driven Society', in Markus D. Dubber and Tatjana Hörnle (eds), *The Oxford Handbook of Criminal Law (*Oxford University Press 2014) 195ff, proposing a 'presumption of innocence by design'.

[53] See Council of Europe, European Commission for the Efficiency of Justice (n 39) 50ff, 56ff, alluding to a 'tyranny of the algorithm'.

[54] See Hildebrandt (n 52) 195ff; Gless (n 51) 250ff; Ligeti (n 10) 12. See also Council of Europe, European Commission for the Efficiency of Justice (n 39) 55, emphasising the difference between Europe and the USA in the regime of access to the algorithm, a difference which stems largely from the protective rules introduced by the General Data Protection Regulation – Regulation (EU) 2016/679 (see especially Article 15 (1) (h)).

[55] See Rodrigues, 'Medida da pena de prisão – desafios na era da Inteligência Artificial' (n 45) 265.

[56] See ibid 270ff; also Luís Greco, *A Impossibilidade Jurídica de Juízes Robôs* (conference, Instituto Eduardo Correia 2020 <http://www.institutoeduardocorreia.com.br/videos/2020> accessed 25 August 2021).

ceedings, the erosion of geographic frontiers (and of national sovereignties thereby) requires increasingly undetermined rules based on highly subjective concepts, because only concepts of this character are capable of grasping the situations in which interstate cooperation is *reasonable*. A clear example of this is precisely the above proposed reform of the principle of dual criminality into a principle of 'qualitative relevance of the acts'. Profoundly and intentionally subjective, concepts like this can hardly be computed by a machine. Thus, generally speaking, we believe that the use of AI in the decision of cooperation requests is even more unfitting than in criminal proceedings proper.

The other question which is transversal to the foregoing analysis is whether it should be possible for a State to cooperate with another State whose criminal procedure is making use of AI systems that raise *doubts of compliance with fundamental individual rights*. For instance, an extradition request issued on account of a criminal procedure where the person was convicted by a judge-robot; or by a human judge, but where nearly all the evidence was gathered by opaque AI systems; or where the person was convicted to a custodial penalty, rather than to a non-custodial penalty (which does not usually admit extradition), based on a risk assessment performed by a system like COMPAS; or the transfer of a prisoner who will likely be placed in a facility with extreme monitoring performed by an AI system. Just to mention a few examples.

In the case of States that already make use of such AI systems, the problem does not really arise: if these States use such systems in their own criminal proceedings, then in principle they will not have a fundamental objection to cooperating with a State that does so as well. However, most States do not yet make such a use of AI, and these States might then have to take a stance on the whole problem not for internal purposes, but for the purposes of deciding whether or not to cooperate. With the additional difficulty that, thus far, regional and international judicial bodies have barely had the chance to deliver case law on the use of these AI systems and can therefore not offer guidance. Most treaties and national statutes on international cooperation do contain rules preventing cooperation if the procedure/enforcement in the requesting State does not attain a certain threshold of respect for fundamental rights. However, as noted, it is still unclear to what extent fundamental rights are indeed breached by the use of such AI systems, and in any case refusal of cooperation on fair trial grounds normally requires a *flagrant* denial of justice in the requesting State.[57]

Therefore, in many cases, even if there are meaningful doubts as to the fairness of the proceedings in the requesting State, no legally prescribed grounds for refusal of cooperation will be met, with the consequence that it will not be possible for the courts of the requested State to refuse cooperation. In our view, this constitutes a particular moment where the intervention of the executive branch in cooperation proceedings should be reinvigorated. Indeed, it is curious to note that, historically, cooperation proceedings shifted from purely political to mixed proceedings in order to ensure adequate protection

[57] See Costa (n 3) 83ff and 108ff.

of fundamental rights, and in some cases (as in the EAW) they even shifted to fully judicial proceedings for the sake of expeditiousness and other interests, which ironically compressed again the protection of the individual against the punitive interests of the requesting State.[58] This is not to say that in particular contexts (as in the EU) this is not justified. However, at least regarding certain factual and normative developments, notably in fast-moving areas like AI, the intervention of the executive – with all its flexibility, and albeit at the price of some arbitrariness – may be the only way to compensate for the fact that the judiciary is bound to apply laws that inexorably develop at a much slower pace.

4 Conclusion

AI brings new challenges to jurisdiction in criminal matters, especially inasmuch as it concerns positive conflicts of territorial jurisdiction, and States should be cautious in exerting their jurisdiction, so as to secure fundamental rights whilst pursuing the public interest in order to ensure a proper administration of justice. Complex new jurisdictional conflicts arise, the solution to which depends heavily on the approach followed by the different States to the liability for crimes involving the use of AI. The dematerialisation of both the criminal conduct and the ensuing damage also cause serious difficulties, and a supranational judicial instance might be necessary at some point.

AI brings new challenges also to international cooperation, notably insofar as it concerns the dual criminality principle and the use of AI tools in criminal proceedings in the requesting State. Regarding the former, the view has been conveyed that more flexible concepts than that of dual criminality could be used in order to accompany the highly dynamic character of these developments. As for the latter, the present moment is marked by undefinition as to the extent to which fundamental rights are being jeopardised, which leads to the proposition that the political branch should take on a more dominant role in cases of questionable fairness by refusing cooperation where the judicial branch cannot.

References

Bernard P, *Traité Théorique et Pratique de l'Extradition*, vol II (Arthur Rosseau Éditeur 1883)

Brodowski D, 'Cybercrime, human rights and digital politics' in Ben Wagner and Matthias C. Kettemann and Kilian Vieth (eds), *Research Handbook on Human Rights and Digital Technology: Global Politics, Law and International Relations* (Edward Elgar Publishing 2019)

Brundage M and Avin S *et al*., 'The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation' (2018) <https://doi.org/10.17863/CAM.22520> accessed 25 August 2021

[58] ibid 384ff.

Caeiro P, 'Editorial of the dossier "International Judicial Cooperation in Criminal Matters" – Current problems in a global perspective' (2019) 5 RBDPP 553

—— 'Jurisdiction in criminal matters in the EU: negative and positive conflicts, and beyond' (2010) 4 KritV 366

—— *Fundamento, Conteúdo e Limites da Jurisdição Penal do Estado. O Caso Português* (Coimbra Editora | Wolters Kluwer 2010)

Costa M J, *Extradition Law: Reviewing Grounds for Refusal from the Classic Paradigm to Mutual Recognition and Beyond* (Brill | Nijhoff 2019)

Council of Europe, European Committee on Crime Problems, *List of Decisions of the 77th Plenary Session* (2019)

—— (Sabine Gless), *Assessment of the Answers to the Questionnaire on Artificial Intelligence and Criminal Justice (using the example of Automated Driving* (2019)

—— *Working Group of Experts on Artificial Intelligence and Criminal Law – Working Paper II, 1st meeting* (2019)

—— *Artificial Intelligence and its Impact on CPDC Work – The case of automated driving – Thematic session on Artificial Intelligence and Criminal Law* (2018)

Council of Europe, European Commission for the Efficiency of Justice, *European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and their environment* (2018)

Council of Europe, *Governing the Game Changer – Impacts of artificial intelligence development on hu-man rights, democracy and the rule of law: Conclusions from the Conference*» (Helsinki, 26-27 February 2019)

Dias J F and Caeiro P, 'Comentário ao Acórdão do Tribunal de Justiça de 3 de Maio de 2007, proc. C-303/ 05, Advocaten voor de Wereld VZW contra Leden van de Ministerraad', in Eduardo Paz Ferreira and Maria Luísa Duarte and Miguel Sousa Ferro (eds), *Jurisprudência Cunha Rodrigues – Comentários* (Associação Académica da Faculdade de Direito de Lisboa Editora 2013)

European Parliament, Panel for the Future of Science and Technology, The impact of the General Data Protection Regulation (GDPR) on artificial intelligence (2020)

——, Committee on Civil Liberties, Justice and Home Affairs (Tudor Ciuhodaru), *Draft Report on artificial intelligence in criminal law and its use by the police and judicial authorities in criminal matters* (2020/2016(INI))

European Parliamentary Research Service (Wouter van Ballegooij), *European Arrest Warrant: European Implementation Assessment* (PE 642.839, Brussels, June 2020)

Fidalgo S, 'A utilização de inteligência artificial no âmbito da prova digital – direitos fundamentais (ainda mais) em perigo' in Anabela Miranda Rodrigues (ed), A *Inteligência Artificial no Direito Penal* (Almedina 2020)

Foucault M, *Discipline and Punish: The Birth of the Prison* (Vintage Books 1977)

Geist M A, 'Is there a there there? Toward Greater Certainty for Internet Jurisdiction' (2001) 16 Berkeley Technol. Law J. 1345

Gless S, 'AI in the Courtroom: A Comparative Analysis of Machine Evidence in Criminal Trials' (2020) 51 Georget. J. Int. Law 195

Greco L, *A Impossibilidade Jurídica de Juízes Robôs* (conference) (Instituto Eduardo Correia 2020 <http://www.institutoeduardocorreia.com.br/videos/2020> accessed 25 August 2021

Habermas J, *The Crisis of the European Union – A Response* (Polity 2012)

Hallevy G, 'The Criminal Liability of Artificial Intelligence Entities' (2010) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1564096> accessed 25 August 2021.

Hildebrandt M, 'Criminal Law and Technology in a Data-Driven Society', in Markus D. Dubber and Tatjana Hörnle (eds), *The Oxford Handbook of Criminal Law (*Oxford University Press 2014)

Interpol / United Nations Interregional Crime and Justice Research Institute (UNICRI), *Artificial Intelligence and Robotics for Law Enforcement* (Torino / Lyon 2019)

Klip A, 'International criminal law. Information society and penal law' (2014) 85 RIDP 381

Kohl U, *Jurisdiction and the Internet - Regulatory Competence over Online Activity* (Cambridge University Press 2007)

Lemos M and Costa M J, 'Inteligência Artificial e Direito da Guerra: Reflexões sobre as Armas Autónomas Mortíferas', in *Revista Julgar – Special Issue: O Direito na Era da Inteligência Artificial*, forthcoming

Ligeti K, 'Artificial Intelligence and Criminal Justice' (2019) <http://www.penal.org/en/information> accessed 25 August 2021

Lima D, 'Could AI agents be held criminally liable? Artificial Intelligence and the challenges for criminal law' (2018) 69 S. C. Law Rev. 677

Maas M, 'International Law does not Compute: Artificial Intelligence and the Development, Displacement or Destruction of the Global Legal Order' (2019) 20 Melb. J. Int. Law 29

Pagallo U and Quattrocolo S, 'The impact of AI on criminal law, and its twofold procedures', in Woodrow Barfield and Ugo Pagallo (eds.), *Research Handbook on the Law of Artificial Intelligence* (Edward Elgar Publishing 2018)

Ramalho D S, *Métodos Ocultos de Investigação Criminal em Ambiente Digital* (Almedina 2017)

Rodrigues A M, 'Medida da pena de prisão – desafios na era da Inteligência Artificial' (2020) 149 RLJ 258

—— A questão da pena e a decisão do juiz – entre a dogmática e o algoritmo', in Anabela Miranda Rodrigues (ed), A *Inteligência Artificial no Direito Penal* (Almedina 2020)

Ryngaert C, *Jurisdiction in International Law* (2nd edn, Oxford: Oxford University Press 2015)

Schwab K, 'The Fourth Industrial Revolution – What It Means and How to Respond' (2015) Foreign Aff. <www.foreignaffairs.com> accessed 25 August 2021

Singh S R, 'Move fast and break societies: the weaponisation of social media and options for accountability under international criminal law' (2019) 8 Camb. Int. Law J. 331

Sousa S A, '"Não fui eu, foi a máquina": Teoria do Crime, Responsabilidade e Inteligência Artificial', in Anabela Miranda Rodrigues (ed), *A Inteligência Artificial no Direito Penal* (Almedina 2020)

Swart B, 'Human Rights and the Abolition of Traditional Principles', in Albin Eser and Otto Lagodny (eds), *Principles and Procedures for a New Transnational Criminal Law* (Max Planck Institut 1992)

United Nations (UN) High-level Panel on Digital Cooperation (Luke Kemp et al.), *A Proposal for International AI Governance* (2019)

Vien M, 'Provisional Arrest, Release and the Role of Interpol' (1991) 62 RIDP 63

Wyngaert C, 'Double criminality as a requirement to jurisdiction', in John Dugard and Christine van den Wyngaert (eds), *International Criminal Law and Procedure* (Dartmouth Pub. Co. 1996)

Završnik A, 'Criminal justice, artificial intelligence systems, and human rights' (2020) 20 ERA Forum 567

Zimmermann F, 'Conflicts of Criminal Jurisdiction in the European Union' (2015) 3 BJCL&CJ 1.

AI-ASSISTED AND AUTOMATED ACTUARIAL JUSTICE OR ADJUDICATION OF CRIMINAL CASES

LOMBROSO 2.0: ON AI AND PREDICTIONS OF DANGEROUSNESS IN CRIMINAL JUSTICE

By Alice Giannini[*]

Abstract

The purpose of this paper is to analyze the legal and ethical issues raised by the use of artificial intelligence (AI) technologies in predicting criminal behavior. In fact, ever since Cesare Lombroso's L'uomo delinquente, scientists, on one side, and jurists, on the other, have been discussing the 'criminal brain'. Violence risk assessment tools have been applied in criminal courts for more than sixty years, yet the discussion has found once again importance thanks to the development of new AI techniques (such as machine learning) and to their application in both the medical and the criminal justice area. The new frontier is represented by the enhancement of these instruments through the combination of AI and neuropredictions. This paper presents critical reflections on the benefits and drawbacks of applying these technologies to predictions of violence in criminal justice. The inquiry concludes with a number of open questions which are hoped will contribute to the ongoing debate and work as a primer for future investigations on the matter.

1 Introduction

Dangerousness is a concept embedded in possibly every modern criminal legal system. Ever since Cesare Lombroso's *L'uomo delinquente*,[1] scientists and jurists have been discussing the 'criminal brain'. Classical questions of such debate include whether it is possible to scientifically map biological characteristics that lead to the commission of a crime, or if 'criminal science' can be used to identify in advance subjects which will incur in criminal behavior. In essence, we could claim that Lombrosianism boils down to one existential question: are we born criminals?

In point of fact, the assessment of violent behavior has been for long now the 'chief battlefield in the struggle between law and psychiatry'.[2] Accordingly, some argue that the most recent advances in behavioral genetics, neuroscience, psychiatry and criminological epidemiology, together with the emergence of neurocriminology, characterize the return

[*] PhD Student in Criminal Law, University of Florence and University of Maastricht (Double PhD Program). In 2021 her research project was awarded with the Giulia Cavallone price. For correspondence: <alice.giannini@unifi.it>.

[1] The book was first published in Italian in 1876 and then in English in 1911, after Lombroso's death, with the title 'Criminal Man'.

[2] Christopher D Webster, Mark H Ben-Aron and Stephen J Hucker, Dangerousness (Cambridge University Press 1987) 14.

of a Lombrosian vision of crime.[3] Better yet, a 'Lombroso 2.0'.[4] Conversely, more and more academics joined a newborn field of investigation named 'neurolaw', addressing the impact of neuroscience on the foundations of criminal responsibility.[5] This trend is being followed by the development of new artificial intelligence (AI) systems[6] and by their application in both the medical and the criminal justice area. The potential of using these technologies is enormous: they are capable of analyzing massive quantities of data at a very high speed, sometimes with little or no human supervision. AI technology is being combined with neuroscience to address future dangerousness and, consequently, never-before-seen correlations between violent behavior and a person's characteristics might be identified.

This paper will be structured as follows. First, it will briefly account for the relationship between science and criminal law, focusing specifically on the concept of dangerousness. Then, the investigation will touch upon algorithmic risk-assessment and AI neuroimaging. They represent two sides of the same medal: on one side, risk-assessment tools have been used in criminal justice since the 1930s and are seeing new potential today thanks to AI powered data collection and analysis. On the other, forensic neuroscience has been

[3] Christian Munthe and Susanna Radovic, 'The Return of Lombroso? Ethical Aspects of (Visions of) Preventive Forensic Screening' (2015) 8 Public Health Ethics 271.
Think for example of the experiment conducted at Cornell University in 2011, where a group of psychologists proved that individuals are capable of making rather accurate inferences about 'criminality' based on someone's facial appearance (ie on a static cropped image of a face); see Jeffrey M. Valla, Stephen J. Ceci and Wendy M. Williams, 'The Accuracy Of Inferences About Criminality Based on Facial Appearance.' (2011) 5 Journal of Social, Evolutionary, and Cultural Psychology. A few years later two researchers from the Shanghai Jiao Tong University developed an AI system which was able to recognize 'criminals' with an accuracy of 89.5%, based on the curvature of the upper lip, the distance between the inner corners of the eyes and the so-called nose-mouth angle. The publication of this study sparked a heated debate which were addressed by the researchers with a subsequent addendum; see Xiaolin Wu and Xi Zhang, 'Automated Inference on Criminality Using Face Images' [2017] arXiv:1611.04135. Everything considered, as it has been stated, '[i]f humans can spot criminals by looking at their faces, as psychologists found in 2011, it should come as no surprise that machines can do it, too', see 'Neural Network Learns to Identify Criminals by Their Faces' (MIT Technology Review, 2021) <https://www.technologyreview.com/2016/11/22/107128/neural-network-learns-to-identify-criminals-by-their-faces/> accessed 2 August 2021.

[4] The term has appeared in a small number of recently published articles, but it has not (yet) been adopted in the legal academic discourse. See Simone Cosimi, 'Gay o Etero, un algoritmo "legge" l'orientamento sessuale sul volto. Il controverso studio di Stanford' (*la Repubblica*, 2021) <https://www.repubblica.it/tecnologia/2017/09/08/news>; Dario Ronzoni, 'Lombroso 2.0: Una Rete Neurale Per Riconoscere I Criminali Dai Tratti Somatici' (DDay.it, 2021) <https://www.dday.it/redazione/21681/lombroso-20-una-rete-neurale-per-riconoscere-i-criminali-dai-tratti-somatici> accessed 9 August 2021.

[5] See Stephen Morse, 'Neuroethics: Neurolaw', *Oxford Handbooks Online* (2017); Ariane Bigenwald and Valerian Chambon, 'Criminal Responsibility and Neuroscience: No Revolution Yet' (2019) 10 Frontiers in Psychology.

[6] For the purpose of this article, we will adopt the following definition of AI systems: 'An AI system is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy". See OECD, Recommendation OECD/LEGAL/0449 of 22 May 2019 R of the Council on Artificial Intelligence (2019).

depicted as a useful tool to improve the reliability of predictions of dangerousness, thanks to the application of AI to brain scanning and diagnosing. Such analysis will serve as a primer for the following sections: in a (not very) dystopian future AI neuroimaging and brain reading technologies could be combined with risk assessment tools in a technique called 'AI neuroprediction'.[7] Could this also lead to totally automated evaluations of dangerousness and the creation of virtual forensic experts?

As predictions of dangerousness are pivotal in many phases of a trial, this paper will not take into consideration the peculiarities of each and all applications of violence risk assessments. Furthermore, it will not account for the particularities distinctive of different jurisdictions. Rather, we will attempt at forming transversal observations which refer to common issues shared by most of these applications.

Within the paper the author will develop critical observations on ethical and legal questions raised in the previous sections. 'Classical' issues regarding the use of AI-systems such as bias and the perception of human vs. algorithmic error will be tackled. The reflection will be concluded with a number of open questions which, as it is hoped, will contribute to the ongoing debate and work as a primer for future investigations on the matter. Does the application of AI systems in these fields fall in the 'New Technology, Old Problems' saying?[8] Or do they pose new issues with regards to principles distinctive of criminal law?

2 Predicting Dangerousness

2.1 Lost in translation

Before embarking on an analysis of algorithmic risk assessment, it is relevant to briefly address the relationship between criminal law and science, focusing on predicting dangerous behavior. Dangerousness is, indeed, a dangerous concept and a label 'which is easy to attach but difficult to remove'.[9] If we take the American criminal justice system as an example, predictions of dangerousness are used as standards for sentencing matters (especially capital sentencing) and for criminal commitment following verdicts of not guilty by reason of insanity; they play an important role in sexual predator statutes and they are decisive in civil commitment procedures.[10] Indeed, in 2017 the American Law Institute in its first official amendment to the Model Penal Code, which is focused

[7] See Thomas Nadelhoffer and others, 'Neuroprediction, Violence, and the Law: Setting the Stage' (2010) 5 Neuroethics 67; Leda Tortora and others, 'Neuroprediction and A.I. in Forensic Psychiatry and Criminal Justice: A Neurolaw Perspective (2020) 11 Frontiers in Psychology 220.

[8] Stephen Morse, 'Neuroprediction: New Technology, Old Problems' (2015) Faculty Scholarship at Penn Law.

[9] P. D. Scott, 'Assessing Dangerousness in Criminals' (1977) 131 British Journal of Psychiatry quoting S. H. Shaw, 'The Dangerousness of Dangerousness' (1973) 13 Medicine, Science and the Law.

[10] See John Monahan and Jennifer L. Skeem, 'Risk Assessment in Criminal Sentencing' (2016) 12 Annu Rev Clin Psychol.

on sentencing, recommended that sentencing commissions 'develop actuarial instruments or processes, supported by current and ongoing research, that will estimate the relative risks that individual offenders pose to public safety'.[11] Said tools ought to be incorporated into sentencing guidelines.

With this in mind, it is important to stress that the legal and the clinical definitions of dangerousness do not match, and this is one of the reasons why definitive scientific answers to legal questions with regards to predicting dangerousness are hard to find – if not unattainable.[12] This issue can be referred to as the *lingua franca* problem.[13] As a matter of fact, in medicine the state of health of a patient can vary along a scale from extremely ill to completely healthy. On the contrary, legal language, especially in the field of criminal responsibility, is a strictly binary language. A person can be either guilty or not guilty, insane or sane, dangerous or not dangerous. The dialogue between these two languages is troublesome and forensic psychiatry has traditionally represented the bridge between the two disciplines when it comes to criminal liability.[14] Lately, it appears as said role is being taken over by neuroscience.

