United Nations
Interregional Crime and Justice
Research Institute

CENTRE FOR
ARTIFICIAL
INTELLIGENCE
AND ROBOTICS

SPECIAL COLLECTION ON ARTIFICIAL INTELLIGENCE

Disclaimer

The opinions, findings, conclusions and recommendations expressed herein are those of the authors and do not necessarily reflect the views and positions of the United Nations and UNICRI, or any other national, regional or international entity involved.

The designation employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of the Secretariat of the United Nations and UNICRI, concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

UNICRI has requested the authors to provide a deep analyses on the topic at hand, this does not imply any kind of endorsement from the part UNICRI neither of the UN Secretariat regarding specific references to Member States. Similarly, the mention of specific institutions, companies or of certain manufacturers' products does not imply that they are endorsed or recommended by the Secretariat of the United Nations or UNICRI in preference to others of a similar nature that are not mentioned.

Acknowledgements

This Special Collection on Artificial Intelligence has been collated by Ms. Sophie van de Meulengraaf with the support of Mr. Odhran McCarthy and Ms. Ana Rodriguez Tamayo, under the overall guidance of Mr. Irakli Beridze. UNICRI would like to express its appreciation and to acknowledge the contributions of all the authors that contributed to this special collection and the external reviewers that supported the process.

Viale Maestri del Lavoro, 10, 10127 Torino – Italy

Tel: + 39 011-6537 111 / Fax: + 39 011-6313 368

Website: www.unicri.it

E-mail: unicri.publicinfo@un.org

FOREWORD

Our world is undergoing a massive technological transformation, involving all levels of public and private life. In particular, developments in artificial intelligence (AI) are challenging traditional perspectives, boundaries and methods and promising enhanced efficiency and effectiveness in the process. Just as this technology is heralding change in healthcare, retail, transportation and the financial services, advancements in AI are coming to and will increasingly play a role in crime prevention and the criminal justice system in the years ahead.

The disruptive nature of these technologies is already being discussed at large, but much work remains to be done in order to advance understanding of the change on the horizon and how communities concerned and society as a whole can prepare for it, particularly from the perspective of adapting and developing policy and legislation. This entails not only understanding how to shape national and international governance frameworks, but also how to ensure that these frameworks remain relevant in light of the pace of technological innovation. At the same time, it is also imperative that we better understand how to safeguard human rights and fundamental freedoms through such frameworks as, indeed, respect for these must be the very foundation upon which we work. If AI falls foul of these, the implications will be far reaching, impacting the lives of individuals and undermining public trust in authorities.

It is undeniably a fascinating time in which we find ourselves. The potential of the AI for law enforcement, legal professionals, the court system and even in penal system to augment human capabilities is enormous. However, we will need to truly test the limits of our creativity and innovation to overcome the challenges that come with these technologies, as well as to develop entirely new approaches, standards and metrics that will be necessitated by them. We must begin to generate more thought on this and on the full range of legal aspects of AI, identify current use-cases and possible future scenarios and test boundaries.

In this regard, contained within the pages of this *UNICRI Special Collection on Artificial Intelligence*, are a selection of articles from innovative minds in academia and it is our sincere hope that this collection will be a valid contribution. We hope these articles will stimulate discussion in this domain and on how to shape the design of the policies and legal frameworks of the future and provide guidance to those who will build the AI-based tools and techniques in question.

It is, however, incumbent upon me to conclude by underscoring that no specific AI use-case mentioned in this collection should be perceived as an endorsement by UNICRI. Our intention is not to suggest what AI should be used for, rather it is to provoke thought, discussions and perhaps even possible solutions to challenges we will face in this emerging domain.

Irakli Beridze

Head of Centre,

Centre for AI and Robotics

UNICRI

TABLE OF CONTENTS

1. BALANCING TESTS AS A TOOL TO REGULATE ARTIFICIAL INTELLIGENCE IN THE FIELD OF CRIMINAL LAW

Francisco Tomás Rizzi* and Agustín Pera**

Abstract

The advent of new technologies entails undeniable benefits, although with intrinsic risks. Having analyzed past experiences, this paper proposes that Criminal Law, as a science, should nourish from the contributions that other sciences have to offer, preserving, however, its constitutional structure. Regarding Artificial Intelligence, specifically when the implementation of AI-rigged systems collides with citizens' rights, the importance of judicial control in concrete cases is emphasized, as well as the way this is done by applying a filter of proportionality in which the rights and the principles in conflict are weighed. Finally, it analyses how judicial control can be applied in the face of the implementation of some AI systems commonly used in security and in jurisdictional areas.