Indeed, even though dangerousness is a concept that is used in numerous legal contexts and that is strongly connected to decisions impacting on a person's freedom, the law does not define dangerous individuals with the same particularity as it is done, for example, by meteorologists when it comes to dangerous storms.[15] Questions like 'What level of violence do we expect from an individual in order for him to be considered dangerous? Is it just physical violence or also psychological violence?' or 'When we say that a person will be dangerous, will it happen tomorrow, in a month or in a year? What is the temporal scope of predictions of dangerousness?' remain unanswered. In other terms, magnitude, imminence, and frequency are not defined in the legal conception of dangerousness and no systematic effort has been done so far to fill these gaps. When it comes to dangerousness the law is permissive and discretionary because it *needs* predictions of violence: criminal law guards its own domains jealously.[16]

Another pivotal issues with regards to the relationship between science and criminal law is the so called 'group to individual problem (G2i)'.[17] What the law asks of science is to

[11] American Law Institute (ALI) Model Penal Code 2017 §6B.09(2).
[12] See N. Pollock and C. Webster, 'The clinical assessment of dangerousness', in Robert Bluglass and Paul Bowden (eds*), Principles and Practice of Forensic Psychiatry* (Churchill Livingstone 1990), 489.
[13] Joshua W. Buckholtz and David L. Faigman, 'Promises, promises for neuroscience and law' (2014) 24 Current Biology R864.
[14] See Zvi Zemishlany and Yuval Melamed, 'The Impossible Dialogue Between Psychiatry and the Judicial System: A Language Problem' (2006) 43 The Israel journal of psychiatry and related sciences 150.
[15] See John Monahan and Henry J. Steadman, 'Violent storms and violent people: How meteorology can inform risk communication in mental health law' (1996) 51 American Psychologist.
[16] See Nigel Eastman and Colin Campbell, 'Neuroscience and Legal Determination of Criminal Responsibility' (2006) 7 Nature Reviews Neuroscience 314.
[17] David L. Faigman and others, 'Group to individual (G2i) inference in scientific expert testimony' (2014) Univ. Chic. Law Rev.

answer the question of whether a 'particular case is an instance of the general phenomenon',[18] where instead science 'is focused on characterizing generalizable phenomena to establish mechanistic explanations that apply within definable population groups and, hence, are generalizable to other members of those populations (who may not yet have been observed)'.[19]

Conclusively, the translation from science to law is not an easy path to walk. Some have described the process of going from science to law as a 'journey for which there is no map'.[20] What role does AI play in this scenario? As a matter of fact, following an expansive trend, AI systems are being used in healthcare in a number of areas such as diagnostics (radiology and medical imaging), surgery and clinical care.[21] What about AI applied to (forensic) psychiatry and risk assessment tools?

These questions will guide us in the following paragraphs.

2.2 Violence risk assessment tools

Violence risk assessment tools can be defined today as instruments 'designed to increase structure, consistency, and accuracy in the evaluation of the likelihood of violent recidivism through consideration of items associated with violence recidivism'.[22] These tools were first used in the 1960s for civil commitment hearings in the American criminal system to predict whether an individual with serious mental illness would be of danger to himself or to others. Subsequently, more tools were validated specifically for predicting the criminal recidivism of 'justice-involved persons with and without mental health problems'.[23]

Regardless of where they were applied, for a long time, scientific predictions of dangerousness shared one characteristic: their fallacy – no different, as some contended, from 'flipping coins in the courtroom'.[24] In the past twenty years, progress has been made and there now seems to be substantial evidence of the increase of the accuracy of predictions associated with the use of these tools.[25]

[18] David Faigman, Philip A. Dawid, and Steven E. Feinberg, 'Fitting Science into Legal Contexts: Assessing Effects of Causes or Causes of Effects?' (2014) 43 Soc. Methods & Res. 385.
[19]Russel A. Poldrack and others, 'Predicting Violent Behavior: What Can Neuroscience Add?' (2018) 22 Trends Cogn Sci. 2, 115.
[20] ibid.
[21] See WHO, Guidance on Ethics and governance of artificial intelligence for health (2021), 6.
[22] Sarah L. Desmarais and Samantha A. Zottola, 'Violence Risk Assessment: Current Status and Contemporary Issues' (2020) 794.
[23] ibid.
[24] Bruce J. Ennis and Thomas R. Litwack, 'Psychiatry and the Presumption of Expertise: Flipping Coins in the Courtroom' (1974) 62 California Law Review.
[25] Desmarais (n 22) 801.

Traditionally, scholars distinguished between two types of violence risk assessments: clinical and actuarial. This distinction is not clear-cut anymore, as recently these tools can present characteristics of both methods. We will enumerate here a number of paramount examples of the different types of violence risk assessments tools, as it is not in the scope of this analysis to provide an exhaustive account.[26]

Clinical predictions of dangerousness are based on observation, personal examination, history taking, and testing carried out by a clinician.[27] In a clinical approach the individual's past behavior is examined and the expert affirms whether the individual will likely act in the same manner in similar circumstances.[28] The factors (ie risk factors) assessed are combined in an intuitive manner in order to evaluate the risk of violence.[29] A risk factor can be defined as a 'variable that precedes and increases the likelihood of criminal behavior'.[30] In the field of predicting recidivism, risk factors have been classified into fixed markers (such as the early onset of antisocial behavior), variable markers (which can be changed over time but not through intervention, such as age), and variable factors (such as employment status).[31] The fourth type is causal risk factors, which are 'variable risk factors that, when changed through intervention, can be shown to change the risk of recidivism'.[32]

With this in mind, the main defect of clinical predictions is that they are highly discretionary and cannot be standardized. Since clinical predictions are not structured tools, each is subjective, and this entails that it may be based on erroneous stereotypes and prejudices.

On another note, actuarial methods are based on a number of pre-identified variables that are correlated statistically to risk and result in producing a probability (or a probability range) of risk.[33] The individual's future dangerousness can then be assessed based on the characteristics that he or she shares with other people for whom a base rate exists. The more factors in the assessment, the more complex the probability rate.

---

[26] For an exhaustive overview of risk assessment tools and of the most relevant literature see *inter alia* Georgia Zara, 'Tra il probabile e il certo' (2016) Diritto Penale Contemporaneo 13-23.

[27] See John Parry and Eric Y. Drogin, Criminal Law Handbook on Psychiatric and Psychological Evidence and Testing (ABA, 2000) 24.

[28] ibid 208.

[29] See John Monahan, 'A Jurisprudence of Risk Assessment: Forecasting Harm Among Prisoners, Predators, and Patients' (2009) 92 Virginia Law Review 405.

[30] Monahan and Skeem (n 10) 497.

[31] See Monahan, 'Violent storms and violent people: How meteorology can inform risk communication in mental health law' (n 14) 497.

[32] ibid.

[33] See Christopher Slobogin, *Proving the unprovable: the role of law, science, and speculation in adjudicating culpability and dangerousness* (Oxford University Press 2006) 101.

The main differences between actuarial and clinical violence risk assessment tools are efficiently explained by Hilton, Harris, and Rice,[34] who do so by identifying two conceptually distinct tasks. The first task is selecting the relevant characteristics, where the second task is combining said characteristics to obtain an interpretation. With regards to the first task, when adopting an actuarial method, the selection is typically based on one or more follow-up studies which map the factors that are related to the violent outcome. The purpose of the first task is to identify 'an optimum set of items on the basis of incremental validity—that is, selecting the most powerful predictors first and then adding items only when they improve prediction'.[35] Instead, when it comes to clinical tools this task is conducted based on 'intuition, non-empirical experience, and one's memory for empirical findings'.[36] In short, according to these authors 'it is the method of selection rather than the items attended to that distinguishes clinical from statistical prediction'.[37] With regards to the second task, clinical judgment leaves the combination rule unspecified and relies on 'gut-level' processes, where instead prototypical actuarial methods 'combine risk factors using item weights derived from empirically established relationship with violent recidivism'.[38]

The most famous actuarial tool is the Violence Risk Appraisal Guide (VRAG-R),[39] which focuses on 12 variables and aims at providing a score that indicates the probability of recidivism. It was developed from a research conducted in Canada on over 600 men committed to a maximum-security hospital. The variables were identified out of fifty predictors, which were coded from institutional files, and they were used to categorize the patients into nine groups based on their actuarial risk of future violence.

A third type of prediction of dangerousness is the adjusted actuarial assessment (or structured professional judgment, SPJ) which, similar to the actuarial approach, is based on a finite number of pre-identified variables connected to risk. Thus, differently from an actuarial tool, the factors and the conclusions are not reached through a mathematical process. Following this kind of approach, the expert will assess a probability of dangerousness using an actuarial method and then adjust the result based on other factors, such as those connected to the offender. An example of this kind of approach is the Historical Clinical Risk Management (HCR-20 V3). It is a violence risk assessment scheme which is made of twenty different ratings based on historical (such as previous violence, age, em-

[34] See N. Zoe Hilton, Grant T. Harris and Marnie E. Rice, 'Sixty-Six Years of Research on the Clinical Versus Actuarial Prediction of Violence' (2006) 34 The Counseling Psychologist 401.
[35] ibid.
[36] ibid.
[37] ibid.
[38] ibid.
[39] See Grant T Harris and others, *Violent Offenders: Appraising and Managing Risk (3rd Ed.)* (American Psychological Association 2015).

ployment, substance use, relational instability...), clinical (lack of insight, active symptoms of major mental illness, unresponsive to treatment...) and risk management variables (lack of personal support, exposure to destabilizers…).[40]

One of the most relevant deficiencies of actuarial methods is that, regardless of the number of factors which can be included in the evaluation, their aggregate determination will never be sufficient to account wholly for the individual involved and '[a]ccordingly, no matter how carefully a forensic expert assembles the available actuarial information, there still is going to be a significant leap in reaching conclusions about a particular individual unless it can be shown that that individual has recently engaged in the behavior at issue or tried to do so, but was denied the opportunity'.[41] In other words, actuarial predictions neglect some characteristics of the individual evaluated which might be deemed relevant, obscuring the person's individuality.

Recently, a new tool was developed by Monahan and his colleagues: the Classification of Violence Risk Software (COVR).[42] It is an interactive software aimed at estimating the risk that a person hospitalized for mental disorder will be violent to others based on 40 risk factors. Its goal is 'offering clinicians an actuarial ''tool''' to assist in their predictive decision making'.[43] The method developed by Monahan is based on classification trees, which means that the sequence of the questions asked after the first is based on the answer given in the previous question. This methodological choice is significant: it entails that the team was able to develop the 'first software application for actuarial risk assessment'[44]. The tool does not replace the clinical decision: in fact, the inventors of the COVR software recommend themselves a review by the clinician responsible for the risk assessment in order to avoid mistakes. Moreover, as the software was trained and validated on data pertaining exclusively to psychiatric inpatients in the US, its reliability in a criminal justice setting remains uncertain.

It is possible to identify a fourth generation of violence risk assessment tools which combine individual risk factors with community-level risk variables. Said tools integrate the prediction of risk with risk management and therefore assist the doctor to develop a case management plan.[45]

To conclude, even though it is well established that the reliability of actuarial methods outperforms clinical evaluations, they are still not clinicians' nor judges' preferred tool and therefore their use is not as pervasive as one ought to think.[46] Why? One possible

[40] See Christopher D. Webster and others, HCR-20 V3: Assessing Risk for Violence. User Guide (Burnaby 2013).
[41] Parry (n 26) 24.
[42] See John Monahan and others, 'The Classification of Violence Risk' (2019) 24 Behavioral Sciences & The Law.
[43] ibid 721.
[44] ibid.
[45] An example is the Italian C-VRR (Checklist di Valutazione del Rischio di Recidiva). See Zara (n 26) 20.
[46] See Hilton (n 32), noting that '[i]f forensic decisions mirror unaided clinical judgment and not actuarial assessments after a well-validated actuarial system has been available for a decade, we must concede that

answer to the question is connected to the G2i problem and to the fact that actuarial assessments are seen as 'too impersonal for the purposes of the law'.[47] Consequently, according to some, '[r]eliable and accurate assessment of violence risk remains an elusive goal for forensic psychiatrists'.[48]

It is in this scenario that neuroprediction first, and then AI, enter the picture.

2.3 Enter: AI

2.3.1 Neuroprediction and AI

The lion's share of the debate on the impact of neurosciences on criminal law is centered on ascertaining criminal responsibility,[49] rather than predictions of violent behavior.[50] Nevertheless, in the last ten years scientists started turning to neuroscience in a quest to improve the reliability of forecasting future violent behavior, by linking the structure of the brain to dangerousness. One of the reasons guiding the inclusion of neuroscientific data in risk assessment is that by adding 'important personalized information about the brain of offenders to the risk assessment equation' said studies might 'make it more likely that legal decision makers rely on the best available tools of violence risk assessment'.[51] Some argue that the future use of neuroprediction is 'not morally worse than the present use of clinical risk assessments', since it regards 'objectionable features of the use of statistical, unlike individualized, evidence' and because it is 'more accurate'.[52] Others harshly criticize arguments in favor of moral permissibility of this practice, based on the fact that they will present the same issues that all actuarial tools present.[53]

The most recent scientific studies take an even forward steps and focus on 'the use of structural or functional brain parameters coupled with machine learning methods to make clinical or behavioral predictions', namely AI neuroprediction.[54] As a matter of fact, Tortora and others predict that AI neuroprediction of recidivism 'is likely to become available in the near feature'.[55] These studies include using Functional Magnetic Resonance Imaging (fMRI) data to predict recidivism based on potential correlations between low activation of the portion of the brain deputed to impulses and error processing (the dorsal anterior cingulate cortex) and recidivism;[56] using machine learning to study

---

the mere availability of actuarial methods—and the careful analysis and dissemination of their consistent advantage—is fundamentally insufficient for their adoption' 404.

[47] Nadelhoffer (n 6) 79.

[48] Richard G. Cockerill, 'Ethics Implications of the Use of Artificial Intelligence in Violence Risk Assessment' (2020) 48 J Am Acad Psychiatry Law 1.

[49] Tortora (n 6) 3.

[50] Nadelhoffer (n 6) 634.

[51] ibid 86.

[52] Lippert-Rasmussen (n 61) 125 summarizing '*Nadelhoffer's Argument*', Nadelhoffer (n 6).

[53] ibid 135.

[54] Tortora (n 6) 2.

[55] ibid.

[56] See Alexandre Abraham and others, 'Machine Learning for Neuroimaging with Scikit-Learn' (2014) 8 Frontiers in neuroinformatics 14.

whether brain age can be used as a predictive factor for rearrest;[57] and evaluating whether including the levels of cerebral blood flow in risk assessment could improve predictions of violent behavior in long-term follow-up forensic psychiatric patients.[58]

One critique to (AI) neuroprediction which is hardly surmountable at this point – but might be in the future – is that current neuropredictions are developed on a very low base rate which often comprises of a homogenous population, such as inmates or psychiatric inpatients. In this regard, finding plausible solutions to the G2i problem becomes troublesome. Consider the following example:

> A neuroscientist uses fMRI to scan 100 participants who are instructed to either lie or tell the truth about a set of facts. Contrasting brain activity during lying with truth telling reveals statistically significant activation in dorsolateral prefrontal cortex (DLPFC). This permits the valid group-level inference that lying is associated with DLPFC engagement. However, examination of each individual's data reveals that, while most subjects exhibited higher DLPFC activity during lying, some participants showed no difference and still others demonstrated lower DLPFC activity during lying compared with truth telling.[59]

Simply put, the risk of false positive and false negatives is prominent.

All things considered, discoveries in neuroscience have traditionally been accompanied by a great level of enthusiasm regarding their potential for explaining causal processes of violence behavior, ie of opening the 'black box' of the human brain.[60] However, by applying complex AI to neuroprediction we might be, in fact, introducing a new black box in the system. We will develop this argument further on in this analysis.

2.3.2 AI: friend or foe?

One common characteristic of the risk assessment tools which were mentioned in section 2.2. is that they are static, not dynamic, systems. They are based on rather simple algorithms which simply weight and combine information to produce the likelihood of future violent behavior.[61] This dynamicity can be explained as follows: by applying deep learning[62] to violence risk assessments 'new data are constantly incorporated to improve and

---

[57] See Kent A. Kiehl and others, 'Age of gray mattes: Neuroprediction of recidivism' (2018) 19 NeuroImagie: Clinical.

[58] See Carl Delfin and others, 'Prediction of recidivism in a long-term follow-up of forensic psychiatric patients: Incremental effects of neuroimaging data' (2019) PLoS One.

[59] Poldrack (n 17) 115.

[60] ibid.

[61] Desmarais (n 22) 813.

[62] Deep learning (DL) is a subset of ML. An algorithm based on Machine Learning (ML) techniques teaches itself rules by learning from the training data through statistical analysis, detecting patterns in large amounts of information and generating outputs. Deep Learning consists of layers of artificial neural networks (ANNs). Neural networks' engineering was inspired by the functioning of biological neurons, hence their basic function is to establish features from input. ANNs are made of layers of functions ("nodes" or "neurons") that perform various operations on the data that they are fed. The main difference

refine a predictive model' as 'correct predictions reinforce the model, while incorrect predictions cause it to recalibrate'.[63] In other words, risk assessment tools based on AI represent an 'enhanced version of … the empirical actuarial approach', as these systems have the potential of combining 'countless data points in complex ways to identify persons at risk of violence', developing models of 'unfathomable complexity'.[64] As such, AI systems could 'amplify or mitigate both the strengths and the weaknesses associated with existing actuarial prediction techniques'.[65]

Let us consider an example. Between 2015 and 2018, a study on how to detect risks of school violence was conducted on 131 students.[66] The aim of the study was to develop an AI system based on natural language processing[67] and machine learning capable of automatically analyzing the contents of interviews and information on the student's household in order to identify risk factors and predict the risk of violent behavior of the single student. The study proved that linguistic patterns are relevant indicators of a student's risk of school violence. The deployment of this system would facilitate immensely school violence risk assessment, which is a costly procedure (each assessment conducted by a clinician costs from $1000 to $3500 dollars).

What does this entail? If currently applied tools for predicting dangerousness are 'merely' an instrument in the hand of judges and psychiatrists, in the future AI systems could be programmed to replace judicial and clinical decision-making. We will analyze this perspective in the following sections.

It should be noted at this point that one of the benefits of applying AI to risk assessment is that they process massive amounts of data in an incredibly fast and accurate way, saving time in often very lengthy criminal trials. Indeed, the revolutionary aspect of machine learning algorithms is that they are capable of identifying unforeseen patterns and correlations (ie predictions) between factors that are not perceived by the human agent (as it would be with a traditional actuarial risk assessment tools), finding new solutions for a given task. The question that rises, then, is the following: should all the links identified by AI through these massive data analysis operations be considered significant links?

---

between ML and DL is that in DL the algorithm is fed raw (unlabeled) data and then identifies by itself which features are relevant. In ML learning, instead, the algorithm is given an established of relevant features to analyze.

[63] Cockerill (n 22) 1.

[64] Neil R. Hogan, Ethan Q. Davidge, Gabriela Corabian, 'On the Ethics and Practicalities of Artificial Intelligence, Risk Assessment, and Race' (2021) 49 J Am Acad Psychiatry Law 3, 2.

[65] ibid.

[66] See Ni Yizhaho and others, 'Finding warning markers: Leveraging natural language processing and machine learning technologies to detect risk of school violence' (2020) 139 International Journal of Medical Informatics.

[67] Process by which the system extracts data from human language and makes decisions based on that information. It enables clear human-to-machine communication. Examples of NLP systems are voice-activated digital assistants such as Alexa, Siri, Cortana and Google Assistant.

Who should decide which kind prediction is acceptable from a criminal law standpoint and which one is not?

If we think of 'traditional' actuarial risk assessment tools, the choice of which risk factors to include in the analysis has always been deemed critical *per se*. Think of gender, race or life history variables such as the level of education or employment or the family environment: all in all, the use of these technologies cannot be perpetrated without taking into account the principles paramount of criminal law. Applying these principles to risk factors entails that in order for a variable to be relevant from a criminal law standpoint, the parameter has to be connected to the blameworthiness of the individual. As it was clarified,

> Demographic and life history variables that characterize an offender may have significant predictive validity in assessing his or her likelihood of recidivism, but no bearing on the ascription of blame for the crime of which he or she was convicted'. Both race and gender correlate significantly with criminal recidivism … However, neither race nor gender is seen as bearing on an offender's blameworthiness for having committed crime—as a class, offenders who are women are seen as no more (or no less) blameworthy than offenders who are men, and offenders who are African American are seen as no more (or no less) blameworthy than offenders who are white.[68]

We can now apply the same kind of argument used as a critique to "classical" violence risk assessment tools to AI based systems. Imagine an AI-system that analyzes huge amounts of data contained in criminal records at a very fast pace. Say that the task given to the system is to identify the factors which show the highest correlation with convictions of rape, as the intended purpose is to prevent the commission of such crimes by felons on a national basis. Imagine now that the algorithm identifies with a certainty of 99.9% that out of 100.000 individuals convicted for rape, 60% had red hair. Should police forces focus more on red haired individuals in their crime prevention activities? Taking this example even further, say that according to the output of the algorithm out of the aforementioned 60% red haired sex offenders, 35% of them reoffended and they were all taller than 1.80 m. Is this connection noteworthy? Should it affect a judge deciding on whether the (poor) red haired tall defendant can be admitted to parole? The issue becomes even thornier if we take into consideration other traits of an individual.

Think of mental illness, which has been carrying the stigma of dangerousness for decades: it follows that there has been a lot more attention (*rectius*, data collection) on dangerous behaviors by mentally ill offenders than there has been on sane offenders, making available databases unbalanced.[69] This is surely an overly simplified case scenario but it

[68] Monahan, 'Risk Assessment in Criminal Sentencing' (n 9) 503.

[69] It has been empirically proven that symptoms or diagnoses of serious mental disorders are not related or inversely related to subsequent violence in a variety of clinical populations such as civil psychiatric patients, forensic patients and mentally disordered offenders, sex offenders and violent offenders in general. See the studies cited in Hilton (n 35) 402.

brings to light important ethical questions which will have to be confronted in the future in order to truly avoid Lombroso 2.0 criminal policies.

To continue, the use of AI and neurodata does not supersede on of the classical issues related to violence risk assessments: the asymmetrical perception of statistic (or actuarial) evidence versus individualized evidence. As humans, we are bound to think that machines cannot make mistakes ('to err is human, not algorithmic'[70]), yet we trust more (fallible) human decisions. This problem is best explained through the famous 'gatecrasher's case',

> It is certain that 1,000 people attended a football match and certain that only 10 people bought a ticket. In one case, John is convicted for gatecrashing solely on the basis of statistical evidence that he, undeniably, went to the football match and accordingly, on the basis of this information, there is a 99% probability that he gatecrashed. In another case, John is convicted solely on the basis of a piece of individualized evidence consisting in an eyewitness report of John's gatecrashing. There is a 99 % probability that the eyewitness report is true.[71]

The critical issue is that most people will believe that John should be acquitted in the first scenario but convicted in the second. Nevertheless, the evidence in the two scenarios is exactly the same with regards as the likelihood that John gatecrashed.

Moreover, the application of AI and the introduction of neuroprediction does not seem to solve the problem of the lack of individualization of actuarial assessments: criminal judgments have to be specific and tailored to the distinctive features of the single perpetrator and cannot be reduced to pointing out that an individual X due to certain characteristics belongs to a general category Y of people who recidivated in the past. Yet, some say that this objection to actuarial assessments can be overcome, since clinicians also apply a G2i logic in their evaluations: their clinical assessment relies on their training and on comparing the individual case to their past experiences with similar cases, hence 'while clinicians look at individual patterns, they do not so in a vacuum'.[72] One thing is for certain: judges and clinicians could be influenced in their decisions by a number of (criminally) irrelevant and unaccounted factors, where instead an AI system might not. In other words: an AI system will not deliver a more lenient judgment based on what it had for lunch. A (human) judge might.[73]

---

[70] Laetitia A. Renier, Marianne Schmid Mast and Anely Bekbergenova, 'To Err Is Human, Not Algorithmic – Robust Reactions to Erring Algorithms' (2021) Computers in Human Behavior.

[71] Kasper Lippert-Rasmussen (n 62) 124.

[72] Christopher Slobogin, 'Dangerousness and Expertise' (1984) 133 University of Pennsylvania Law Review, 126.

[73] See Myles Udland, 'Want a Favorable Ruling in Court? Catch A Judge Right After Lunch.' (Business Insider, 2021) <https://www.businessinsider.com/court-leniancy-improves-after-judges-eat-2015-11?r=US&IR=> accessed 19 July 2021.

Further concerns regarding AI-based risk assessments include the 'GIGO' problem (Garbage In, Garbage Out)[74] and bias: biased data leads to biased predictions. Be that as it may, one must also account for two common misconceptions with regards to bias in statistical predictions systems which are paramount of criminal justice. First, that these systems will produce a biased output only if they are trained on inaccurate or incomplete datasets.[75] The second misconception is referred to as 'fairness through unawareness' and indicates the belief that 'predictions can be made unbiased by avoiding the use of variables indicating race, gender, or other protected classes'.[76]

Admittedly, there is no such thing as an 'easy fix'.[77] If we take race as an example, scholars argue that the root of racial inequality in risk assessment lies in the 'nature of the prediction itself', [78] rather than in the databases or in the structure of the algorithm. Since predictions 'look to the past to make guesses about future events', it is only natural that in 'a racially stratified world, any method of prediction will project the inequalities of the past into the future'.[79] In this regard, some claim that one of the benefits of including neuroimaging in risk assessment evaluations would represent as a way to decrease bias in risk assessments, provided that neuroprediction is not just incorporated in (already biased) existing risk assessment tools.[80]

A problematic feature which is distinctive of AI systems – and not of traditional actuarial systems – is the one of black box: the process of creation of outputs produced by complex AI systems, such as the one based on deep learning, might not be explainable (to the laymen, but even to its creator) since the algorithm teaches itself a rule without any kind of instruction. In other words, we are dealing with a 'system [that] is so complicated that even the engineers who designed it may struggle to isolate the reason for any single action. And you can't ask it: there is no obvious way to design such a system so that it could always explain why it did what it did.' [81]

For example, a recent study on the application of deep learning to detect Covid-19 in chest radiographs proved that certain AI systems produced a diagnosis by relying on 'confounding factors rather than medical pathology' so that the results of the systems appear accurate at first sight, but they are not when tested in a different medical facility. Simply put, these systems used 'shortcuts': instead of learning the real underlying pathology which could prove the presence of COVID-19, they used 'spurious associations

[74] See, amongst others, 'Report On Algorithmic Risk Assessment Tools in the U.S. Criminal Justice System - The Partnership On AI' (The Partnership on AI, 2021) <https://www.partnershiponai.org/report-on-machine-learning-in-risk-assessment-tools-in-the-u-s-criminal-justice-system/> accessed 19 July 2021.
[75] ibid.
[76] ibid.
[77] Sandra G. Mayson, 'Bias In, Bias Out' (2018) 128 The Yale Law Journal 8, 2218.
[78] ibid.
[79] ibid.
[80] Tortora (n 6) 5.
[81] Will Knight, 'The Dark Secret at the Heart of AI' (MIT Technology Review, 2021) <https://www.technologyreview.com/2017/04/11/5113/the-dark-secret-at-the-heart-of-ai/> accessed 19 July 2021.

between the presence or absence of COVID-19 and radiographic features that reflect variations in image acquisition'.[82] Notably, the study proved that this kind of shortcut learning might occur even when data of better quality (and not data that 'confuse' the system) is used to train the system. The perils of these 'shortcuts' and the lack of transparency, were the AI be applied for risk assessment in a field as delicate as criminal justice, are self-evident.

3 A Glance to the Future: Towards Virtual Forensic Experts?

According to a global survey conducted in 2020, psychiatrist are not that worried that AI could steal their jobs: '[t]he mental status examination, evaluation of dangerous behavior and formulation of a personalized treatment plan, all essential roles of a psychiatrist, were … felt to be tasks that a future AI/ML technology would be unlikely to perform'.[83] As a matter of fact, the work of a forensic psychiatrist is 'rife with ethical dilemmas'.[84] Differently from other branches of medicine, psychiatrists have to balance the duty of care of their patient with issues regarding public safety almost on a daily basis. How could AI assist psychiatrists through this minefield? Possible solutions might lie ahead.

For instance, in January 2021 the Department of Neuroscience of the University La Sapienza in Rome published an application for a research grant titled 'Artificial intelligence in forensic psychiatry: the development of a new algorithm to guide and structure forensic evaluations of criminal responsibility and social dangerousness'. The aim of this project is to overcome ethical issues related to the application of machine learning techniques in the forensic field by 'developing models and an (Explainable & Trustworthy) AI-based Decision Support System (DSS) (Virtual Forensic Experts) to guide and support forensic psychiatric evaluations of criminal responsibility and social dangerousness, in order to make them more objective, transparent, and reliable'.[85]

The themes which emerge from these few sentences are innumerable. We will only mention a few here: What would be the role of Virtual Forensic Experts in future society? We have touched upon some of the most futuristic developments of violence risk assessment (AI neuroprediction) and, as we know, these systems could produce a reliable output without being able to provide an explanation of the underlying logical process. How will a forensic psychiatrist interact with the machine? Will the AI systems be faithful assistants of human forensic experts or will they take their place completely? We mentioned earlier in this paper that AI systems could be fairer than a human judge or a human

[82] Alex J. DeGrave, Joseph D. Janizek and Su-In Lee, 'AI for radiographic COVID-19 detection selects shortcuts over signal (2021) 3 Nature Machine Intelligence 610-619.
[83] P. Murali Doraiswamy, Charlotte Blease and Kaylee Bodner, 'Artificial Intelligence and the Future of Psychiatry: Insights from a Global Physician Survey' (2020) 102 Artificial Intelligence in Medicine, 12.
[84] Hogan (n 67) 6.
[85] 'Artificial Intelligence in Forensic Psychiatry: The Development of New Algorithm to Guide and Structure Forensic Evaluations of Criminal Responsibility and Social Dangerousness' (*EURAXESS*, 2021) <https://euraxess.ec.europa.eu/jobs/598745> accessed 19 July 2021.

psychiatrist. However, can AI systems make ethical decisions? Could we think of a Hippocratic Oath for 'artificial psychiatrists'? These will be the topics of future conversations between psychiatrists and criminal legal scholars.

4 Conclusion

In this paper we have attempted at breaking down the issues related to applying AI to risk assessment tools for predicting future dangerous behavior. We started by addressing a classical debate, namely the dialogue between criminal law and science. Subsequently, we provided a short state of the art with regards to violence risk assessment tools. Then, we introduced AI to our reflection: we first focused on a niche area, namely the intersection between AI and neurosciences applied to predictions of recidivism. Building on these notions, we conducted a number of critical reflections on the benefits and drawbacks of applying AI to predictions of violence in criminal justice. We concluded this investigation with the introduction of a futuristic scenario such as the creation of virtual forensic experts and its possible impact from a legal and an ethical point of view.

Many questions remain open. Hopefully, they will work as a primer for the debate that is coming. For example, it will be relevant to analyze what the role of judges would be in a future where AI neuroprediction is applied. How binding will AI neuropredictions be in a criminal trial? There is no transversal solution to the issues that emerged in this short analysis. The answers to the questions we asked will not only depend on the possible legal implications, but also on social concerns. The key factor in the future discussion will be, for example, the degree of development of AI techniques in a certain country, the degree of trust placed in such techniques by the population (and therefore by the legislator) and the degree of confidence that the judge will have when having to decide on AI-related matters. We end where we began, as we will be bound to ask ourselves one final (recurrent) question: are we criminals because of what we do or because of who we are?

References

Abraham A, and others, 'Machine Learning for Neuroimaging with Scikit-Learn' (2014) 8 Frontiers in neuroinformatics

American Law Institute (ALI) Model Penal Code [2017]

'Artificial Intelligence in Forensic Psychiatry: The Development of a New algorithm to Guide and Structure Forensic Evaluations of Criminal Responsibility and Social Dangerousness' (*EURAXESS*, 2021) <https://euraxess.ec.europa.eu/jobs/598745> accessed 19 July 2021

Bigenwald A, and Chambon V, 'Criminal Responsibility and Neuroscience: No Revolution Yet' (2019) 10 Frontiers in Psychology

Bluglass R, and Bowden P (eds), *Principles and Practice of Forensic Psychiatry* (Churchill Livingstone 1990)

Buckholtz J W, and Faigman L D, 'Promises, promises for neuroscience and law' (2014) 24 Current Biology R864

Cockerill R, 'Ethics Implications of the Use of Artificial Intelligence in Violence Risk Assessment' (2020) 48 J Am Acad Psychiatry Law

Cosimi S, 'Gay o etero, un algoritmo "legge" l'orientamento sessuale sul volto. Il controverso studio di Stanford' (la Repubblica, 2021) <https://www.repubblica.it/tecnologia/2017/09/08/news/gay_o_etero_un_algoritmo_legge_l_orientamento_sessuale_sul_volto_il_controverso_studio_di_stanford-174923406/> accessed 9 August 2021

DeGrave A J, Janizek J D, and Lee S, 'AI for radiographic COVID-19 detection selects shortcuts over signal (2021) 3 Nature Machine Intelligence

Delfin C, and others, 'Prediction of recidivism in a long-term follow-up of forensic psychiatric patients: Incremental effects of neuroimaging data' (2019) PLoS One

Desmarais S L, and Zottola S A, 'Violence Risk Assessment: Current Status and Contemporary Issues' (2020) 794

Doraiswamy P, Blease C, and Bodner K, 'Artificial Intelligence and the Future of Psychiatry: Insights from a Global Physician Survey' (2020) 102 Artificial Intelligence in Medicine

Douglas T, and others, 'Risk Assessment Tools in Criminal Justice and Forensic Psychiatry: The Need for Better Data' (2017) 42 European Psychiatry

Eastman N, and Campbell C, 'Neuroscience and Legal Determination of Criminal Responsibility' (2006) 7 Nature Reviews Neuroscience

Ennis B, and Litwack T, 'Psychiatry and the Presumption of Expertise: Flipping Coins in the Courtroom' (1974) 62 California Law Review

Faigman D, Dawid P A, and Feinberg S E, 'Fitting Science into Legal Contexts: Assessing Effects of Causes or Causes of Effects?' (2014) 43 Soc. Methods & Res.

—— and others, 'Group to individual (G2i) inference in scientific expert testimony' (2014) Univ Chic Law Rev

Glenn A, and Raine A, 'Neurocriminology: Implications for the Punishment, Prediction and Prevention of Criminal Behaviour' (2013) 15 Nature Reviews Neuroscience

Harris G T and others, 'Violent Recidivism of Mentally Disordered Offenders: The Development of a Statistical Prediction Instrument' (1993) 20 Crim Just & Behav

—— *Violent Offenders: Appraising and Managing Risk (*3Rd Ed., American Psychological Association 2015)

Hilton N, Harris G T, and Rice M, 'Sixty-Six Years of Research on the Clinical Versus Actuarial Prediction of Violence' (2006) 34 The Counseling Psychologist

Hoffman M, 'Nine Neurolaw Predictions' (2018) 21 New Criminal Law Review

Hoga N R, Davidge E Q, Corabian G, 'On the Ethics and Practicalities of Artificial Intelligence, Risk Assessment, and Race' (2021) 49 J Am Acad Psychiatry Law

Kiehl A K, and others 'Age of gray mattes: Neuroprediction of recidivism' (2018) 19 NeuroImagie: Clinical

Lippert-Rasmussen K, 'Neuroprediction, Truth-Sensitivity, and the Law' (2014) 18 The Journal of Ethics 2

Monahan J, and Steadman H J, 'Violent storms and violent people: How meteorology can inform risk communication in mental health law' (1996) 51 American Psychologist

—— 'A Jurisprudence of Risk Assessment: Forecasting Harm among Prisoners, Predators, and Patients' (2009) 92 Virginia Law Review 405

—— and others 'The Classification of Violence Risk' (2006) 24 Behavioral Sciences & the Law.

—— and Skeem J L, 'Risk Assessment in Criminal Sentencing' (2016) 12 Annu Rev Clin Psychol

Morse S, 'Neuroprediction: New Technology, Old Problems' (2015) Faculty Scholarship at Penn Law

—— 'Neuroethics: Neurolaw', Oxford Handbooks Online (2017)

Munthe C, and Radovic S, 'The Return of Lombroso? Ethical Aspects of (Visions of) Preventive Forensic Screening' (2015) 8 Public Health Ethics

Musumeci E, 'Against the Rising Tide of Crime: Cesare Lombroso and Control of the "Dangerous Classes" in Italy, 1861-1940' (2018) Crime, Histoire & Sociétés

Nadelhoffer T, and Sinnott-Armstrong W, 'Neurolaw and Neuroprediction: Potential Promises and Perils' (2012) Philosophy Compass 7/9

'Neural Network Learns to Identify Criminals by Their Faces' (MIT Technology Review, 2021) <https://www.technologyreview.com/2016/11/22/107128/neural-network-learns-to-identify-criminals-by-their-faces/> accessed 2 August 2021

OECD, Recommendation OECD/LEGAL/0449 of 22 May 2019 R of the Council on Artificial Intelligence (2019)

Parry J, and Drogin E Y, *Criminal Law Handbook on Psychiatric and Psychological Evidence and Testing* (ABA, 2000)

Renier L, Schmid Mast M, and Bekbergenova A, 'To err is human, not algorithmic – Robust reactions to erring algorithms' [2021] Computers in Human Behavior

Poldrack R A, and others, 'Predicting Violent Behavior: What can Neuroscience add?' (2018) 22 Trends Cogn Sci 2

'Report on Algorithmic Risk Assessment Tools in the U.S. Criminal Justice System - The Partnership on AI' (The Partnership on AI, 2021) <https://www.partnershiponai.org/report-on-machine-learning-in-risk-assessment-tools-in-the-us-criminal-justice-system/> accessed 19 July 2021

Ronzoni D, 'Lombroso 2.0: una rete neurale per riconoscere i criminali dai tratti somatici' (DDay.it, 2021) < https://www.dday.it/redazione/21681/lombroso-20-una-rete-neurale-per-riconoscere-i-criminali-dai-tratti-somatici> accessed 9 August 2021

Scott P, 'Assessing Dangerousness in Criminals' (1977) 131 British Journal of Psychiatry

Slobogin C, 'Dangerousness and Expertise' (1984) 133 University of Pennsylvania Law Review

—— *Proving the unprovable: the role of law, science, and speculation in adjudicating culpability and dangerousness* (Oxford University Press 2006)

Stone A, and Stromberg D C, *Mental Health and Law: A System in Transition* (National Institute of Mental Health, Center for Studies of Crime and Delinquency 1976)

Strano M, 'A Neural Network Applied to Criminal Psychological Profiling: An Italian Initiative' (2004) 48 Int J Offender Therapy & Comp Criminology 495

Tortora L, and others, 'Neuroprediction and A.I. in Forensic Psychiatry and Criminal Justice: A Neurolaw Perspective' (2020) 11 Frontiers in Psychology

Udland M, 'Want A Favorable Ruling in Court? Catch a Judge Right After Lunch.' (Business Insider, 2021) <https://www.businessinsider.com/court-leniancy-improves-after-judges-eat-2015-11?r=US&IR=> accessed 19 July 2021

Walsh C G, Ribeiro D J, Franklin J C, 'Predicting Risk of Suicide Attempts Over Time Through Machine Learning' (2017) 5 Clinical Psychological Science 3

Webster C, Ben-Aron M, and Hucker S, Dangerousness (Cambridge University Press 1987)

——and others, *HCR-20 V3: Assessing Risk for Violence. User Guide (*Burnaby 2013)

WHO, Guidance on Ethics and governance of artificial intelligence for health (2021)

Will Knight, 'The Dark Secret at the Heart of AI' (MIT Technology Review, 2021) <https://www.technologyreview.com/2017/04/11/5113/the-dark-secret-at-the-heart-of-ai/> accessed 19 July 2021

Yizhaho N, and others, 'Finding warning markers: Leveraging natural language processing and machine learning technologies to detect risk of school violence' (2020) 139 International Journal of Medical Informatics

Zara G, 'Tra il probabile e il certo' (2016) Diritto Penale Contemporaneo

Zemishlany Z, and Melamed Y, 'The impossible dialogue between psychiatry and the judicial system: a language problem' (2006) 43 The Israel journal of psychiatry and related sciences

Mayson G S, 'Bias In, Bias Out' (2018) 128 The Yale Law Journal 8

Valla J, Ceci S, and Williams W, 'The accuracy of inferences about criminality based on facial appearance.' (2011) 5 Journal of Social, Evolutionary, and Cultural Psychology

Wu X, and Zhang X, 'Automated Inference on Criminality using Face Images' [2017] arXiv:1611.04135v1

THE USE OF AI TOOLS IN CRIMINAL COURTS: JUSTICE DONE AND SEEN TO BE DONE?

By Vanessa Franssen[*] and Alyson Berrendorf[**]

Abstract

Artificial intelligence (AI) is impacting all sectors of society these days, including the criminal justice area. AI has indeed become an important tool in this area, whether for citizens seeking justice, legal practitioners or police and judicial authorities. While there is already a large body of literature on the prediction and detection of crime, this article focuses on the current and future role of AI in the adjudication of criminal cases. A distinction is made between AI systems that facilitate adjudication and those that could, in part or wholly, replace human judges. At each step, concrete examples are given, and it is evaluated what are, or could be, the advantages and disadvantages of such systems when used in criminal courts.

1 Introduction

Artificial intelligence (AI) has become a subject that is difficult to ignore these days. All sectors of society seem to be impacted: the financial sector, health care, national security, public administration, (social) media, and even the legal area. In this area in particular, AI seems to have become a major and perhaps unavoidable tool, whether for citizens seeking justice, legal practitioners or police and judicial authorities.

Without a doubt, this new form of intelligence makes it possible to perform more quickly and effectively certain basic and tedious tasks that, prior to the arrival of AI systems, used to consume a lot of time. Data analysis and cross-referencing have achieved results in terms of performance and accuracy that would have been hard to imagine some years ago. For instance, predictive policing tools such as PredPol (for Predictive Policing),[1] Pal-

[*] Professor of criminal law and criminal procedure, University of Liège; Affiliated Senior Researcher, KU Leuven. For correspondence: <vanessa.franssen@uliege.be>.

[**] PhD Candidate (Research fellow F.R.S.-FNRS) in the field of criminal law and criminal procedure, University of Liège. For correspondence: <alyson.berrendorf@uliege.be>.

[1] For some basic information about the functioning of this system, see PredPol, 'What. Where. When' <https://www.predpol.com/> accessed 14 July 2021.

antir,[2] CAS (for Criminality Anticipation System)[3] or CloudWalk,[4] are already widely used by police authorities around the globe to predict where future crime will take place.[5] Supporters of such technologies highlight the effectiveness and efficiency of predictive policing, enabling police authorities to better target their interventions and even anticipate certain criminal phenomena, whereas critics argue that these systems are less effective than they appear to be, resulting in a self-fulfilling prophesy and targeting mainly lower social classes and ethnic minorities,[6] and that they contain biases and have disempowering effects.[7] Moreover, a human decision, based on digitized and potentially reductive information, remains necessary.[8] Regardless of who is right or wrong in this particular debate, the development of AI tools creates high expectations, also in the field of criminal justice, as they can support and perhaps even replace human actors in preparing and trying criminal cases.

This raises the question whether AI tools will enable us to achieve, if not 'ultimate' justice, at least greater justice at a less high cost and more equity. The development of AI indeed looks promising in terms of speed, efficiency and effectiveness of justice, but it may well prove to be Pandora's box. What is the potential of AI tools for adjudication by criminal courts and what are the pitfalls? These are the questions that we will address in this short article, thereby explicitly focusing on the *adjudication* of criminal cases, not on the prediction and detection of crime as there is already a large body of literature on this

[2] See eg Pascal Hérard, 'Technologies de prédiction du crime: Palantir a scruté les citoyens de la Nouvelle-Orléans en secret pendant 6 ans' (*TV5Monde*, 3 March 2018) <https://information.tv5monde.com/info/technologies-de-prediction-du-crime-palantir-scrute-les-citoyens-de-la-nouvelle-orleans-en> accessed 14 July 2021; Ali Winston, 'New Orleans ends its Palantir predictive policing program' (*TheVerge*, 15 March 2018) <https://www.theverge.com/2018/3/15/17126174/new-orleans-palantir-predictive-policing-program-end> accessed 14 July 2021.

[3] See eg Serena Oosterloo and Gerwin van Schie, 'The Politics and Biases of the "Crime Anticipation System" of the Dutch Police' in Jo Bates, Paul D. Clough, Robert Jäschke and Jahna Otterbacher (eds), *Proceedings of the International Workshop on Bias in Information, Algorithms, and Systems* (CEUR Workshop Proceedings 2018) 30, 41.

[4] See eg Yuan Yang, Yingzhi Yang and Sherry Fei Ju, 'China seeks glimpse of citizens' future with crime predicting AI' (*Financial Times*, 23 July 2017) <https://www.ft.com/content/5ec7093c-6e06-11e7-b9c7-15af748b60d0> accessed 14 July 2021.

[5] See eg Walter L Perry, Brian McInnis, Carter C Price, Susan Smith and John S. Hollywood (2013), 'Predictive Policing. Forecasting Crime for Law Enforcement' (*Rand Research Brief*, 2018) <https://www.rand.org/ pubs/research_briefs/RB9735.html> accessed 15 July 2021.

[6] See eg Cathy O'Neil, *Algorithmes – La Bombe à retardement* (Les Arènes 2018) 135-137; Mehdi Harmi, 'Algorithmes – Peuvent-ils prédire l'avenir ?' (2021) Science et Vie 76; Interview with Angèle Christin: Hubert Guillaud, 'La justice prédictive (2/3) : prédictions et régulations' (*Le Monde*, 13 September 2017) < https://www.lemonde.fr/blog/internetactu/2017/09/13/la-justice-predictive-23-predictions-et-regulations /> accessed 5 July 2021.

[7] See eg Rosamunde Elise van Brakel, 'Pre-emptive Big Data Surveillance and its (Dis)empowering Consequences: The Case of Predictive Policing' in Bart van der Sloot, Dennis Broeders and Erik Schrijvers (eds), *Exploring the Boundaries of Big Data* (Amsterdam University Press 2016) 117, 141.

[8] Emre Bayamlioglu and Ronald Leenes, 'The "rule of law" implications of data-driven decision-making: a techno-regulatory perspective' (2018) Law, Innov Technol 295, 313.

subject.[9] To this end, we will make a distinction between AI systems that facilitate adjudication (Part 3) and those that could, in part or wholly, replace human judges (Part 4). At each step, we will give some concrete examples and evaluate what are, or could be, the advantages and disadvantages of such systems in the area of criminal justice. However, before doing so, it may be good to clarify some elementary concepts related to AI, necessary for a good understanding of the problems inherent to software developed for and used in criminal justice (Part 2).

2 Some Basic Notions

Despite the growing interest in AI, at all levels, there still exists quite some confusion about certain basic concepts.[10] In fact, this is not entirely surprising, considering the lack of international consensus on what exactly constitutes an AI system. At EU level, recent attempts have been made to define more clearly this notion. According to the High-Level Expert Group on AI set up by the European Commission, AI systems are:

> software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal (…).[11]

In its proposal for an AI Act,[12] made public in February 2021, the European Commission put forward a more precise and simpler definition of AI systems. While one may welcome the Commission's attempt to define AI in a narrower way, the proposed definition has also provoked quite some critical reactions.[13]

---

[9] See eg Neil Shah, Nandish Bhagat and Manan Shah, 'Crime forecasting: a machine learning and computer vision approach to crime prediction prevention' (2021) 4 VCIBA 9; Rohit Patil, Muzamil Kacchi, Pranali Gavali, Komal Pimparia, 'Crime Pattern Detection, Analysis & Prediction using Machine Learning' (2020) 7 IRJET 119; Shruti Gosavi and Shraddha Kavathekar, A Survey on Crime Occurrence Detection and prediction Techniques' (2018) 8 Int. j. eng. technol. manag. appl. sci. 1405.

[10] Rembrandt Deville, Nico Sergeyssels and Catherine Middag, 'Basic concepts of AI for legal scholars' in Jan De Bruyne and Cedric Vanleenhove (eds), *Law & Artificial Intelligence* (Intersentia 2021) 1.

[11] High-Level Expert Group on Artificial Intelligence, 'Ethics Guidelines for Trustworthy AI' (*European Commission*, 8 April 2019) <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai> accessed 5 July 2021.

[12] European Commission, 'Proposal for a regulation of the European parliament and of the council laying down harmonized rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts', COM(2021), Brussels.

[13] See eg Michael Vaele and Frederik Zuiderveen Borgesius, 'Demystifying the Draft EU Artificial Intelligence, Analysing the good, the bad, and the unclear elements of the proposed approach' (2021) 22 Computer L. Rev. Int. 97; Nathalie Smuha, Emma Ahmed-Rengers, Adam Harkens, Wenlong Li, James MacLaren, Riccardo Piselli and Karen Yeung, 'How the EU can achieve Legally Trustworthy AI: A Response to the European Commission's Proposal for an Artificial Intelligence Act' (*SSRN*, 5 August 2021) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3899991&download=yes> accessed 19 October 2021.

While the debate on the definition of AI is ongoing, one can clearly observe two approaches to AI in the scientific literature, namely 'knowledge-based learning' and 'data-based learning'.

The first approach is based on explicit knowledge in the form of a model:

> Typically, an expert in the field trie[s] to pour his knowledge into a model (eg a set of rules, patterns or logical statements). This model [i]s subsequently implemented as a series of instructions – and thus as an algorithm – in the machine to obtain its goal.[14]

Under this approach, the input of the developer or programmer of the AI system is thus detrimental for its output.

In contrast, under the data-based approach, the system itself recognizes patterns, based on the numerous examples of inputs received. This approach has become predominant today and is obviously at the heart of machine learning.[15] Machine learning becomes 'deep learning'[16] when tasks become more complex and the system more autonomous and therefore, inevitably, also opaquer.[17] When such self-directed learning processes are used in a legal context, they are likely to raise significant questions. For example, risk assessment tools that predict future reoffending (*infra*) are based on machine learning techniques, involving a statistical analysis of large datasets on (past) criminal conduct.[18] However, this complexity cannot be an excuse for the opacity of the legal process, as will be discussed below.

3 AI as a Tool to Facilitate Adjudication

After these short terminological explanations, we can now focus on the first use of AI in the field of criminal justice: AI as a support tool for the adjudication of criminal cases. Three types of tools can be distinguished. In order of growing complexity, these are: 1) tools that make existing legal information more easily accessible and searchable, 2) tools that make predictions about the outcome of legal cases, and 3) tools that aim to predict human behaviour relevant for sentencing purposes.

---

[14] Rembrandt Deville, Nico Sergeyssels and Catherine Middag, 'Basic concepts of AI for legal scholars' in Jan De Bruyne and Cedric Vanleenhove (eds), *Law & Artificial Intelligence* (Intersentia 2021) 1, 4.

[15] ibid 1.

[16] The main difference between machine learning and deep learning is that machine learning algorithms will process quantitative and structured data whereas the algorithms used in deep learning will be based on unstructured data.

[17] Emmanuel Barthe, 'L'intelligence artificielle et le droit' (2017) 54 Information, données & documents 23, 24; Morgane Hubert, 'Les algorithmes prédictifs au service du juge : vers une déshumanisation de la justice pénale ? Regards critiques de juges d'instruction' (Master thesis, Catholic University of Leuven 2020) 29.

[18] ibid 28.

In the following sections, we will briefly present each type of tool on the basis of some concrete examples (Section 3.1), before examining the added value (Section 3.2) and potential risks (Section 3.3) these tools entail for the criminal justice system.

3.1 Some examples of AI support tools

3.1.1 Legal search engines

In the legal area, one of the first applications of algorithms consisted in the creation of legal databases, containing legislation, case law and/or legal literature. While some of these databases are quite simple, offering only limited search functions (eg on the basis of the date of a judgment, the case number or the parties' names), others are more sophisticated and powerful, combining several parameters. Clearly, these tools make legal information accessible in a more efficient manner. By now, they have become part of legal practitioners and judges every-day toolset to prepare their cases. Some of these databases have been developed by or for public authorities and are publicly available, free of charge; others are owned by private companies (eg legal publishers) and require prior subscription, sometimes at a relatively high price. In recent years, there seems to be a trend among public authorities to outsource the development and maintenance of certain databases to the private sector for cost-efficiency and/or technical reasons.[19] As we will discuss later, this trend raises a number of concerns.

3.1.2 Tools predicting legal outcomes

In a next step, AI tools based on assisted machine learning[20] have been developed for the purpose of predicting the most likely solution to a dispute. Thanks to various mathematical and statistical tools, it would be possible to assess, with varying degrees of accuracy, the probability of success of certain proceedings.

Examples of such prediction tools in the area of civil and administrative law are Predictice, Supra Legem or Case Law Analytics. In the case of Predictice, an algorithm will calculate the probability of resolution of a dispute as well as the range of compensation

[19] For instance, in the United States, the publication of judgments and their electronic access is largely enabled by legal publishers. Similarly, but much more recently, in Belgium, the rulings of the Court of Cassation are no longer published by the Court itself but by Larcier and the electronic access to older judgments has been made possible thanks to the efforts of the law library of the KU Leuven (see <https://justice.belgium.be/fr/ordre_judiciaire/cours_et_tribunaux/cour_de_cassation/jurisprudence> and <https:// bib.kuleuven.be/rbib/collectie/archieven/arrcass/arresten-van-het-hof-van-cassatie>). Furthermore, in order to make all judicial decisions publicly available, as the Belgian legislator ambitiously set forward in 2019, the Ministry of Justice is exploring public-private partnerships. For a critical analysis, see Jean De Codt, 'La parole du juge sous le boisseau de sa quantification numérique. À propos de la publicité des jugements à l'ère 2.0' (2021) Journal des Tribunaux 22, 24.

[20] Emmanuel Barthe, 'L'intelligence artificielle et le droit' (2017) 54 Information, données & documents 23, 24.

amounts, and all this will be exported and made comprehensible in the form of a customizable report.[21] The predictive algorithm of Supra Legem analyzes hundreds of thousands of administrative decisions (for instance, in the field of migration law), applying different criteria.[22] This analysis would allow to identify trends in case law that are otherwise invisible.[23] Furthermore, the Case Law Analytics software examines the risks regarding a contract or a litigation. According to the programme description, the combination of mathematics and law renders it possible to measure in a precise way the influence of a specific element of a case, in the decision of the judge in charge of the case, or to know how to adjust in the best way a clause in the contract.[24]

Case Crunch, yet another AI software, makes predictions about, for instance, insurance claims, based on the analysis of the outcome of past claims. In a 2017 competition, the system proved to be right about the expected outcome in 87%, compared to 62% of correct assessments by 100 top lawyers of London law firms.[25]

Similarly, in the field of labour law, Legal Insights, an AI tool developed by Wolters Kluwer for the Belgian legal market, makes predictions about the chances of winning a dismissal case.[26]

Closer to the criminal law field are studies using AI tools that involve machine learning and natural language processing to predict the outcome of a case before certain courts, such as the United States Supreme Court[27] or the European Court of Human Rights.[28] The accuracy of these predictions ranges between 70% and 79%.

Clearly, such prediction tools enable legal practitioners (and their clients) to assess, *ex ante*, the success rate of future litigation and to prepare their cases more effectively. Yet for judges too, these tools may prove to be useful as they allow them to position themselves against previous decisions taken by other courts in their legal system and thus to

[21] Predictice, 'Au Cœur de la justice' <https://predictice.com/> accessed 5 July 2021.
[22] For example, the subject matter of the decision, the characterics of the claimant as well as the defendant, and the meaning of the legal provision.
[23] Michaël Benesty, 'Supra Legem' <https://www.data.gouv.fr/fr/reuses/supra-legem/> accessed 5 July 2021.
[24] Case Law Analytics, 'Analysez votre risque juridique grâce à l'IA' <https://www.caselawanalytics.com/> accessed 5 July 2021.
[25] Case Crunch, 'Lawyer Challenge' <https://www.case-crunch.com> accessed 14 July 2021.
[26] Wolters Kluwer, 'Legal Insights' <https://www.wolterskluwer.com/fr-be/solutions/legal-insights/> accessed 14 July 2021.
[27] Daniel Martin Katz, Michael J Bommarito II and Josh Blackman, 'A General Approach for Predicting the Behavior of the Supreme Court of the United States' (*PLoS ONE*, 12 April 2017) <https://doi.org/10.1371/journal.pone.0174698> accessed 14 July 2021.
[28] Nikaloas Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro and Vasileios Lampos, 'Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective' (*PeerJ Computer Science*, 24 October 2016) <https://doi.org/10.7717/peerj-cs.93> accessed 14 July 2021.

ensure more consistency in case law, even in legal systems that do not adhere to the doctrine of binding precedent.

3.1.3 Tools predicting future behaviour of suspects or convicts

Still less frequently applied are AI tools that make *ex ante* individual[29] predictions about future behaviour, in particular criminal behaviour. These predictive algorithms have been used 'to predict (…) who is likely to fail to appear at their court hearing, and who is likely to reoffend at some point in the future.'[30] Such predictions may indeed be relevant for the execution of judicial decisions (eg parole decisions), but potentially also at the sentencing stage as they give the judge an evidence-based indication about the suspect's likelihood to reoffend. Based on this prediction, the judge could pronounce a more adequate sentence.

Among the best-known programmes is COMPAS (acronym for Correctional Offender Management Profiling for Alternative Sanctions), used in the United States. It is a decisional support tool[31] developed by a private company[32] that makes a prediction on the basis of the defendant's criminal file and a questionnaire-based interview. More precisely, 'this software predicts a defendant's risk of committing a misdemeanor or felony within 2 years of assessment from 137 features about an individual and the individual's past criminal record.'[33] These features include, for instance, age, sex and criminal history, but not race. While originally designed for pre-trial release decisions and post-sentencing decisions (eg parole), the tool's use has been gradually expanded to sentencing decisions, as the *Loomis* case (*infra*) illustrates.

COMPAS is a classic example of supervised machine learning. The algorithm is trained with past data that are analyzed on the basis of a decision tree and develops model relationships between independent and dependent variables. In a next step, this model is tested on new cases to improve its performance. Once sufficiently trained, the algorithm is applied in individual cases to determine the defendant's recidivism score on a scale of

---

[29] Unlike predictive policing tools (*supra*), which focus on geographical areas. As indicated in the introduction, this article will not focus on predictive policing tools, but rather on tools that could help judges assess the risk of future criminal behaviour to the extent that this is relevant for their decision-making in a particular case.

[30] Julia Dressel and Hany Farid, 'The accuracy, fairness, and limits of predicting recidivism' (*Science Advances*, 17 January 2018) <https://advances.sciencemag.org/content/4/1/eaao5580> accessed 5 July 2021, quoting Walter L Perry, Brian McInnis, Carter C Price, Susan Smith and John S Hollywood, *Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations* (*Rand Corp,* 2013) <https://www.rand.org/pubs/research_reports/RR233.html> accessed 5 July 2021.

[31] Interview with Angèle Christin: Hubert Guillaud, 'La justice prédictive (2/3) : prédictions et régulations' (*Le Monde*, 13 September 2017) <https://www.lemonde.fr/blog/internetactu/2017/09/13/la-justice-predictive-23-predictions-et-regulations/> accessed 5 July 2021.

[32] Originally named Northpointe, rebranded to Equivant in 2017, shortly after the critical ruling of the Wisconsin Supreme Court in the notorious *Loomis* case (*infra*).

[33] Julia Dressel and Hany Farid, 'The accuracy, fairness, and limits of predicting recidivism' (*Science Advances*, 17 January 2018) <https://advances.sciencemag.org/content/4/1/eaao5580> accessed 5 July 2021.

ten. In the *Loomis* case, the result of this assessment was added to the case file as part of the presentencing investigation report submitted to the sentencing court, which informs the latter in view of imposing a sentence.

It is important to emphasize that, even though the defendant's criminal record and his/her answers to the questionnaire are taken into account, the risk assessment is *not* specific to the defendant personal situation but based on a comparison of his/her data (at a certain point in time) with a similar data group collected in the past. Nor does the tool allow to account for future changes in the individual's personal situation or at policy-making level (eg increased efforts to reintegrate sentenced persons in society, improvements in the educational system, better economic conditions).

COMPAS is not entirely unique, but definitely the most discussed (and, as we will see, the most criticized) AI tool in its kind. Other more or less comparable tools are HART (for Hart Assessment Risk tool), used by the Durham police,[34] and CAIS (for Correctional Assessment and Intervention System), developed by the National Council on Crime and Delinquency in the United States for post-sentencing treatment by social services agencies.[35] The earlier mentioned CloudWalk technology would also enable individual risk assessments thanks to facial recognition, thereby sophisticating traditional predictive policing.[36]

All these tools are instruments that, in an ideal scenario, provide evidence-based information to judges (and other authorities) and thus enable them to make better informed, allegedly more objective and less biased decisions. Yet is this really true? This question will be answered in the next two sections.

3.2 Advantages

The application of AI as a support tool for adjudication in criminal cases definitely entails a number of advantages, as private companies developing such tools are quick to point out. Generally speaking, the appeal of algorithmic processing lies in its speed, efficiency and (apparent) objectivity.[37]

---

[34] See eg Chris Baraniuk, 'Durham Police AI to help with custody decisions' (*BBC*, 10 May 2017) <https://www.bbc.com/news/technology-39857645> accessed 14 July 2021.

[35] See eg Elgin Karls, Eric La Nguyen, Dan Spika and Kendall Vega, 'A Demonstration Analysis of the Correctional Assessment and Intervention Analysis (CAIS)', Report prepared for the National Council on Crime and Delinquency (University of Wisconsin Madison 2018) <https://lafollette.wisc.edu/images/publications/workshops/2018-NCCD_final.pdf> accessed 14 July 2021.

[36] See eg Daniel Faggella, 'AI for Crime Prevention and Detection – 5 Current Applications' (*Emerj*, 2 February 2019) <https://emerj.com/ai-sector-overviews/ai-crime-prevention-5-current-applications/> accessed 14 July 2021.

[37] Sonia Desmoulin-Canselier, 'L'emprise des algorithmes – A propos de Frank Pasquale, The Black Box Society. The Secret Algorithms That Control Money and Information' (*La vie des idées*, 20 June 2018) <https://laviedesidees.fr/L-emprise-des-algorithmes.html> accessed 14 July 2021.

The first advantage of AI as a tool in criminal justice is the underlying computing power and speed of execution. Without a doubt, the evolution of technologies, in particular machine learning, has allowed the automated processing of a large amount of data, consisting of judicial decisions, legal rules, but also examples of cases.[38] Legal search engines are a straight-forward example thereof, but only a first step. The ultimate goal and advantage would be to automate all repetitive tasks and eradicate time-consuming tasks.

Moreover, a second advantage of these intelligent decision-making tools would be the speed of the proceedings, which would have the correlative advantage of relieving the courts and which could tackle the (systemic) problem of judicial delays.

Equally importantly, the use of these tools by judges in their decision-making would also prove – and this constitutes a third advantage – to be a factor of consistency for judicial practices. This is particularly true for AI tools predicting legal outcomes in fairly simple legal proceedings that require limited human judgment. Typically, this concerns highly regulated fields of law with quite precise legal requirements such as traffic law, labour law or immigration law.

Finally, a fourth argument put forward in favour of the introduction of AI tools in the criminal justice area would be the neutrality and accuracy of these systems. For example, according to the companies developing these new software programmes, they would be more objective and accurate than human beings in analyzing a criminal's chances of reoffending. This assessment has, however, been questioned and invalidated by independent actors, as we will discuss below.

3.3 Disadvantages

Notwithstanding the above advantages, the use of AI in courts also presents various and non-negligible drawbacks, which arguably are more important in the field of criminal law than in the area of civil or administrative law. In this Section, we will identify three main disadvantages. The first one is probably intrinsic to the operation of AI systems, especially those involving machine learning, and concerns the data used for their operation. Systems that are data-based require the collection and exploitation of a large set of data. Problems may arise in the selection of data and/or in their subsequent exploitation. Second, we will address the opacity of AI systems – a problem better known under the concept of 'black box'. On the one hand, this opacity is due to the way in which AI systems are designed and function, hampering the interpretability and explicability of decisions based on algorithms. On the other hand, this 'black box' phenomenon is further increased by the fact that many AI tools are created by private companies, invoking their business secrecy to shield off the design and functioning of those systems. Third, the effectiveness and accuracy of at least some tools seem to be overrated too.

[38] Serge Abiteboul and Florence G'Sell, 'Les algorithmes pourraient-ils remplacer les juges?' *Le Big data et le droit* (Dalloz 2019) 12 <https://hal.inria.fr/hal-02304016v2/document> accessed 25 October 2021.

3.3.1 The issue of data quality and risks of errors and biases

AI systems essentially require the input and processing of data. At the level of the collection and exploitation of data, several issues may arise.

As pointed out by Arthur Holland Michel, these issues 'can be categorized as incomplete data, low quality data, incorrect or false data, and discrepant data (data that differ from the data the system was designed for).'[39] The output of the system inherently depends on the quality of the input. It is a common mistake to believe that the more data you have, the higher your success rate will be. Of course, a large amount of data is needed to train the algorithms, but if the data are of poor quality, incomplete, incorrect and/or badly transcribed, this will impact the quality of the result too.[40]

In addition, there may be errors and/or biases built into the system. Errors are generally due to the data that are used, the functioning of the algorithm, or security flaws. Biases, whether explicit or implicit, may result from the (past) data used to train the system or the factors and their relative weight in the decision tree.[41] Whereas machine learning techniques appear to be value-neutral,[42] the same cannot be said of the people who design or programme them. Consequently, 'a predictive algorithm's recommendation actually masks an underlying series of subjective judgments on the part of the system designers about what data to use, include or exclude, how to weight the data, and what information to emphasize or deemphasize.'[43] In the same way, 'being aware that these "implicit biases" exist, and that everyone possesses them – even scientists – is an important step toward drawing fair and unbiased conclusions.'[44] Therefore, once the data have been collected and integrated, it is still necessary to check whether the process is free of errors and biases, in order to achieve the expected result.

The aforementioned COMPAS software provides an apt illustration of these issues. As explained, this software was designed without taking into account the ethnicity of individuals and thus one could legitimately believe that this type of data would not play a

---

[39] Arthur Holland Michel, 'Known Unknowns: Data Issues and Military Autonomous Systems' (2021) United Nations Institute for disarmament research 1, 3.

[40] Serge Abiteboul and Florence G'Sell, 'Les algorithmes pourraient-ils remplacer les juges?' *Le Big data et le droit* (Dalloz 2019) 12 <https://hal.inria.fr/hal-02304016v2/document> accessed 25 October 2021.

[41] For other sources of bias, see eg Dana Pessach and Erez Shmueli, 'Algorithmic Fairness' (*Cornell University*, 21 January 2020) <https://arxiv.org/abs/2001.09784> accessed 21 October 2021.

[42] Yannick Meneceur, 'Les systèmes judiciaires européens à l'épreuve du développement de l'intelligence artificielle' (2018) 2 Revue pratique de la prospective de l'innovation 13.

[43] Robert Brauneis and Ellen Goodman, 'Algorithmic Transparency for the Smart City' (2018) 20 Yale J. L. & Tech. 103, 119; Harry Surden, 'Values Embedded in Legal Artificial Intelligence' (2017) Univ. Colo. Law Legal Studies 5.

[44] Hanna Wallach, 'Big Data, Machine Learning, and the Social Sciences: Fairness, Accountability, and Transparency' (*Hanna Wallach*, 19 December 2014) <https://hannawallach.medium.com/big-data-machine-learning-and-the-social-sciences-927a8e20460d> and <https://www.microsoft.com/en-us/research/publication/big-data-machine-learning-and-the-social-sciences-fairness-accountability-and-transparency/> accessed 15 July 2021.

role in the subsequent recidivism risk assessment. Nevertheless, a study by ProPublica, an independent non-profit newsroom, revealed that, indirectly, ethnicity did matter and even outweighed other explicitly included factors due to the cross-referencing of different data such as place of residence or profession, but also because certain ethnicities were overrepresented in the data used to train the system.[45] As a result, black suspects were more likely to obtain a high-risk score than white ones. This shows the risk of biases is significant and not easy to address.[46]

The use of COMPAS in the notorious *Loomis* case,[47] however, also shows that the quality (and accuracy) of the risk assessment is determined by the quality of the collected data. In particular, the Wisconsin Supreme Court pointed out that the tool had been trained with group data collected nation-wide, without cross-validation for the Wisconsin population.[48] Consequently, certain (demographic, social, economic, legal, etc.) specificities of the State in which the tool was implemented might not be sufficiently reflected. Therefore, the Wisconsin Supreme Court required that the presentencing investigation report including a risk assessment based on the COMPAS tool entails a notification of the limitations of the system.

Moving beyond the COMPAS example, in some legal systems, there are simply not enough and/or not sufficiently precise data available to feed a recidivism prediction tool.[49] What is more, AI tools trained with past data inevitably create models that replicate what happened in the past. They are unable to account for new policies or circumstances. Therefore, the data used to train the AI system need to be updated regularly to account for changing legislation and policies, even if they are not directly linked to sen-

---

[45] Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, 'Machine Biais – There's software used across the country to predict future criminals. And it's biased against blacks' (*ProPublic*a, 23 May 2016) <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> accessed 5 July 2021.

[46] As Angèle Christin points out, 'many of these tools today are tainted by a suspicion of unconstitutionality and removing the postcode or job will not remove their bias, particularly because the history of data and judicial systems is steeped in discrimination.' Interview with Angèle Christin: Hubert Guillaud, 'La justice prédictive (2/3): prédictions et régulations' (*Le Monde*, 13 September 2017) <https://www.lemonde.fr/blog/internetactu/2017/09/13/la-justice-predictive-23-predictions-et-regulations/> accessed 5 July 2021.

[47] For a summary of the case, see Adrien van den Branden, *Les robots à l'assaut de la justice* (1st ed., Bruylant 2019) 4-5. For a critical analysis, see Katherine Freeman, 'Algorithmic Injustice: How the Wisconsin Supreme Court Failed to Protect Due Process Rights in State v. Loomis' (2016) 18 N.C. J.L. & Tech. On. 75.

[48] Wisconsin Supreme Court, State v. Loomis, 881 N.W.2d 749, 770–71 (Wis. 2016).

[49] In Belgium for instance, there is (still) no nation-wide database on criminal convictions and subsequent reoffending. Available empirical data are based on criminal records of individual convicts (which contain limited information and, until recently, were not always up-to-date or fully correct) and on dedicated academic studies (which usually focus on specific populations and/or types of offences, and which often use different concepts of repeat offending). For a (fairly) recent study, see Benjamin Mine and Luc Robert, 'Recidive na een rechterlijke beslissing. Nationale cijfers op basis van het Centraal Strafregister'/'La récidive après une décision judiciaire. Des chiffres nationaux sur la base du Casier judiciaire central' (2015) Final report, Institut National de Criminalistique et de Criminologie 1, 12-27 <https://nicc.fgov.be/upload/publicaties/rapport_38_1.pdf> accessed 25 October 2021.

tencing and the execution of sentences, such as better education or an improved economic situation. Yet, even then, AI may 'blin[d] us to everything that is not quantifiable and digitizable.'[50]

Therefore, all in all, such prediction tools should be used with particular precaution and judges should be duly informed of the origin and quality of the data, and the way in which they are processed.

For less sophisticated AI support tools, such as legal search engines, the problem regarding the quality of the data is likely to be more easily detectable and can be more easily remedied. Imagine a legal database that only contains judgments of one single court or a limited number of journals. Then, it is quite obvious that the search results will be less rich and will only offer a partial understanding of the matter than with a legal database covering all case law and a large body of legal literature.

3.3.2 Opacity of AI systems

Machine learning systems as inherent black boxes?

Originally referring to the on-board recording system in transportation means, the term 'black box' also refers to an opaque and closed device, inaccessible to the eye.[51] When transposed to AI systems, black box means that 'the processes happening inside of them are difficult – and sometimes impossible – to fully understand.'[52] The problem of opacity is particularly pressing in case of machine learning, especially systems based on neural networks.[53]

The black box problem is intrinsically linked to the requirements of explanation and justification, which are highly important in a legal context and even more so in criminal cases where the defendant's liberty and other fundamental rights are at risk. Considering that justice must not only be done, but also seen to be done, it is crucial that the technical functioning of the tool, the data selected by the algorithm, the decision-making factors and process can be explained in a language that all parties are able to understand.[54] Put

[50]Free translation from the original quote « qui nous rend aveugles à tout ce qui n'est pas quantifiable ou numérisiable ». See Sonia Desmoulin-Canselier, 'L'emprise des algorithmes – A propos de Frank Pasquale, 'The Black Box Society. The Secret Algorithms That Control Money and Information' (*La vie des idées*, 20 June 2018) <https://laviedesidees.fr/L-emprise-des-algorithmes.html> accessed 27 July 2021.

[51] ibid.

[52] Mira Ortegon, 'Dismantling the Black Box: Why Governments Should Demand Algorithmic Accountability' (*Brown Political Review*, 30 March 2019) <https://brownpoliticalreview.org/2019/03/dismantling-black-box-governments-demand-algorithmic-accountability/> accessed 25 July 2021.

[53] Moustafa Zouinar, 'Evolutions de l'Intelligence Artificielle : quels enjeux pour l'activité humaine et la relation Humain-Machine au travail ?' (*Open Edition Journals*, 15 April 2020) <https://journals.openedition.org/activites/4941#quotation> accessed 21 June 2021. See also Jenna Burrell, 'How the Machine Thinks: Understanding Opacity in Machine Learning Algorithms' (2016) Big Data & Society 1.

[54] Lêmy Godefroy, 'L'office du juge à l'épreuve de l'algorithme' in Jean-Pierre Clavier (eds), *L'algorithmisation de la justice* (Larcier 2020) 111.

differently, it must be possible to 'reconstruct the relevance of the pathways'[55] operated by the algorithm. Yet, most importantly, judges need to understand the functioning of the AI system that supports them in order to be able to assess its added value. Without such understanding, the prediction made by the system, whether regarding the probable outcome of a case or the risk of reoffending, could lead to undesirable situations and ill-founded or even arbitrary decision-making, heavily impacting the future of the convicted person.[56]

Again, the *Loomis* case provides a good illustration of this problem.[57] Mr Loomis argued that his due process (or, transposed to the European legal context, fair trial) rights had been violated because the sentencing court based its decision on the risk assessment made by COMPAS, notwithstanding it was impossible, for him and the court, to review how factors are weighed by the system and how risk scores are produced. Since he could not review the functioning of the AI system, it was also impossible for him to challenge the risk score the system had put forward. The Wisconsin Supreme Court dismissed this argument, but in a concurring opinion, Justice Abraham pointed out that the 'lack of understanding of COMPAS was a significant problem in the instant case' and argued that 'in considering COMPAS (or other risk assessment tools) in sentencing, a circuit court must set forth on the record a *meaningful process of reasoning* addressing the relevance, strengths, and weaknesses of the risk assessment tool.'[58]

Other positive examples, imposing higher transparency requirements upon the use of AI systems, can be found outside the ('hard core'[59]) criminal law field. Certain public administrations that use AI tools to detect administrative infringements (eg illegal parking or illegal renting of holiday housing) have started to publish so-called 'algorithm registers' providing more information on the systems that are used and their functioning.[60] These examples prove that opacity of AI systems is not a kind of fatality, but an issue that can be addressed if there is a will to do so.[61]

---

[55] Emmanuel Poinas, *Le tribunal des algorithme. Juger à l'ère des nouvelles technologies* (Berger Levrault 2019) 240.

[56] Yannick Meneceur, 'Les systèmes judiciaires européens à l'épreuve du développement de l'intelligence artificielle' (2018) 2 Revue pratique de la prospective de l'innovation 7, 14.

[57] See also Alyssa M. Carlson, 'The Need for Transparency in the Age of Predictive Sentencing Algorithms' (2017) 103 Iowa L. Rev. 303.

[58] Emphasis added.

[59] *Jussila v Finland* App no 73053/01 (ECtHR, 23 November 2006), para 43. For a comprehensive analysis of the difference between criminal law and quasi-criminal law, see Vanessa Franssen and Christopher Harding (eds), *Criminal and Quasi-criminal Enforcement Mechanisms in Europe. Origins, Concepts, Future* (Hart Publishing *forthcoming* 2022).

[60] See eg the algorithm registers published by the cities of Amsterdam and Helsinki <https://algoritme-register.amsterdam.nl/en/ai-register/> ; <https://ai.hel.fi/en/ai-register/> accessed 15 July 2021.

[61] Nazrin Huseinzade, 'Algorithm transparency: How to Eat the Cake and Have It Too' (*European Law Blog*, 27 January 2021) <https://europeanlawblog.eu/2021/01/27/algorithm-transparency-how-to-eat-the-cake-and-have-it-too/#comments> accessed 21 July 2021.

An additional complicating factor in this black box discussion is, however, that AI systems are often developed by private companies and that these actors are reluctant to communicate about the functioning of their systems, hiding behind business secrecy to protect their proprietary code. As a result, 'transparency is hard to come by.'[62] Once again, the *Loomis* case aptly illustrates this issue. The proprietary nature of COMPAS prevented the sentencing court and Mr Loomis to understand how the system functions, to evaluate its scientific validity (or accuracy) and, subsequently, its contestability. Although the Wisconsin Supreme Court recognised the software's limitations, it found that Northpointe had a clear financial interest in not disclosing the algorithm that it had itself developed.[63] In addition, the company also pointed out that with knowledge of the algorithm, criminals could potentially distort the risk assessment and exploit the model to their advantage, which would make the algorithm ineffective.[64]

Moreover, AI systems provided by the private sector also create other impediments to information. As indicated above, legal search engines have become an indispensable tool for legal practitioners and judges to prepare their cases. The more powerful ones are, however, expensive and thus not accessible to all. Suspects who are able to afford a lawyer from a top law firm are thus likely to get a better defence. Considering the criminal justice system is often underfunded, judges too will probably have more limited access to useful legal information. If one adds to this the trend to outsource certain public services, including databases providing access to case law, one will understand why the argument that AI renders justice more cost-efficient should be somewhat mitigated.

In conclusion, it seems fundamental that the algorithmic tools underlying legal decisions are accessible, or at least, explicable, at the end of the process. The complexity of the decision-making process should not be used as an excuse for the opacity of the system, and private companies' business interests should be adequately balanced against suspects' fundamental rights.[65]

3.3.3 Overstated effectiveness

Finally, a third concern is the accuracy and effectiveness of AI tools predicting legal outcomes and future risks. While the developers of AI tools like to emphasize how accurate

[62] Mira Ortegon, 'Dismantling the Black Box: Why Governments Should Demand Algorithmic Accountability' (*Brown Political Review*, 30 March 2019) <https://brownpoliticalreview.org/2019/03/dismantling-black-box-governments-demand-algorithmic-accountability/> accessed 21 July 2021.

[63] Adrien van den Branden, *Les robots à l'assaut de la justice* (1st ed., Bruylant 2019) 8.

[64] ibid.

[65] For some interesting examples outside the criminal law field on how data protection rules can be used to enhance AI transparency, see Nazrin Huseinzade, 'Algorithm transparency: How to Eat the Cake and Have It Too' (*European Law Blog*, 27 January 2021) <https://europeanlawblog.eu/2021/01/27/algorithm-transparency-how-to-eat-the-cake-and-have-it-too/#comments> accessed 21 July 2021.

and reliable the predictions made by these tools are, independent studies show a different reality.

For instance, with respect to the COMPAS tool, an academic study published in 2018 analyzed the reliability of the software by comparing the results obtained by the COMPAS to those of different individuals with no or very little legal background, estimating the likelihood of reoffending by the same convicted persons. To this end, randomly selected individuals were asked one simple question: 'Do you think this person will commit another crime within two years?' In contrast, COMPAS analyzed 137 factors to arrive at a conclusive result. It turned out that the accuracy of both assessments, made by AI and human beings, was very similar: While COMPAS obtained a 65,2% efficiency score, the human beings were right in 67% of the cases. Thus, even people without any criminal justice expertise would have achieved the same result as a computer programme designed and trained to make such risk assessments.[66]

As regards AI tools predicting legal outcomes, it is important to highlight that they are usually designed for very specific legal disputes, typically in fields of law that are highly regulated and technical, thus requiring limited human judgment (eg traffic law), or disputes where the number of legal factors to account for is relatively limited (eg insurance claims, labour law disputes, return decisions in the field of migration law). The more complex the nature of the dispute and the more open-ended legal norms are, the less likely such tools will lead to a satisfactory prediction. This point will be further elaborated in the next part.

4 Judge-Made vs AI-Made Criminal Justice

In this last part, we will take the analysis a step further and focus on AI tools that, rather than supporting human judges, could replace them. While examples of such 'robot judges' are still fairly limited[67] and, to our knowledge, inexistent in the field of criminal justice, we can nevertheless already examine, on a theoretical level, some advantages these tools are likely to bring compared to a human judge (Section 4.1). Next, we will present a number of disadvantages that can result from automated adjudication (Section 4.2).

4.1 Advantages

A robot judge potentially has several advantages over the human judge, in particular in terms of consistency in decision-making, reliability, cost and speed.

[66] Julia Dressel and Hany Farid, 'The accuracy, fairness, and limits of predicting recidivism' (*Science Advances*, 17 January 2018) <https://advances.sciencemag.org/content/4/1/eaao5580> accessed 5 July 2021.

[67] So far, some proposals have been made for disputes with small values and to extend some online dispute resolution mechanisms. See eg Michael Grupp 'How to Build a Robot Lawyer' in Markus Hartung, Micha-Manuel Bues and Gernot Halbleib (eds), *Legal Tech. How Technology is Changing the Legal World. A Practioner's Guide* (Beck-Hart-Nomos 2018) 249, 257.

First, in terms of consistency and reliability, it is well-known that human judges are not in a constant mood and thus that their decision may vary depending on the day and the hour of the hearing or decision, or the events that may occur in their lives.[68] In contrast, the algorithm will work every day, at every time of day, in the same way. This coherence is based on the proper functioning of the AI system; it is indeed a safe bet that systems implementing algorithmic justice will base their analysis on the same data recorded in their digital library, even the least known. Consistent decision-making contributes to the coherence and stability of the legal system, and thereby enhances legal certainty.

Next, robot judges may be more cost-efficient than human judges. For sure, the initial investments may be high, but once those are made, the robot judge has a lower operational cost than a human one, for a similar number of cases handled.[69] Moreover, thanks to the robot judge, physical trials and the costs related to their organisation could perhaps be abandoned in a number of cases (to the extent, of course, that this is compatible with the right to a fair trial),[70] which would lead to a substantial reduction in the costs of legal proceedings.[71]

Finally, we could mention the speed and effectiveness of robot judges. The AI system would outperform the human judge for certain tasks, in particular for dealing with repetitive demands and the calculation of damages or interests.

4.2 Disadvantages

Despite the aforementioned advantages, the disadvantages of AI decision-making should neither be neglected, nor underestimated.

4.2.1 Law is not code

First of all, our law contains a lot of open-ended norms, requiring a certain degree of human appreciation. To take just one example, the application of fundamental rights

[68] Shai Danziger, Jonathan Levav and Liora Avnaim-Pesso, 'Extraneous factors in judicial decision' (*PNAS*, 26 April 2011) <http://www.pnas.org/content/108/17/6889> accessed 5 July 2021; Serge Abiteboul and Florence G'Sell, 'Les algorithmes pourraient-ils remplacer les juges?' *Le Big data et le droit* (Dalloz 2019) 9 <https://hal.inria.fr/hal-02304016v2/document> accessed 25 October 2021.

[69] Adrien van den Branden, *Les robots à l'assaut de la justice* (1st ed., Bruylant 2019) 48.

[70] The European Court of Human Rights has ruled on a number of occasions that for some criminal cases an oral hearing may not be indispensable, in particular 'where there are no issues of credibility or contested facts which necessitate a hearing,' where an oral hearing would be 'an obstacle to the particular diligence required' in certain cases and where the legal questions raised are not particularly complicated. In such cases, 'the courts may fairly and reasonably decide the case on the basis of the parties' submissions and other written materials.' See eg *Jussila v Finland* App no 73053/01 (ECtHR, 23 November 2006), paras 40-48; *Nusret Kaya and Others v Turkey* App nos 43750/06, 43752/06, 32054/08, 37753/08 and 60915/08 (ECtHR, 22 April 2014), para 84. This case law may offer useful guidance when considering to introduce robot judges in certain criminal cases.

[71] Loïck Gérard and Dominique Mougenot, 'Titre 1 – Justice robotisée & droits fondamentaux' in Jean-Benoit Hubin, Hervé Jacquemin and Benoît Michaux (eds), *Le juge et l'algorithme : juges augmentés ou justice diminuée ?* (1st ed., CRIDS 2019) 39.

such as the right to a fair trial or the right to privacy implies a balancing of interests and almost inevitably entails a certain level of uncertainty as to the outcome of a particular case. Such norms cannot easily be coded in algorithmic language.[72]

What is more, legal norms are based on values and social norms. When applied by human judges, 'they identify which values are at stake in a given decisional environment and ask, where necessary, if those values have been properly balanced.'[73] Yet when designers and programmers of AI systems are led to arbitrate on values, the importance of legal rules and the interpretation of these rules, this may be highly problematic, especially since they have not been legally trained and are thus unfamiliar with the technique of law and its purpose.[74]

4.2.2 Static decision-making

As explained above, data-driven learning systems are, for the most part, backward-looking: 'algorithms can only learn from existing datasets, which are grounded in past experiences and past trends.'[75] This has the advantage of reliability and predictability, but also entails the risk of static decision-making. Indeed, an AI system will struggle when faced with new situations or evolving legal norms. One may thus wonder whether AI is capable of producing new and innovative decisions. For instance, the European Convention on Human Rights constitutes a living instrument and the European Court of Human Rights regularly adjusts its case law to keep pace with societal evolutions, as well as international and national law. Recent examples concern the principle of *non bis in idem*[76] and the protection of the right to privacy in the context of mass surveillance practices.[77]

---

[72] Kareb Hao, 'AI still doesn't have the common sense to understand human language' (*MIT Technology Review*, 31 January 2020) <https://www.technologyreview.com/2020/01/31/304844/ai-common-sense-reads-human-language-ai2/> accessed 14 July 2021.

[73] Kiel Brennan-Marquez, 'Plausible Cause: Explanatory Standards in the Age of Powerful Machines' (2017) 70 Vand. L. Rev. 1249.

[74] Engerrand Marique and Alain Strowel, 'Gouverner par la loi ou les algorithmes: de la norme générale de comportement au guidage rapproché des conduites' (2017) 10 Dalloz IP/IT 517, 521.

[75] Mira Ortegon, 'Dismantling the Black Box: Why Governments Should Demand Algorithmic Accountability' (*Brown Political Review*, 30 March 2019) <https://brownpoliticalreview.org/2019/03/dismantling-black-box-governments-demand-algorithmic-accountability/> accessed 25 July 2021.

[76] *A and B v Norway* App nos 24130/11 and 29758/11 (ECtHR, 15 November 2016). For a critical analysis of the Court's new approach, see Michiel Luchtman, 'Ne bis in idem at the Interface of Administrative and Criminal Law Enforcement – Sufficiently Connected in Substance, Time and Space?' (2019) 90 Revue internationale de droit pénal 335, 343-347.

[77] *Big Brother Watch and Others v The United Kingdom* App nos 58170/13, 62322/14 and 24960/15 (ECtHR, 25 May 2021). For a critical analysis of the Court's new approach, see Nóra Ni Loideain, 'Not So Grand: The Big Brother Watch ECtHR Grand Chamber Judgment' (*Information Law and Policy Centre*, 28 May 2021) <https://infolawcentre.blogs.sas.ac.uk/2021/05/28/not-so-grand-the-big-brother-watch-ecthr-grand-chamber-judgment/> accessed 25 October 2021.

Slightly older examples regard the acceptability of irreducible life sentences without parole possibility in terrorism and other serious crime cases in the light of Article 3 of the European Convention on Human Rights.[78]

4.2.3 Independence and impartiality

In the context of algorithmic justice, the questions of independence and impartiality of the robot judge cannot be ignored. Impartiality refers to the judge's own qualities and denotes the absence of prejudice or bias. It has a *subjective* (ie with regard to the judge's personal behaviour) and an *objective* dimension (ie the appearance created toward the parties or the general public).[79] Independence is more about the status or position of the judge, especially in relation to other branches of power, but also in relation to the influence of third parties.

When transposed to robot judges, one could believe that the digital process is free of any reproach. The robot judge would be incorruptible, emotionless, uninfluenced and neutral. But is this really true?

On the one hand, the impartiality of the AI system can be problematic, particularly because of the opacity of the system and its operation (*supra*). From the outside, this process seems inaccessible or impenetrable, and may thus create the impression of arbitrary or unfair decision-making.[80] This impression is likely to be exacerbated by the aforementioned risk of systemic biases.[81] On the other hand, with respect to independence, how can we check and control the private actors who finance and design the AI decision-making tools to make sure they do not influence judicial decisions? Is there not a risk that justice will lose its independence to a private justice system, intrinsically linked to the companies that create the algorithms?[82]

As the Consultative Committee of European Judges pointed out, these technologies 'must not prevent judges from applying the law independently and impartially (...). An

---

[78] See eg *Babar Ahmad and Others v The United Kingdom* App nos 24027/07, 11949/08, 36742/08, 66911/09 and 67354/09 (ECtHR, 10 April 2012); *Vinter and Others v The United Kingdom* App nos 66069/09, 130/10 and 3896/10 (ECtHR, 9 July 2013).

[79] For a fairly recent criminal case where the Grand Chamber found a violation of the impartiality requirement, see *Morice v France* App no 29369/10 (ECtHR, 23 April 2015).

[80] See eg Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford and Hanna Wallach, 'A Reductions Approach to Fair Classification' (*PMLR*, 2018) 1 <http://proceedings.mlr.press/v80/agarwal18a.html> accessed 21 October 2021.

[81] That said, interpretability and understandability of the AI system alone will not solve the problem of biases. See eg Connor O'Sullivan, 'Interpretability in Machine Learning' (*Towards data science*, 21 October 2020) <https://towardsdatascience.com/interpretability-in-machine-learning-ab0cf2e66e1> accessed 25 October 2021; Connor O'Sullivan, 'What is Algorithm Fairness? An introduction to the field that aims at understanding and preventing bias in machine learning models' (*Towards data science*, 5 March 2021) <https://towardsdatascience.com/what-is-algorithm-fairness-3182e161cf9f> accessed 25 October 2021.

[82] Morgane Hubert, 'Les algorithmes prédictifs au service du juge : vers une déshumanisation de la justice pénale ? Regards critiques de juges d'instruction' (Master thesis, Catholic University of Leuven 2020) 48, 49.

excessive dependence on technology and those who control it is a risk for justice. (...) Judges should not be subject, for reasons of efficiency alone, to the imperatives of technology and those who control technology.'[83]

4.2.4 Transparency and contestability

When a suspect is convicted or acquitted, it is important for him/her to understand that decision, either to accept it or, on the contrary, to identify the means that would allow him/her to contest it. The court's reasoning is essential to respect the right to a fair trial, including the equality of arms, the adversarial principle, and the right to a legal remedy. For this reason, the General Data Protection Regulation and the so-called Law Enforcement Directive are very strict on applying automated decision-making to individuals.[84]

To the extent AI tools are used in the decision-making process, it is therefore crucial that the defendant has access to all the stages of the algorithmic reasoning that led to the decision. As explained earlier, this is far from obvious due to the black box phenomenon and the private interests of the company that developed the AI tools. Therefore, it is important to minimize the risks generated by the multiplication of black boxes that algorithmically determine individual trajectories. A balance must be struck between business secrecy protecting the creation of software by private companies and the transparency requirements intended to protect the individual.[85] Finally, it is also necessary to rethink and establish control and surveillance mechanisms to prevent justice from moving from the public to the private domain.[86]

5 Conclusion

As we have seen, AI tools can have substantial advantages, but also considerable disadvantages, both for justice in general and for individual defendants, that should not be minimized. However, it would be a mistake to categorically reject these AI tools in the area of criminal justice. It is indeed necessary to strike the right balance, which would

[83] Consultative Committee of European Judges, 'Justice et technologies de l'information', Opinion No 14, 2011, paras 8 and 34; Commission européenne pour l'efficacité de la justice, 'Lignes directrices sur la conduite du changement vers la Cyberjustice' (*CEPEJ-GT-QUAL*, 7 December 2016) <https://rm.coe.int/1680748154#_Toc461547117> accessed 23 June 2021.

[84] 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), recital 71 and art 22; 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, recital 38 and art 11.

[85] Georges De Leval and Frédérique Georges, *Droit judiciaire, t. 1, Institutions judiciaires et éléments de compétence* (Larcier 2014) 101.

[86] Morgane Hubert, 'Les algorithmes prédictifs au service du juge : vers une déshumanisation de la justice pénale ? Regards critiques de juges d'instruction' (Master thesis, Catholic University of Leuven 2020) 96.

combine the positive elements of the one and the other and limit their negative features. If these tools can provide valuable support to human judges on the basis of reliable, transparent and verifiable information, one should not deprive criminal justice actors of these instruments. These tools could constitute an additional guarantee, not a danger for parties, provided they are properly mastered.[87] Ultimately, 'if we can preserve a human-centric approach to leveraging this powerful technology, predictive analysis can be one of the building blocks of a more transparent and more efficient legal system.'[88]

References

Abiteboul S and G'Sell F, 'Les algorithmes pourraient-ils remplacer les juges?' *Le Big data et le droit* (Dalloz 2019) <https://hal.inria.fr/hal-02304016v2/document> accessed 25 October 2021

Agarwal A, Beygelzimer A, Dudík M, Langford J and Wallach H, 'A Reductions Approach to Fair Classification' (*PMLR*, 2018) 1 <http://proceedings.mlr.press/v80/agarwal18a.html> accessed 21 October 2021

Aletras N, Tsarapatsanis D, Preoţiuc-Pietro D and Lampos V, 'Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective' (*PeerJ Computer Science*, 24 October 2016) <https://doi.org/10.7717/peerj-cs.93> accessed 14 July 2021

Angwin J, Larson J, Mattu S and Kirchner L, 'Machine Biais – There's software used across the country to predict future criminals. And it's biased against blacks' (*ProPublic*a, 23 May 2016) <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> accessed 5 July 2021

Baraniuk C, 'Durham Police AI to help with custody decisions' (*BBC*, 10 May 2017) <https://www.bbc.com/news/technology-39857645> accessed 14 July 2021

Barthe E, 'L'intelligence artificielle et le droit' (2017) 54 Information, données & documents 23

Bayamlioglu E and Leenes R, The "rule of law" implications of data-driven decision-making: a techno-regulatory perspective' (2018) Law, Innov Technol 295

Benesty M, 'Supra Legem' <https://www.data.gouv.fr/fr/reuses/supra-legem/> accessed 5 July 2021

[87] Serge Abiteboul and Florence G'Sell, 'Les algorithmes pourraient-ils remplacer les juges?' *Le Big data et le droit* (Dalloz 2019) 14-15 <https://hal.inria.fr/hal-02304016v2/document> accessed 25 October 2021.

[88] Roland Vogl, 'Legal Tech in the USA' in Markus Hartung, Micha-Manuel Bues and Gernot Halbleib (eds), *Legal Tech. How Technology is Changing the Legal World. A Practioner's Guide* (Beck-Hart-Nomos 2018) 380, 392.

Brauneis R and Goodman E, 'Algorithmic Transparency for the Smart City' (2018) 20 Yale J. L. & Tech. 103

Brennan-Marquez K, 'Plausible Cause: Explanatory Standards in the Age of Powerful Machines' (2017) 70 Vand. L. Rev. 1249

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

Carlson A, 'The Need for Transparency in the Age of Predictive Sentencing Algorithms' (2017) 103 Iowa L. Rev. 303

Case Crunch, 'Lawyer Challenge' <https://www.case-crunch.com> accessed 14 July 2021

Case Law Analytics, 'Analysez votre risque juridique grâce à l'IA' <https://www.caselawanalytics.com/> accessed 5 July 2021

Commission européenne pour l'efficacité de la justice, 'Lignes directrices sur la conduite du changement vers la Cyberjustice' (*CEPEJ-GT-QUAL*, 7 December 2016) <https://rm.coe.int/1680748154#_Toc461547117> accessed 23 June 2021

Consultative Committee of European Judges, 'Justice et technologies de l'information', Opinion No 14, 2011

Danziger S, Levav J and Avnaim-Pesso L, 'Extraneous factors in judicial decision' (*PNAS*, 26 April 2011) <http://www.pnas.org/content/108/17/6889> accessed 5 July 2021

De Leval G and Georges F, *Droit judiciaire, t. 1, Institutions judiciaires et éléments de compétence* (Larcier 2014)

Desmoulin Canselier S, 'L'emprise des algorithmes – A propos de Frank Pasquale, The Black Box Society. The Secret Algorithms That Control Money and Information' (*La vie des idées*, 20 June 2018) <https://laviedesidees.fr/L-emprise-des-algorithmes.html> accessed 14 July 2021

Deville R, Sergeyssels N and Middag C, 'Basic concepts of AI for legal scholars' in De Bruyne J and Vanleenhove C (eds), *Law & Artificial Intelligence* (Intersentia 2021)

Dressel J and Farid H, 'The accuracy, fairness, and limits of predicting recidivism' (*Science Advances*, 17 January 2018) <https://advances.sciencemag.org/content/4/1/eaao5580> accessed 5 July 2021

European Commission, 'Proposal for a regulation of the European parliament and of the council laying down harmonized rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts', COM(2021), Brussels

Faggella D, 'AI for Crime Prevention and Detection – 5 Current Applications' (*Emerj*, 2 February 2019) <https://emerj.com/ai-sector-overviews/ai-crime-prevention-5-current-applications/> accessed 14 July 2021

Freeman K, 'Algorithmic Injustice: How the Wisconsin Supreme Court Failed to Protect Due Process Rights in State v. Loomis', 2016 18 N.C. J.L. & Tech. On. 75

Gérard L and Mougenot D, 'Titre 1 – Justice robotisée & droits fondamentaux' in Hubin J-B, Jacquemin H and Michaux B (eds), *Le juge et l'algorithme : juges augmentés ou justice diminuée ?* (1st ed., CRIDS 2019)

Godefroy L, 'L'office du juge à l'épreuve de l'algorithme' in Clavier J-P (ed.), *L'algorithmisation de la justice* (Larcier 2020)

Gosavi S and Kavathekar S, A Survey on Crime Occurrence Detection and prediction Techniques' (2018) 8 Int. j. eng. technol. manag. appl. sci. 1405

Grupp M, 'How to Build a Robot Lawyer' in Hartung M, Bues M-M and Halbleib G (eds), *Legal Tech. How Technology is Changing the Legal World. A Practioner's Guide* (Beck-Hart-Nomos 2018)

Guillaud H, 'La justice prédictive (2/3) : prédictions et régulations' (*Le Monde*, 13 September 2017) <https://www.lemonde.fr/blog/internetactu/2017/09/13/la-justice-predictive-23-predictions-et-regulations/> accessed 5 July 2021

Hao K, 'AI still doesn't have the common sense to understand human language' (*MIT Technology Review*, 31 January 2020) <https://www.technologyreview.com/2020/01/31/304844/ai-common-sense-reads-human-language-ai2/> accessed 14 July 2021

Hérard F, 'Technologies de prédiction du crime: Palantir a scruté les citoyens de la Nouvelle-Orléans en secret pendant 6 ans' (*TV5Monde*, 3 March 2018) <https://information.tv5monde.com/info/technologies-de-prediction-du-crime-palantir-scrute-les-citoyens-de-la-nouvelle-orleans-en> accessed 14 July 2021

Holland Michel A, 'Known Unknowns: Data Issues and Military Autonomous Systems' (2021) United Nations Institute for disarmament research 1

Hubert M, 'Les algorithmes prédictifs au service du juge : vers une déshumanisation de la justice pénale ? Regards critiques de juges d'instruction' (Master thesis, Catholic University of Leuven 2020)

Huseinzade N, 'Algorithm transparency: How to Eat the Cake and Have It Too' (*European Law Blog*, 27 January 2021) <https://europeanlawblog.eu/2021/01/27/algorithm-transparency-how-to-eat-the-cake-and-have-it-too/#comments> accessed 21 July 2021

Karls E, La Nguyen E, Spika D and Vega K, 'A Demonstration Analysis of the Correctional Assessment and Intervention Analysis (CAIS)', Report prepared for the National Council on Crime and Delinquency (*University of Wisconsin Madison*, 2018) <https://lafollette.wisc.edu/images/publications/workshops/2018-NCCD_final.pdf> accessed 14 July 2021

Katz D M, Bommarito II M J and Blackman J, 'A General Approach for Predicting the Behavior of the Supreme Court of the United States' (*PLoS ONE*, 12 April 2017) <https://doi.org/10.1371/journal.pone.0174698> accessed 14 July 2021

Luchtman L, 'Ne bis in idem at the Interface of Administrative and Criminal Law Enforcement – Sufficiently Connected in Substance, Time and Space?' (2019) 90 Revue internationale de droit pénal 335

Marique E and Strowel A, 'Gouverner par la loi ou les algorithmes: de la norme générale de comportement au guidage rapproché des conduites' (2017) 10 Dalloz IP/IT 517

Meneceur Y, 'Les systèmes judiciaires européens à l'épreuve du développement de l'intelligence artificielle' (2018) 2 Revue pratique de la prospective de l'innovation 13

Ni Loideain N, 'Not So Grand: The Big Brother Watch ECtHR Grand Chamber Judgment' (*Information Law and Policy Centre*, 28 May 2021) <https://infolawcentre.blogs.sas.ac.uk/2021/05/28/not-so-grand-the-big-brother-watch-ecthr-grand-chamber-judgment/> accessed 25 October 2021

Mine B and Robert L, 'Recidive na een rechterlijke beslissing. Nationale cijfers op basis van het Centraal Strafregister'/'La récidive après une décision judiciaire. Des chiffres nationaux sur la base du Casier judiciaire central'(2015) Final report, Institut National de Criminalistique et de Criminologie 1 <https://nicc.fgov.be/upload/publicaties/rapport_38_1.pdf> accessed 25 October 2021

O'Neil C, *Algorithmes – La Bombe à retardement* (Les Arènes 2018)

O'Sullivan C, 'Interpretability in Machine Learning' (*Towards data science*, 21 October 2020) <https://towardsdatascience.com/interpretability-in-machine-learning-ab0cf2e66e1> accessed 25 October 2021

—— 'What is Algorithm Fairness? An introduction to the field that aims at understanding and preventing bias in machine learning models' (*Towards data science*, 5 March 2021) <https://towardsdatascience.com/what-is-algorithm-fairness-3182e161cf9f> accessed 25 October 2021

Oosterloo S and Van Schie G, 'The Politics and Biases of the "Crime Anticipation System" of the Dutch Police' in Bates J, Clough P D, Jäschke R and Otterbacher J (eds), *Proceedings of the International Workshop on Bias in Information, Algorithms, and Systems* (CEUR Workshop Proceedings 2018) 30

Ortegon M, 'Dismantling the Black Box: Why Governments Should Demand Algorithmic Accountability' (*Brown Political Review*, 30 March 2019) <https://brownpoliticalreview.org/2019/03/dismantling-black-box-governments-demand-algorithmic-accountability/> accessed 25 July 2021

Patil R, Kacchi M, Gavali P and Pimparia K, 'Crime Pattern Detection, Analysis & Prediction using Machine Learning' (2020) 7 IRJET 119

Perry W L, McInnis B, Price C C, Smith S and Hollywood J S, *Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations* (*Rand Corp*, 2013) <https://www.rand.org/pubs/research_reports/RR233.html> accessed 5 July 2021

Pessach D and Shmueli E, 'Algorithmic Fairness' (*Cornell University*, 21 January 2020) <https://arxiv.org/abs/2001.09784> accessed 21 October 2021

Poinas E, *Le tribunal des algorithme. Juger à l'ère des nouvelles technologies* (Berger Levrault 2019)

Predictice, 'Au Cœur de la justice' <https://predictice.com/> accessed 5 July 2021

PredPol, 'What. Where. When' <https://www.predpol.com/> accessed 14 July 2021

Shah N, Bhagat N and Shah M, 'Crime forecasting: a machine learning and computer vision approach to crime prediction prevention' (2021) 4 VCIBA 9

Smuha N, Ahmed-Rengers E, Harkens A, Li W, MacLaren J, Piselli R and Yeung K, 'How the EU can achieve Legally Trustworthy AI: A Response to the European Commission's Proposal for an Artificial Intelligence Act' (*SSRN*, 5 August 2021) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3899991&download=yes> accessed 19 October 2021

Surden H, 'Values Embedded in Legal Artificial Intelligence' (2017) Univ. Colo. Law Legal Studies 5

Vaele M and Zuiderveen Borgesius F, 'Demystifying the Draft EU Artificial Intelligence, Analysing the good, the bad, and the unclear elements of the proposed approach', (2021) 22 Computer L. Rev. Int. 97

Van Brakel R. E, 'Pre-emptive Big Data Surveillance and its (Dis)empowering Consequences: The Case of Predictive Policing' in van der Sloot B, Broeders D and Schrijvers E (eds), *Exploring the Boundaries of Big Data* (Amsterdam University Press 2016)

Van den Branden A, *Les robots à l'assaut de la justice* (1st ed., Bruylant 2019)

Vogl R, 'Legal Tech in the USA' in Hartung M, Bues M-M and Halbleib G (eds), *Legal Tech. How Technology is Changing the Legal World. A Practioner's Guide* (Beck-Hart-Nomos 2018)

Wallach H, 'Big Data, Machine Learning, and the Social Sciences: Fairness, Accountability, and Transparency' (*Hanna Wallach*, 19 December 2014) <https://hannawallach.medium.com/big-data-machine-learning-and-the-social-sciences-927a8e20460d> and <https://www.microsoft.com/en-us/research/publication/big-data-machine-learning-and-the-social-sciences-fairness-accountability-and-transparency/> accessed 15 July 2021

Winston A, 'New Orleans ends its Palantir predictive policing program' (*TheVerge*, March 2018) <https://www.theverge.com/2018/3/15/17126174/new-orleans-palantir-predictive-policing-program-end> accessed 14 July 2021

Wolters Kluwer, 'Legal Insights' <https://www.wolterskluwer.com/fr-be/solutions/legal-insights/> accessed 14 July 2021

Yang Y, Yang Y and Fei Ju S, 'China seeks glimpse of citizens' future with crime predicting AI' (*Financial Times*, 23 July 2017) <https://www.ft.com/content/5ec7093c-6e06-11e7-b9c7-15af748b60d0> accessed 14 July 2021

Zouinar M, 'Evolutions de l'Intelligence Artificielle : quels enjeux pour l'activité humaine et la relation Humain-Machine au travail ?' (*Open Edition Journals*, 15 April 2020) <https://journals.openedition.org/activites/4941#quotation> accessed 21 June 2021

AUTOMATED JUSTICE AND ITS LIMITS: IRREPLACEABLE HUMAN(E) DIMENSIONS OF CRIMINAL JUSTICE

By Nina Peršak[*]

Abstract

Artificial intelligence (AI) is increasingly transforming our lives and everyday decision-making processes. Its rapid increase raises several ethical, legal and practical questions – particularly regarding its use in criminal justice. While AI is likely to play a key role in economic development and societal well-being, including assisting humans in the domain of justice, concerns relating to fundamental rights, equality, criminal law standards, procedural safeguards and potentially diminishing human engagement present challenges to the idea of (criminal) justice as we know it. Without diminishing the value of artificial intelligence in assisting the human decision-making in the judicial context, the aim of this article is to address two features or dimensions of such criminal justice that automated decision-making – if it were ever to be fully implemented – would upend, namely, the affective dimension and the human (interactive) dimension, which encompass essential elements, requirements and values of the contemporary (and traditional) criminal justice systems in many jurisdictions across the world.

1 Introduction

Artificial intelligence (AI) is progressively transforming our lives and everyday decision-making processes. Advances in machine learning techniques, which have become possible through increased computer processing power and the availability of large training datasets, have spurred 'AI spring'. It has been pointed out that we are already living in 'algorithmic society', ie society organised around automated social and economic decision-making by algorithms and AI agents, which collect vast amounts of data about individuals and enable new forms of surveillance, control and manipulation.[1] With information technology everywhere and AI's progress so inspiring, many legal professional harbour high expectations and, at the same time, concerns as to whether AI applications help promote a good society or lead to harmful consequences.[2] Are they a road to eunomia and eudaimonia or rather to dystopia?

The rapid increase of AI raises several ethical, legal and practical questions – particularly regarding its use in criminal justice. On the one hand, AI systems and techniques create exceptional opportunities for investigation and prosecution of criminal offences as well

[*] Scientific Director and Senior Research Fellow, Institute for Criminal-Law Ethics and Criminology, Ljubljana; law professorship habilitation, University of Maribor; Member of the European Commission's Expert Group on EU Criminal Policy; Independent Ethics Adviser; Co-Editor-in-Chief of the RIDP. All views expressed are the author's own. For correspondence: <nina.persak@criminstitute.org>.

[1] Jack M. Balkin, 'Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation' (2018) 51 UC Davis Law Rev. 1149, 1153.

[2] Bart Verheij, 'Artificial Intelligence as Law: Presidential Address to the Seventeenth International Conference on Artificial Intelligence and Law' (2020) 28 Artif Intell Law 181, 181.

as for the functioning and efficiency of the criminal justice system, for example, through digitalisation of criminal justice and quicker access to data relevant for decision-making.[3] AI has also been recognised as being of strategic importance, playing a key role in economic development and societal well-being, and praised for its potential to assist humans in a variety of domains, including justice. On the other hand, concerns relating to fundamental rights, privacy and data protection, biases in algorithmic decision-making and consequent discrimination, due process, equality of arms, access to justice as well as fairness, transparency[4] and accountability – essentially the rule of law concerns[5] – abound. Some of the frequent questions raised by jurists relate to the legal profession and legal decision-making as such: Will AI profoundly disrupt the legal profession or help it? Will AI show bias? Will 'robots' or 'algorithms' render judges largely redundant? Would AI judges be less or more biased than their human counterparts? Can AI even 'reason' in any meaningful sense?

In view of the stakes involved, the need for regulation of AI and its alignment with human rights, democracy and the rule of law standards has been recognised by EU and a number of international organisations. Council of Europe has in its 2018 study assessed

[3] In 2017, AI competed with more than 100 commercial lawyers in a Case Crunch Lawyer Challenge in London and won: in more than 775 predictions based on facts provided, the AI programme Case Crunch Alpha, the brainchild of four Cambridge law students, correctly predicted the outcome in 86.6% of cases, while top London lawyers in only 66.3%. In the area of human rights law, some have, reported that their AI models predicted the European Court of Human Rights' decisions (on Art. 3, 6 and 8 of ECHR) with a strong accuracy of 79% on average (Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preotiuc-Pietro and Vasileios Lampos, 'Predicting Judicial Decisions of the European Court of Human Rights: A Natural Language Processing Perspective' (2016) 2:e93 PeerJ Comput. Sci. 1. Another study revealed a 75% average accuracy on nine articles of the ECHR (Masha Medvedeva, Michel Vols and Martijn Wieling, 'Using Machine Learning to Predict Decisions of the European Court of Human Rights' (2020) 28 Artif Intell Law 237). Some legal professionals also see the potential of Legal AI, particularly Natural Language Processing, to free them from a maze of paperwork (Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu and Maosong Sun, 'How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence' in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics 2020) 5218.

[4] One of the major challenges and knowledge deficits involves understanding what AI exactly is or does and how it does it. The fact that information on the functioning of some of the existing predictive systems, such as COMPAS recidivism algorithm, has not been made public and can be protected with IP laws against it, has been strongly criticised. The problem is, however, further compounded by the fact that once an algorithm is learning, even skilled computer scientists no longer know to any degree of certainty what it does exactly and how it does it, ie what its rules and parameters are. 'At which point we can't be certain of how it will interact with other algorithms, the physical world, or us' (Matthew Griffin, 'Our Algorithmic Society, the Deadly Consequences of Unpredictable Code' (*Intelligence and the Senses*, 10 April 2020) <https://www.311institute.com/our-algorithmic-society-the-deadly-consequences-of-unpredictable-code/>).

[5] Stanley Greenstein, 'Preserving the Rule of Law in the Era of Artificial Intelligence (AI)' (2021) Artif Intell Law 1 (DOI 10.1007/s10506-021-09294-4); Markku Suksi, 'Administrative Due Process When Using Automated Decision-Making in Public Administration: Some Notes from a Finnish Perspective' (2021) 29 Artif Intell Law 87.

the impact of algorithms on various human rights, and highlighted the need for transparency, accountability, ethics and improved risk assessment among regulatory implications of the use of automated data processing techniques and algorithms.[6] In late 2018, the Council of Europe's European Commission for the Efficiency of Justice (CEPEJ) adopted an Ethical charter containing five principles on the use of AI in judicial systems and their environment, namely: the principle of respect for fundamental rights, the principle of non-discrimination, the principle of quality and security, the principle of transparency, impartiality and fairness, and the principle 'under user control'.[7] In April 2020, the Committee of Ministers adopted a Recommendation to the Council of Europe member States on the human rights impacts of algorithmic systems, which included, among others, a recommendation to engage in regular, inclusive and transparent consultation with all relevant stakeholders 'paying particular attention to the needs and voices of vulnerable groups, with a view to ensuring that human rights impacts stemming from the design, development and ongoing deployment of algorithmic systems are comprehensively monitored, debated and addressed'.[8] The second report by INTERPOL and UNICRI's Centre for AI and Robotics focused on responsible AI innovation and the use of AI for law enforcement purposes, flagging lawfulness, social acceptance, trustworthiness, responsibility and ethics as key concepts throughout the report.[9] Setting out the European Union's approach to artificial intelligence and building upon its Communication on AI in Europe (European strategy for AI) presented in April 2018,[10] the European Commission published a White paper on AI in February 2020, which presents policy options on how to achieve the two objectives of promoting the uptake of AI and of addressing the risks associated with certain uses of this new technology.[11] In April 2021, European Commission tabled a proposal for a Regulation laying down harmonised rules on artificial intelligence,[12] which is the world's first legal framework on AI. The proposal defines 'artificial intelligence system' (AI system) as 'software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of

[6] Council of Europe, *Algorithms and Human Rights: Study on the Human Rights Dimensions of Automated Data Processing Techniques and Possible Regulatory Implications*, prepared by the Committee of experts on Internet intermediaries (MSI-NET), DGI(2017)12 (Council of Europe 2018). For the past and on-going work of the Council of Europe bodies or committees on artificial intelligence more generally, see: <https://www.coe.int/en/web/artificial-intelligence>.

[7] CEPEJ, 'European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment' (3-4 December 2018).

[8] Committee of Ministers, 'Recommendation CM/Rec(2020)1 of the Committee of Ministers to Member States on the Human Rights Impacts of Algorithmic Systems' (8 April 2020), point 5.

[9] INTERPOL-UNICRI, 'Towards Responsible AI Innovation: Second INTERPOL-UNICRI Report on Artificial Intelligence for Law Enforcement' (19 May 2020).

[10] European Commission, 'Communication from the Commission to the European Parliament, the European Council, the Council, The European Economic and Social Committee and the Committee of Regions: Artificial Intelligence for Europe', COM(2018) 237 final, 25.4.2018.

[11] European Commission, 'White Paper on Artificial Intelligence – A European Approach to Excellence and Trust', COM(2020) 65 final, 19.2.2020.

[12] European Commission, 'Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts', COM(2021) 206 final, 21.4.2021.

human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with' (Art. 3(1)).[13] It sets harmonised rules for the development, placement on the market and use of AI systems in the Union, following a risk-based approach, which differentiates between uses of AI that create (i) an unacceptable risk, (ii) a high risk, and (iii) low or minimal risk. It lays down a methodology to define 'high-risk' AI systems that pose significant risks to the health and safety or fundamental rights of natural persons. Annex III lists 'high-risk AI systems referred to in Article 6(2)' – which are stand-along high-risk AI systems with mainly fundamental rights implications – and includes among them 'AI systems intended to assist a judicial authority in researching and interpreting facts and the law and in applying the law to a concrete set of facts'.[14] Such systems shall be subject to a strict regulatory framework where 'predictable, proportionate and clear' obligations will be imposed on providers and users of these systems.

One way in which the rules and principles mentioned above are implemented 'on the ground' can be seen in the way EU now evaluates new research proposals eligible for EU funding. Awareness of ethical issues of AI and its use, and of the links between the scientific excellence of the proposed AI system and ethical repercussions thereof, has become a priority high on the list of issues to be examined during EU project proposal evaluations in the new EU's key funding programme for research and innovation – Horizon Europe. The ethics review of such EU project proposals, for example, evaluates all ethical aspects related to the development, deployment and use of AI system or techniques. Ethically sound AI thus supports human agency and oversight, guarantees privacy and data protection, ensures transparency and accountability, avoids any possible risk of harm to the individual, society and environment, and is designed in a way to avoid bias, discrimination and stigmatisation.[15]

---

[13] This is a simpler definition compared to the one adopted by the High-Level Expert Group on AI, established by the European Commission (which expanded the initial definition of Artificial Intelligence, as proposed within the European Commission's Communication on AI), according to which AI systems are: 'software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions.' High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI* (8 April 2019) <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

[14] European Commission (n 12) point 8. The proposed Regulation has started the regular legislative procedure before the European Parliament and the Council of the EU and the incumbent Slovenian Presidency of the Council of the EU has included the ethical use and development of AI among its priorities.

[15] European Commission, 'How to Complete Your Ethics Self-Assessment', Version 2.0 (13 July 2021) <https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/common/guidance/how-to-complete-your-ethics-self-assessment_en.pdf> 39-41. The similarity with ethical principles applicable to AI systems, proposed by the Council of Europe Ad Hoc Committee on Artificial Intelligence (CAHAI) in

Discrimination is clearly a major legal concern that transpires also through most of the legal literature on AI. Human biases are coded into the datasets algorithms use and therefore render skewed, discriminatory outputs. In this way, we are replicating our inequalities in the digital world. However, some see automated decision-making as a positive development precisely because of human biases. AI as a judge would not have emotional biases that humans possess, rendering the process fairer because it would not be steered by emotion. However, this may be too simplistic a view or at least one that requires some nuancing, as it will be argued. What further remains understudied is how, or instances where, human input and engagement, fallible as it might be on occasion, is part and parcel of the justice delivered and one which therefore cannot be easily replaced by AI or automated justice.

Without diminishing the value of AI in assisting the human decision-making in the judicial context, the aim of this article is to address precisely these two features of justice, namely: (a) the affective dimension of justice, where emotion is not only an asset or requirement for an effective judicial deliberation but also structurally reflected as a value in of humane penal policy and criminal justice, and (b) the human (interactive) dimension of criminal justice. These two dimensions contain essential elements, requirements and values of the modern (and traditional) criminal justice systems in many jurisdictions across the world – they are, however, also aspects that automated decision-making would upend – if it were ever to be fully implemented.

2 Judicial Decision-Making: What's Emotion Got to Do with It?

2.1 Eliminating emotion: a desirable path?

Some see the development towards an automated justice, delivered by AI, as a positive development, since, on their view, the latter would not be guided by emotion.[16] This is not the only argument offered *for* the automated justice, but an important and often repeated one, which will be addressed here. The suggestion here is that unemotional AI may be the solution, leading towards increased fairness. The underlying reasoning is therefore based on the premise that emotion in judicial decision-making is 'bad', ie that emotion equals bias, unreasonableness and flawed reasoning. While this assumption is misguided and needs much further nuancing, it reflects a long-held view, particularly persistent in certain legal doctrinal circles.[17]

'AI Ethics Guidelines: European and Global Perspectives', CAHAI(2020)07-fin (15 June 2020), demonstrate that there is a widespread agreement on the core content of such principles.

[16] Jaap van den Herik, Full Professor of Law and Informatics (University of Leiden) in the podcast 'Kan een computer een rechter vervangen?' (Can a computer replace a judge?), De Universiteit van Vlaanderen Podcast, no. 284 (15 November 2020). Similarly, 'corporations like Facebook and Google have sold and defended their algorithms on the promise of objectivity, an ability to weigh a set of conditions with mathematical detachment and absence of fuzzy emotion' (Griffin (n 4)).

[17] It has been observed that legal studies or legal doctrine, particularly on the European Continent (less so in the USA and UK, where educational modules on 'law and emotion' are more common), seem to predominantly ignore the topic or continue to display ambivalence towards emotion as an element of

Law has traditionally been regarded as the preserve of reason, and emotion as the enemy of reason. On the conventional view, the law is seen as rational and unemotional.[18] Emotions are to be avoided in legislating and adjudicating, as they bring in passions and distort rational decision-making. Emotions have been described as impulsive, unpredictable, 'dysfunctional and irrational',[19] 'antithetical to reasons, disorienting and distorting practical thought'[20] and leading to logical fallacies in argumentation.[21] Emotional appeals – today so often witnessed in the courtroom (by parties and their attorneys) and even officially invited, eg through the instrument of Victim Impact Statements – have been labelled as 'illegitimate substitutes for proper argument'.[22] However, this taken-for-granted supposition does not reflect the reality and knowledge generated by social sciences, particularly developments in cognitive psychology and neuroscience. Emotion theory informs us that emotions can be considered 'rational much of the time'[23] and that they are, in fact, crucial in the correct processing of information.

Emotions already penetrate into the law-making as well as laws themselves.[24] The law as such often expresses emotional attachments and affective elements of culture.[25] Crim-

law-making and adjudication. See eg Susan A. Bandes and Jeremy A. Blumenthal, 'Emotion and the Law' (2012) 8 The Annual Review of Law and Social Science 161; Neil Feigenson and Jaihyun Park, 'Emotions and Attributions of Legal Responsibility and Blame: A Research Review' (2006) 30 Law and Human Behaviour 143. Such a predisposition, however, can be more easily found among 'legal formalists' who believe that judges merely apply, and should merely apply, rules to the facts of a case in a rather mechanistic way, using the methods of deductive logic (syllogism) to obtain the correct result, while 'legal realists' tend to recognise judges as human beings with moods, political preferences and emotions who are influenced not only by legal rules but by law-extrinsic factors as well.

[18] Jack M. Barbalet, 'Moral Indignation, Class Inequality and Justice: An Exploration and Revision of Ranulf' (2002) 6 Theoretical Criminology 279; Daniel Z. Epstein, 'Rationality, Legitimacy, & the Law' (2014) 7 Washington University Jurisprudence Review 1; Terry Maroney, 'A Field Evolves: Introduction to the Special Section on Law and Emotion' (2016) 8 Emotion Review 3.

[19] Klaus R. Scherer, 'On the Rationality of Emotions: Or, When are Emotions Rational?' (2011) 50 Social Science Information 330, 331.

[20] Patricia Greenspan, 'Practical Reasoning and Emotion' in: Mele, A.R., Rawling, P. (eds), *The Oxford Handbook of Rationality* (Oxford University Press, 2014) 206.

[21] Alan Brinton, 'Pathos and the 'Appeal to Emotion': An Aristotelian Analysis' (1988) 3 History of Philosophy Quarterly 207.

[22] Raphaël Micheli, 'Emotions as Objects of Argumentative Constructions' (2010) 24 Argumentation 1, 16.

[23] Scherer (n 19) 331. According to Scherer and his Component Process Model (CPM), the correctness or appropriateness of emotions depends on the accuracy of the underlying appraisal, the appropriateness of response pattern and the efficacy of emotion regulation. Within this framework, emotions can be considered 'rational' when they fulfil at least one of the three criteria of rationality, namely, functionality, inference, and reasonableness. Emotions are thus rational if, in a particular situation, (a) they are adaptive (functional, purposeful), (b) they are based on well-grounded, accurate inference from available information, and (c) they are considered as reasonable (understandable, sensible) reactions by others.

[24] Nina Peršak, 'Beyond Public Punitiveness: The Role of Emotions in Criminal Law Policy' (2019) 57 International Journal of Law, Crime and Justice 47.

[25] Roger Cotterrell, 'Theory and Values in Socio-Legal Studies' (2017) 44 Journal of Law and Society S19.

inal law, in particular, incorporates emotions in various ways. For example, it criminalises the causing of emotional distress to another person.[26] It also takes emotion into account when crimes are committed 'in the heat of passion' through the 'crimes-of-passion defence', challenging the *mens rea* element, specifically premeditation. Emotions can, further, act as exculpating factors and as aggravating or mitigating circumstances.[27] Judges are therefore required to be able to recognise emotion (as part of the element of an offence), distinguish between different emotions, assess them, and – it will be argued – be able to empathise with various subjects or their testimonials in the proceedings before them to be able to render a just verdict.[28] In order to be capable of doing all this, however, a certain Emotional Intelligence (EI)[29] is indispensable. Needless to say, AI has no EI (yet) and is therefore not capable of emotional rationality required for complex, socially embedded decision-making such as adjudication.

2.2 Intuition and empathy in penal policy and criminal justice

According to Norbert Elias, any investigation trying to understand human beings that disregards the structure, direction and form of human affects, drives and passions can only be of limited value.[30] Being able to understand human beings – which few would argue is not a desirable trait in a good judge – therefore presupposes the ability to understand human affect. Let us first consider intuition, for which emotions are crucial, as they provide it with affective cues and pieces of information.[31] Analysing the European Court of Human Rights' anti-discrimination jurisprudence, Wachter et al. assert that the

---

[26] For example, Section 222, para. 1, of the Criminal Code of Hungary (Act C of 2012 on the Criminal Code) stipulates that '[a]ny person who engages in conduct intended to intimidate another person, to disturb the privacy of or to upset, or cause emotional distress to another person arbitrarily […] is guilty of a misdemeanour punishable by imprisonment not exceeding one year, insofar as the act did not result in a more serious criminal offence'. Art. 7, para. 3 of the Slovenian Public Order and Peace Act; criminalises indecent exposure, intrusive offering of sexual services or having sexual intercourse in a public place, insofar as this conduct 'disturbs people, causes alarm or indignation'. Anti-social behaviour in England and Wales is defined as '(a) conduct that has caused, or is likely to cause, harassment, alarm or distress to any person, […]' (Part 1, Section 2 of Anti-social Behaviour, Crime and Policing Act 2014).

[27] According to Art. 20, point 6, of the Spanish Criminal Code, a person whose actions were 'driven by insurmountable fear' is 'not criminally accountable'. Emotions also sometimes aggravate a killing into a murder, leading to the death penalty, whereas in other times mitigating a murder into manslaughter (Norman J. Finkel and W. Gerrod Parrott, *Emotions and Culpability: How the Law is at Odds with Psychology, Jurors, and Itself* (American Psychological Association 2006)).

[28] While AI may be trained to 'read' certain emotions, eg via the sentiment analysis software, it cannot (yet) experience and act with appropriate emotions.

[29] Emotional Intelligence (EI) has been defined as the ability to perceive, use, understand, evaluate and control emotions. EI is typically associated with empathy, intelligence and emotions because it involves connecting one's personal experiences with those of others to enhance thought and understanding of interpersonal dynamics. See: John D. Mayer, Richard D. Roberts and Sigal G. Barsade, 'Human Abilities: Emotional Intelligence' (2008) 59 Annual Review of Psychology 507.

[30] Norbert Elias, *The Civilizing Process* (Blackwell Publishers 2000, orig. 1939), 408.

[31] See eg Mark Fenton-O'Creevy, Emma Soane, Nigel Nicholson and Paul Willman, 'Thinking, Feeling and Deciding: The Influence of Emotions on the Decision Making and Performance of Traders' (2011) 32 Journal of Organizational Behavior 1044.

latter is mostly intuitive, ie based on the intuitive understanding of fairness and discrimination.[32] Comparing it to AI, they conclude that the progressive use of algorithms 'disrupts traditional legal remedies and procedures for detection, investigation, prevention, and correction of discrimination which have predominantly relied upon intuition'[33] – the intuitive experiences of discrimination being essential to bringing claims under non-discrimination law.[34] As emotions are needed for intuition, and intuition is required to detect much discrimination, it follows that emotions are needed to detect such discrimination.[35]

Furthermore, emotions are necessary in order to understand other people's emotions, eg in 'crimes of passion', offensive behaviour, hate speech, Victim Impact Statements, repentance and apology, and so forth. To understand the emotional impact a crime had on its victim, the judge (and/or jury) must be able to feel similar emotions, to put herself or himself in the victim's shoes or in their position, in short, to empathise. Empathy is a complex emotion, a vicarious sensation of the other's emotional state or life condition.[36] Empathy for another's life condition, which develops by late childhood, enables one to represent others' distress in more sophisticated ways, eg to acknowledge that certain forms of distress can have long-term consequences for the victim.[37] The ability to grasp such consequences, eg when listening to the victim's testimonial or the offender's plea, is clearly crucial to understanding the relevant facts of the case (such as harm, context,

[32] Sandra Wachter, Brent Mittelstadt and Chris Russell, 'Why Fairness Cannot be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI' (2020) SSRN Electronic Journal 1, 69. DOI: 10.2139/ssrn.3547922

[33] ibid 2.

[34] Such intuitive experiences are diminished in the case of AI discrimination. While humans have various signalling mechanisms when they discriminate (eg negative attitudes based on prejudice and stereotypes), which makes it easier to spot discrimination, AI systems do not. Intuitive or *prima facie* experiences of discrimination, which are essential to bringing claims under EU non-discrimination law, are thus diminished. 'By definition claimants must experience or anticipate inequality. Compared to traditional forms of discrimination, automated discrimination is more abstract and unintuitive, subtle, and intangible. These characteristics make it difficult to detect and prove as victims may never realise they have been disadvantaged. Seeing colleagues getting hired or promoted, or comparing prices in supermarkets, help us to understand whether we are treated fairly. In an algorithmic world this comparative element is increasingly eroded; it will be much harder, for example, for consumers to assess whether they have been offered the best price possible or to know that certain advertisements have not been shown to them.' Ibid., at 10. Similarly, '[b]ias or prejudice related, for example, to racial or ethnic background, may not be recognised as such by the police when integrated into an automated computer program that is deemed independent and neutral […]. As a result, bias may become standardised and may then be less likely to be identified and questioned as such' (Council of Europe (n 6) 11).

[35] This is not to say, however, that detection of *all* types of discrimination requires intuition: AI tools such as Conversation AI or Perspective API are, for example, rather good (and clearly faster or more efficient than their human counterparts) at automatically spotting and moderating the crudest forms of hate speech online.

[36] Nancy Eisenberg and Paul H Mussen, *The Roots of Prosocial Behavior in Children* (CUP 1989); Kevin Durkin, *Developmental Social Psychology: From Infancy to Old Age* (Blackwell Publishers 1995).

[37] Martin L. Hoffman, 'The Contribution of Empathy to Justice and Moral Judgment' in Eisenberg N and Strayer J (eds), *Empathy and its Development* (CUP 1987) 47; Durkin (n 36) 448-449.

mitigating circumstances) and, ultimately, to the delivery of justice. The ability to understand the motivations, intentions, and goals of others is 'a prerequisite to fair and accurate legal decision-making' and essential to structuring fair and effective legal institutions.[38]

Indeed, some emotions – like empathy – are actually a systemic orientation, part of the structures and penal policy in some countries. Emotions of compassion and solidarity[39] have historically played a role, or still do, in the general orientation of criminal policy in several EU countries, for example, in the Nordic countries. The latter were steered by the 'rational and humane criminal policy' developed in the 60s and 70s.[40] The Finnish Criminal Law Committee tasked with the Criminal Code reform in the 70s was described as standing for a 'humane, just and effective criminal justice system'.[41] The sense of 'shared responsibility and solidarity', espoused by the Nordic welfare state, has been seen as 'essential for less punitive penal policies', which led those countries (together with Slovenia and Switzerland) to having the lowest level of prisoners in Europe.[42] Furthermore, even in more retributive and crime control-oriented jurisdictions, restorative justice mechanisms require for their successful functioning a sufficient demonstration of emotion (eg repentance shown in a sincere apology), empathy and emotional understanding on the sides of both, the offender as well as the victim (and their community).[43] Developments in 'therapeutic jurisprudence'[44] have similarly challenged the traditional judicial role, requiring from judges more active and emotional engagement with the defendant to help their rehabilitation.[45]

---

[38] Susan A. Bandes, 'Taz and Empathy' (2015) 58 Howard Law Journal 397, 397.

[39] Solidarity has been described as a 'reflective emotion', ie emotion that is typically mediated by thought, belief and ideology. David Heyd, 'Solidarity: A Local, Partial and Reflective Emotion' (2015) 43 Diametros 55.

[40] Raimo Lahti, 'Towards a Rational and Humane Criminal Policy – Trends in Scandinavian Penal Thinking' (2000) 1 Journal of Scandinavian Studies in Criminology and Crime Prevention 141; Tapio Lappi-Seppäla, 'Penal Policies in the Nordic Countries 1960–2010' (2012) 13(sup1) Journal of Scandinavian Studies in Criminology and Crime Prevention 85; Raimo Lahti, *Towards an Efficient, Just and Humane Criminal Justice: Nordic Essays on Criminal Law, Criminology and Criminal Policy 1972-2020* (Suomalainen Lakimiesyhdistys 2021).

[41] Inkeri Antilla and Patrik Törnudd, 'Current Criminal Code Reforms: The Examples of Finland, Norway and Israel' in Lahti R and Nuotio K (eds), *Criminal Law Theory in Transition: Finnish and Comparative Perspectives* (Finnish Lawyers' Publishing Company 1992) 11, 13.

[42] Lappi-Seppäla (n 40) 99, 107.

[43] See eg Bas van Stokkom, 'Moral Emotions in Restorative Justice Conferences: Managing Shame, Designing Empathy' (2002) 6 Theoretical Criminology 339; Meredith Rossner, 'Restorative Justice in the 21st Century: Making Emotions Mainstream' in Alison Liebling, Shadd Maruna and Lesley McAra (eds), *The Oxford Handbook of Criminology* (Oxford University Press 2017) 967.

[44] Therapeutic jurisprudence 'concentrates on the law's impact on emotional life and psychological well-being' (David B. Wexler and Bruce J. Winick, *Law in Therapeutic Key: Developments in Therapeutic Jurisprudence* (Carolina Academic Press 1996) xvii). Its tenets have, however, been more easily accepted in Anglo-American legal systems than in Europe.

[45] 'Therapeutic judging is regarded as an important element in the rehabilitation of offenders for several reasons. The judge-defendant relationship can, for example, increase motivation, manage risk, enhance

Emotions are thus engrained in the very fabric of our modern penal institutions, structures, laws, principles and rules that govern our criminal justice systems. Replacing human agents with AI or autonomous agents with automated decision-making would disrupt or transform many existing structures and procedures that rely on emotions or emotionally intelligent processing of information.

3 The Human Dimension of Criminal Justice

3.1 Trial by peers

Another aspect of modern criminal justice systems that would be difficult to see replaced by automatons is the notion of being tried by one's 'peers', fellow citizens, members of the community to which one belongs. The rationale behind this legal institution is ultimately that of legitimacy – legitimacy of the process and verdict, that is consequently more likely to be respected and complied with. Although the legal institution of a jury – a group of laypersons participating in deciding cases brought to trial – continues to attract controversy, the proponents argue that it provides an important civic experience, guarantees judicial integrity (as it is said to be more difficult to bribe 12 people than one), provides (as a group) wisdom and strength beyond that of its individual members, that its common sense and experience make up for what it lacks in training, that its judicial inexperience is an asset because it secures a fresh perception of each trial (escaping the stereotypes that may infect the judicial eye), and that its flexibility ensures that the rigidity of the general rule can be shaped to justice in a particular case, governed rather by the spirit of the law than by its letter.[46] Another argument put forward in favour of the desirability of trial by one's peers rests on the notion that '(a) it is good to be tried by a group of individuals who are representative of one's community; and (b) that 'representativeness' makes for impartial, objective, just and fair jury verdicts'.[47] Even though the portraits of the ideal juror differ and may even conflict – eg juror as peer and neighbour encapsulating community consciousness vs. juror as a tabula rasa guaranteeing impartiality[48] – the idea of jury as embodiment of democratic justice persists.

Even in non-jury criminal justice systems, eg in many Continental European countries, the participation of 'lay judges' pursues the same rationale. The judgment is often delivered 'in the name of the people'. This clearly cannot be replicated by AI, at least when AI is delivering decisions that affect humans – decisions affecting eg rogue automated

offender compliance and build on the offender's strengths. While some of these tasks are taken upon by correctional officers, the theory of therapeutic jurisprudence – more influential in Anglo-American legal systems – maintains that 'it is the element of judicial authority that is key to offenders' motivation to change their behaviour'. Arie Freiberg, 'Post-Adversarial and Post-Inquisitorial Justice: Transcending Traditional Penological Paradigms' (2011) 8 *European Journal of Criminology* 81, 96.

[46] Hans Zeisel, 'Jury' (*Encyclopaedia Britannica*) <https://www.britannica.com/topic/jury>

[47] Andreas Kapardis, *Psychology and Law: A Critical Introduction*, 2nd ed. (CUP 2003), 129.

[48] Jeffrey Abramson, *We, the Jury: The Jury System and the Ideal of Democracy* (Harvard University Press 2001), 17-18.

agents (if given legal subjectivity)[49] would probably fulfil the 'peers' requirement. Should AI be allowed to render verdicts guilty/not guilty to humans without substantial human input in terms of evaluation of the information presented, the nature and rationale of some of the most essential elements of criminal justice system, as we know it, would be therefore undermined. As such legal institutions reflect our understanding of justice, the repercussions would be far-reaching. Many agree that the use of technology challenges our understanding of what the objectives of justice are: 'to process as many cases as possible with a view to punishing; or to rehabilitate and restore social ties through an interactive process that aims to restore "social peace". The latter requires humanity that technology cannot offer.'[50] While AI cannot replace such human element, it can, of course, be of assistance, for example, aiding juries to do their jobs or even assisting in the selection of jury members.[51]

3.2 Human interaction as a prerequisite for trust, fair trial and effective defence

Giving substance to many due process standards, physical presence, personal encounters and human interaction have been described as being central to a fair trial. 'Individuals, especially those who are vulnerable, disadvantaged or deprived, feel more "heard" in face-to-face conversations during in-person encounters.'[52] Feeling 'heard' is, however, an important part of procedural justice, which affects trust and consequent compliance with eg the court order, judgment, law, as well as the perceived legitimacy of the relevant authority.[53] Furthermore, even with simpler technology, such as videoconferences, the effectiveness of the defence is lessened. Research and practice reveal that 'defendants are

---

[49] For arguments that it is possible to hold autonomous agents themselves, and not only their makers, users or owners, responsible for the acts of these agents, and that despite potential dangers connected with endowing AI with some type of legal subjectivity, such a course is inescapable, see Jaap Hage, 'Theoretical Foundations for the Responsibility of Autonomous Agents' (2017) 25 Artif Intell Law 255 and Sylwia Wojtczak, 'Endowing Artificial Intelligence with Legal Subjectivity' (2021) AI & Society 1 (DOI 10.1007/s00146-021-01147-7), respectively. For an argument against AI legal personhood, see eg Joanna J. Bryson, Mihailis E. Diamantis and Thomas D. Grant, 'Of, For, and By the People: The Legal Lacuna of Synthetic Persons' (2017) 25 Artif Intell Law 273.

[50] Sergio Carrera, Valsamis Mitsilegas and Marco Stefan, *Criminal Justice, Fundamental Rights and the Rule of Law in the Digital Age* (CEPS 2021).

[51] Voltaire, for example, is an application offered to legal professionals 'who need to instantly search for information on potential jurors and other individuals related to their cases' (<https://voltairapp.com/about/>). It starts searching for basic information such as age, address, education, jobs, then business and legal history of the person, moving on to examining social media posts (Twitter, Facebook and similar) for any indication of a particular attitude or sentiment toward certain issues that may be relevant to the trial (eg a juror's comment that they greatly dislike jury duty or feel very antagonistic about a certain trial-relevant issue). All of this is performed within seconds and the results then presented on a lawyer's iPad or laptop, allowing them to access and assess the data as they stand before juries in the courthouse (artificiallawyer, 'Voltaire Uses AI and Big Data to Help Pick Your Jury' (*Artificial Lawyer*, 26 April 2017) <https://www.artificiallawyer.com/2017/04/26/voltaire-uses-ai-and-big-data-to-help-pick-your-jury/> accessed 28 September 2021).

[52] Carrera et al. (n 50) 10.

[53] Tom R. Tyler, *Why People Obey the Law* (Yale University Press 2006); Tom R. Tyler, 'Legitimacy and Criminal Justice: The Benefits of Self-Regulation' (2009) 7 Ohio State Journal of Criminal Law 307.

less likely to take up legal advice during videoconference hearings', that videoconferencing 'does not always provide the opportunity to entertain the kind of confidential exchanges between lawyers and their clients that are required to deliver an effective defence' and that 'defendants who appear on video from police stations are more likely to get prison sentences.'[54] For example, a recent overview of the functioning of the justice systems, the 2021 EU Justice Scoreboard, shows that only in seven EU Member States the defendants can communicate in *all* criminal cases confidentially with their lawyers during remote hearings.[55]

Having just gone through an involuntary immersive experience with Covid-19 pandemic, lockdowns, remote work and schooling, it is not difficult to comprehend how technology can be helpful to connect with each other and yet, how drastically it limits our human experience. Technology also alienates, leads to 'technology fatigue' and all sorts of psychological and mental hardships. It is thus easy to imagine how a defendant who may go to prison for life must eagerly await to have their day in court, tell their story and try to connect with the judge and/or jury with their personal plea, and how disheartening it must feel if they were to be invited to testify to eg an automated judicial chatbot clerk instead. An image of Kafka's Josef K, a man arrested and prosecuted by a remote, inaccessible authority, with the nature of his crime never revealed to him, permanently waiting at the door, to be let in, to speak, to get some answers, show despair, frustration or appeal for mercy, immediately comes to mind, mixed with the experience of exasperation all of us have probably felt when calling some helpline or customer service to speak to a person who could help us with an impending problem, only to be waiting for ages on the phone, pressing all sorts of keys (meant to increase efficiency) in order to get to a live, human being on the other side, before being hung up upon or rerouted to an answering machine or advised to fill out a website form. If one feels frustrated by the lack of human interaction in such everyday occurrences, it does not take much imagination to sense the sort of despair, feeling of anguish and hopelessness a defendant would face whose fate was being determined by a remote, non-human (possibly inhuman) entity.

4 Conclusion

In this paper, we shed light on two main aspects of criminal justice (and criminal law and policy, more generally) where humans could not be replaced with automated decision-making agents without significant adverse impact on the individual's rights and the very concept of 'justice', as we know it; namely, in the area of judicial deliberation and human interaction during the trial. While algorithms as such do not by themselves have adverse human rights impacts, their implementation and application to human interaction may.

[54] ibid. 10-11.

[55] European Commission, 'Communication from the Commission to the European Parliament, the Council, the European Central Bank, the European Economic and Social Committee and the Committee of the Regions: The 2021 EU Justice Scoreboard', COM(2021) 389, 8.7.2021 <https://ec.europa.eu/info/sites/default/files/eu_justice_scoreboard_2021.pdf>

It is the latter that invests them with social meaning and value that has implications for the individual's rights. It would thus be wrong to blame the algorithms themselves; rather, it is the specific norms and values embedded in algorithms and the decision-making processes around algorithms that need to be scrutinised and challenged in terms of their results and how these affect the defendant's procedural rights, victim's rights and, more generally, human rights.[56] Algorithms should best be seen as complementary tools that help people make better decisions[57] or 'flourish',[58] rather than agents fit to replace human decision-making.

In its 2018 study, the Council of Europe has mentioned a couple of contexts or areas of societal and human interaction where algorithmic decision-making systems may not be appropriate; namely, for promotion of societal development or resolution of complex new challenges for future generations. Relying heavily on algorithms in these contexts is, in their view, likely to do more harm than good.[59] In view of the importance of human judicial deliberation, including intuitive and affective reasoning, and of being heard and judged by one's peers that involves human interaction with authorities, it could be argued that certain, or perhaps most, criminal justice settings ought to be added to this list – particularly those settings that have direct, and possibly detrimental, implications on the various subjects involved in the criminal process and their rights.[60] Nevertheless, our 'primeval fear about technological change'[61] should not impede the uptake of AI systems and techniques where they can significantly and positively aid the human processes, including those in criminal justice.

[56] Council of Europe (n 6) 8.

[57] This is already being utilised in US courtrooms. In a 2017 high-profile 2017, Eric Loomis was sentenced to six years in prison owing partly to recommendations from AI algorithms. The algorithm used, which was built into the software COMPAS, analysed data about Loomis and indicated *inter alia* to the human judge that Loomis had a high risk of recidivism. This influenced the six-year sentence he received, although the sentencing judges were advised to take note of the algorithm's limitations. Criminal justice algorithms like the one in the Loomis case use personal data such as age, sex, and employment history to recommend sentencing, and this technology is reportedly relatively common in the US legal system. Logan Kugler, 'AI Judges and Juries' (2018) 61 Communications of the ACM 19.

[58] The concept of 'human flourishing', originating from virtue ethics, has been proposed as a guide to understanding the ethics in AI. See Bernd Carsten Stahl, *Artificial Intelligence for a Better Future: An Ecosystem Perspective on the Ethics of AI and Emerging Digital Technologies* (Springer 2021).

[59] Council of Europe (n 6) 44.

[60] A different sentiment, but one that is potentially leading to similar results, can be distilled from the AI-ethics principle of 'responsibility', which includes the requirement that 'a human being should be responsible for any decision affecting individual rights and freedoms, with defined accountability and legal liability for those decisions'. PACE (Parliamentary Assembly of the Council of Europe), Committee on Legal Affairs and Human Rights, 'Report on Justice by Algorithm (The Role of Artificial Intelligence in Policing and Criminal Justice Systems)', rapporteur: Mr Boriss Cilevičs (1 October 2020) 18.

[61] Michael R. McGuire, *Technology, Crime and Justice: The Question Concerning Technomia* (Routledge 2012), 221.

References

Abramson J, *We, the Jury: The Jury System and the Ideal of Democracy* (Harvard University Press 2001)

Aletras N, Tsarapatsanis D, Preotiuc-Pietro D and Lampos V, 'Predicting Judicial Decisions of the European Court of Human Rights: A Natural Language Processing Perspective' (2016) 2:e93 PeerJ Comput. Sci. 1

Antilla I and Törnudd P, 'Current Criminal Code Reforms: The Examples of Finland, Norway and Israel' in Lahti R and Nuotio K (eds), *Criminal Law Theory in Transition: Finnish and Comparative Perspectives* (Finning Lawyers' Publishing Company 1992) 11

Balkin JM, 'Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation' (2018) 51 UC Davis Law Review 1149

Bandes SA and Blumenthal JA, 'Emotion and the Law' (2012) 8 The Annual Review of Law and Social Science 161

—— 'Taz and Empathy' (2015) 58 Howard Law Journal 397

Barbalet JM, 'Moral Indignation, Class Inequality and Justice: An Exploration and Revision of Ranulf' (2002) 6 Theoretical Criminology 279

Brinton A, 'Pathos and the "Appeal to Emotion": An Aristotelian Analysis' (1988) 3 History of Philosophy Quarterly 207

Bryson JJ, Diamantis ME and Grant TD, 'Of, For, and By the People: The Legal Lacuna of Synthetic Persons' (2017) 25 Artif Intell Law 273

CAHAI, 'AI Ethics Guidelines: European and Global Perspectives', CAHAI(2020)07-fin (15 June 2020)

Carrera S, Mitsilegas V and Stefan M, *Criminal Justice, Fundamental Rights and the Rule of Law in the Digital Age* (CEPS 2021)

Cotterrell R, 'Theory and Values in Socio-Legal Studies' (2017) 44 Journal of Law and Society S19

CEPEJ, 'European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment' (3-4 December 2018)

Committee of Ministers, 'Recommendation CM/Rec(2020)1 of the Committee of Ministers to Member States on the Human Rights Impacts of Algorithmic Systems' (8 April 2020)

Council of Europe, *Algorithms and Human Rights: Study on the Human Rights Dimensions of Automated Data Processing Techniques and Possible Regulatory Implications*, DGI(2017)12 (Council of Europe 2018)

Durkin K, *Developmental Social Psychology: From Infancy to Old Age* (Blackwell Publishers 1995)

Eisenberg N and Mussen PH, *The Roots of Prosocial Behavior in Children* (CUP 1989)

Elias N, *The Civilizing Process* (Blackwell Publishers 2000, orig. 1939)

Epstein DZ, 'Rationality, Legitimacy, & the Law' (2014) 7 Washington University Jurisprudence Review 1

European Commission, 'Communication from the Commission to the European Parliament, the European Council, the Council, The European Economic and Social Committee and the Committee of Regions: Artificial Intelligence for Europe', COM(2018) 237 final, 25.4.2018

European Commission, 'White Paper on Artificial Intelligence – A European approach to Excellence and Trust', COM(2020) 65 final, 19.2.2020

—— 'Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts', COM(2021) 206 final, 21.4.2021 <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206>

—— 'Communication from the Commission to the European Parliament, the Council, the European Central Bank, the European Economic and Social Committee and the Committee of the Regions: The 2021 EU Justice Scoreboard', COM(2021) 389, 8.7.2021 <https://ec.europa.eu/info/sites/default/files/eu_justice_scoreboard_2021.pdf>

—— 'How to Complete Your Ethics Self-Assessment', version 2.0 (13 July 2021) <https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/common/guidance/how-to-complete-your-ethics-self-assessment_en.pdf>

Feigenson N and Park J, 'Emotions and Attributions of Legal Responsibility and Blame: A Research Review' (2006) 30 Law and Human Behaviour 143

Fenton-O'Creevy M, Soane E, Nicholson N and Willman P, 'Thinking, Feeling and Deciding: The Influence of Emotions on the Decision Making and Performance of Traders' (2011) 32 Journal of Organizational Behavior 1044

Finkel NJ and Parrott WG, *Emotions and Culpability: How the Law is at Odds with Psychology, Jurors, and Itself* (American Psychological Association 2006)

Freiberg A, 'Post-Adversarial and Post-Inquisitorial Justice: Transcending Traditional Penological Paradigms' (2011) 8 European Journal of Criminology 81

Greenspan P, 'Practical Reasoning and Emotion' in Mele AR and Rawling P (eds), *The Oxford Handbook of Rationality* (Oxford University Press 2004), 206

Greenstein S, 'Preserving the Rule of Law in the Era of Artificial Intelligence (AI)' (2021) Artificial Intelligence and Law 1. DOI 10.1007/s10506-021-09294-4

Griffin M, 'Our Algorithmic Society, the Deadly Consequences of Unpredictable Code' (*Intelligence and the Senses*, 10 April 2020) <https://www.311institute.com/our-algorithmic-society-the-deadly-consequences-of-unpredictable-code/>

Hage J, 'Theoretical Foundations for the Responsibility of Autonomous Agents' (2017) 25 Artificial Intelligence and Law 255

Heyd D, 'Solidarity: A Local, Partial and Reflective Emotion' (2015) 43 Diametros 55

High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI* (8 April 2019) <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

Hoffman ML, 'The Contribution of Empathy to Justice and Moral Judgment' in Eisenberg N and Strayer J (eds), *Empathy and its Development* (CUP 1987) 47

INTERPOL-UNICRI, 'Towards Responsible AI Innovation: Second INTERPOL-UNICRI Report on Artificial Intelligence for Law Enforcement' (19 May 2020)

Kapardis A, *Psychology and Law: A Critical Introduction*, 2nd ed. (CUP 2003)

Kugler L, 'AI Judges and Juries' (2018) 61 Communications of the ACM 19

Lahti R, 'Towards a Rational and Humane Criminal Policy – Trends in Scandinavian Penal Thinking' (2000) 1 Journal of Scandinavian Studies in Criminology and Crime Prevention 141

—— *Towards an Efficient, Just and Humane Criminal Justice: Nordic Essays on Criminal Law, Criminology and Criminal Policy 1972-2020* (Suomalainen Lakimiesyhdistys 2021)

Lappi-Seppäla T, 'Penal Policies in the Nordic Countries 1960–2010' (2012) 13(sup1) Journal of Scandinavian Studies in Criminology and Crime Prevention 85

Maroney T, 'A Field Evolves: Introduction to the Special Section on Law and Emotion' (2016) 8 Emotion Review 3

Mayer JD, Roberts RD and Barsade SG, 'Human Abilities: Emotional Intelligence' (2008) 59 Annual Review of Psychology 507

McGuire MR, *Technology, Crime and Justice: The Question Concerning Technomia* (Routledge 2012)

Medvedeva M, Vols M and Wieling M, 'Using Machine Learning to Predict Decisions of the European Court of Human Rights' (2020) 28 Artificial Intelligence and Law 237

Micheli R, 'Emotions as Objects of Argumentative Constructions' (2010) 24 Argumentation 1

PACE – Committee on Legal Affairs and Human Rights, 'Report on Justice by Algorithm (The Role of Artificial Intelligence in Policing and Criminal Justice Systems)' (1 October 2020).

Peršak N, 'Beyond Public Punitiveness: The Role of Emotions in Criminal Law Policy' (2019) 57 International Journal of Law, Crime and Justice 47

Rossner M, 'Restorative Justice in the 21st Century: Making Emotions Mainstream' in Liebling A, Maruna S and McAra L (eds), *The Oxford Handbook of Criminology* (Oxford University Press 2017) 967

Scherer KR, 'On the Rationality of Emotions: Or, When are Emotions Rational?' (2011) 50 Social Science Information 330

Stahl BC, *Artificial Intelligence for a Better Future: An Ecosystem Perspective on the Ethics of AI and Emerging Digital Technologies* (Springer 2021)

Suksi M, 'Administrative Due Process When Using Automated Decision-Making in Public Administration: Some Notes from a Finnish Perspective' (2021) 29 Artificial Intelligence and Law 87

Tyler TR, *Why People Obey the Law* (Yale University Press 2006)

—— 'Legitimacy and Criminal Justice: The Benefits of Self-Regulation' (2009) 7 Ohio State Journal of Criminal Law 307

Van Stokkom B, 'Moral Emotions in Restorative Justice Conferences: Managing Shame, Designing Empathy' (2002) 6 Theoretical Criminology 339

Verheij B, 'Artificial Intelligence as Law: Presidential Address to the Seventeenth International Conference on Artificial Intelligence and Law' (2020) 28 Artificial Intelligence and Law 181

Wachter S, Mittelstadt B and Russell C, 'Why Fairness Cannot be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI' (2020) SSRN Electronic Journal 1. DOI 10.2139/ssrn.3547922

Wexler DB and Winick BJ, *Law in Therapeutic Key: Developments in Therapeutic Jurisprudence* (Carolina Academic Press 1996)

Wojtczak S, 'Endowing Artificial Intelligence with Legal Subjectivity' (2021) AI & Society. DOI 10.1007/s00146-021-01147-7

Zeisel H, 'Jury' (*Encyclopaedia Britannica*) <https://www.britannica.com/topic/jury>

Zhong H, Xiao C, Tu C, Zhang T, Liu Z and Sun M, 'How Does NLP Benefit Legal System: a summary of Legal Artificial Intelligence' in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics 2020)

Subscriptions & Membership Applications

AIDP/IAPL Membership

Annual Contribution of € 110

Benefactor Member – A member wishing to provide extra financial support to the Association. This type of membership includes subscription to the RIDP as well as online access.

Collective Member – Universities, associations, institutes, etc. This type of membership includes subscription to the RIDP.

National group – AIDP Members have established in numerous countries a National Group, which carries out its own scientific activities. Each National Group, in addition to the fees for individual members, has to pay to the AIDP a membership fee which entitles the national group to participate in the activities of the Association. This type of membership includes subscription to the RIDP.

Annual Contribution of € 85

Individual Member – The AIDP membership includes subscription the RIDP as well as online access to the RIDP archives and the RIDP *libri* series.

Annual Contribution of € 45

Young Penalist – AIDP members under the age of 35 may join the Young Penalist Group, which carries out its own activities and elects representatives to the organs of AIDP. This type of membership includes full access to the electronic archive (incl. RIDP *libri*) but no paper version of the RIDP.

Student or retiree – AIDP membership for a reduced contribution. This type of membership includes full access to the electronic archive (incl. RIDP *libri*) but no paper version of the RIDP.

Reduced-fee countries – If you are residing in a country listed on the reduced country fee list, you will be entitled to membership including a subscription to the RIDP for a limited membership fee. The list can be consulted on the AIDP website under the section 'About Us' – guidelines to establish a national group. http://www.penal.org/en/guidelines-establishment-national-groups . This type of membership includes full access to the electronic archive (incl. RIDP *libri*) but no paper version of the RIDP.

Annual Contribution of € 40

AIDP Individual Membership without RIDP subscription - mere AIDP membership without RIDP subscription and no access to electronic archives.

Membership Application instructions

The membership application form can be downloaded at the AIDP website (http://www.penal.org/en/user/register) and returned by email or mail to the address below:

Email: secretariat@penal.org. Secretariat: AIDP, c/o The Siracusa International Institute, Via Logoteta 27, 96100 Siracusa, Italy

BNP PARIBAS Bordeaux C Rouge N° IBAN : FR76 3000 4003 2000 0104 3882 870

Payment instructions

By check: Join your check to your membership application form and mail it to: AIDP secretariat, c/o The Siracusa International Institute, Via Logoteta 27, 96100 Siracusa, Italy.

Bank transfer: The bank and account details are on the membership application form. Once the bank transfer is done, send your membership application form together with a copy of the bank transfer order by fax or email, or by mail to the secretariat of the Association. The identity of the sender does not appear on the bank statement and if you do not send a copy of the bank transfer separately, we will not be able to credit the transfer to your membership.

Payment by credit card: The cryptogram is the three-digit number on the reverse side of your credit card. It is necessary for payment. Do not forget to sign your application. Please return the form by fax or email, or by mail.

For further information please consult the AIDP website http://www.penal.org/.

Subscription to the RIDP

Single Issue – price indicated for each issue on MAKLU website.

Annual Subscription – For the price of € 85, an annual subscription to the RIDP can be obtained which includes the print and free online access to the RIDP back issues. This subscription does not include AIDP Membership.

For RIDP subscription, please follow the instructions on the MAKLU publisher's website: http://www.maklu-online.eu

Artificial intelligence (AI) is impacting our everyday lives in a myriad of ways. The use of algorithms, AI agents and big data techniques also creates unprecedented opportunities for the prevention, investigation, detection or prosecution of criminal offences and the efficiency of the criminal justice system. Equally, however, the rapid increase of AI and big data in criminal justice raises a plethora of criminological, ethical, legal and technological questions and concerns, eg about enhanced surveillance and control in a pre-crime society and the risk of bias or even manipulation in (automated) decision-making. In view of the stakes involved, the need for regulation of AI and its alignment with human rights, democracy and the rule of law standards has been amply recognised, both globally and regionally. The lawfulness, social acceptance and overall legitimacy of AI, big data and automated decision-making in criminal justice will depend on a range of factors, including (algorithmic) transparency, trustworthiness, non-discrimination, accountability, responsibility, effective over-sight, data protection, due process, fair trial, access to justice, effective redress and remedy. Addressing these issues and raising awareness on AI systems' capabilities and limitations within criminal justice is needed to be better prepared for the future that is now upon us.

This special issue on 'Artificial intelligence, big data and automated decision-making in criminal justice' comprises topical and innovative papers on the above issues, centred around AI and big data in predictive detection and policing, liability issues and jurisdictional challenges prompted by crimes involving AI, and AI-assisted and automated actuarial justice or adjudication of criminal cases.

Gert Vermeulen is Senior Full Professor of European and international Criminal Law and Data Protection Law, Director of the Institute for International Research on Criminal Policy (IRCP), Di-rector of the Knowledge and Research Platform on Privacy, Information Exchange, Law Enforcement and Surveillance (PIXLES) and Director of the Smart Solutions for Secure Societies (i4S) business development center, all at Ghent University, Belgium. He is also General Director Publications of the AIDP and Editor-in-Chief of the RIDP.

Nina Peršak is Scientific Director and Senior Research Fellow, Institute for Criminal-Law Ethics and Criminology (Ljubljana), Advanced Academia Fellow (CAS Sofia), Member of the European Commission's Expert Group on EU Criminal Policy, Independent Ethics Adviser, and Co-Editor-in-Chief of the RIDP.

Nicola Recchia is Postdoc Researcher in Criminal Law at the Goethe-University Frankfurt, Germany. He is also member of the Young Penalists Committee and of the Scientific Committee of the AIDP.

www.maklu.be
ISBN 978-90-466-1130-2