Keywords: Criminal Law, New Technologies, Artificial Intelligence, Judicial Control, Filter of Proportionality, Balancing Test.

Introduction

The advances in the field of Artificial Intelligence (AI)[1] signified a change of paradigm in the development and study of different sciences. Criminal Law was not an exception. In terms of security, the possibility of carrying out arduous research studies and intelligence tasks massively, and without regulation limits, frequently resulted in the lack of the proper control to guarantee the respect for Human Rights.

1 The authors would like to thank Clara Rizzi for her comments and suggestions regarding the writing of the present paper
*Criminal Law Specialist and Magister in Judiciary Law. PhD in Law candidate. Currently serving as Secretary in the General Prosecutor Office of the Department of San Isidro, Buenos Aires.
** Lawyer, graduated with honours from the Buenos Aires University (UBA), currently pursuing a Master's degree in Criminal Law (UBA). Legal Clerk at a Buenos Aires Criminal Court.

As we know, there is no single accepted and rigid definition of AI; nevertheless, in the present paper we are going to refer to AI systems in a broad sense. The European Economic and Social Committee divides AI into narrow AI and general AI: narrow AI is capable of carrying out specific tasks and general AI is capable of carrying out any mental task that can be carried out by a human being. See Catelijne Muller, "European Economic and Social Committee. 526th EESC plenary session of 31 May and 1 June 2017" in *Official Journal of the European Union*, 08/2017: C 288/1–288/9, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52016IE5369&from=EN

By analyzing the result of past experiences, we propose a way of making a judicial control in the concrete case of the implementation of new technologies under the existing legal principles, ensuring, in such way, that rationality would not be overlooked.

This underlying idea maintains that AI must work as an instrument to the service of humanity and not as a double-edged sword that might turn against us.[2]

Criminal Law regarding Medicine and Biology

As we have mentioned above, in the 19th Century, the impact of Medicine and Biology on Criminal Law resulted in the creation of a new school known in Continental Europe as special prevention.

In this field, therapeutic doctrines of social defense were developed,[3] as was the case of the so-called *Scuola Positiva* in Italy, whose main authors were Cesare Lombroso, Enrico Ferri, Raffaele Garófalo, Eugenio Florian and Filipi Grispigni.[4] Backed by the medical developments of the time, they argued that offenders were inferior beings, congenitally determined, perverted in different degrees. They concluded that an offense could be anthropologically explained[5] and that the punishment should be imposed by society to their own defense, giving way to hygienic-preventive practices, therapeutic- repressive or surgical-eliminative measures, depending on each case.

Likewise, corrective doctrines were developed, as was the case of the famous *Marburger Programm* launched in 1882 by the German author Franz Von Liszt, who postulated that punishment should have the immediate effect of correction, intimidation and neutralization. In such a way, the punishment should be determined by the type of criminal:

a) rehabilitation for those capable and corrigible;

b) deterrence for criminals who do not need to be corrected; and

c) incapacitation of criminals who are considered incorrigible.[6]

In Common Law, Jeremy Bentham developed the theory of utilitarianism which had the 'utility criterion' as a guideline, even within the concept of morality, according to which, all actions could be measured based on their results. He believed that the punishment had to have a double reformatory function: on the one hand to influence society, and on the other hand, to rectify the offender. From this idea he afterwards developed his famous panopticon.[7]

At that time, the perception of the biologist and mechanist arose globally, which led to the development of utilitarianism-type theories, and threatened -at least- the long standing Kantian concept of the human being as an end in itself.[8] Therefore, punishments started to be applied, not proportional to their culpability, but in accordance to a person's character in consideration of their posed 'danger'.

These theories arose basically due to two interests that were vying in parallel: the scholar and the politician. On the one side, only the natural sciences and exact sciences were considered actual 'sciences' at the time, and Law was perceived as a non-scientific discipline. These authors tried to overpass this juncture and provide Law with the content of the most prestigious sciences by reinventing it into an exact science. On the other hand, on a political level, the subordination of Law to exact sciences responded to what, at that

2 Stephen Hawking, *Breves respuestas a las grandes preguntas* (Buenos Aires, Crítica, 2019), 230-233. Stephen Hawking points out the importance of planning in advance and of averting inherent risks in respect of the development of AI, since doing the opposite may result in "the worst thing ever to happen to humanity", speculating about a world with a super-intelligent AI capable of developing a self-will in conflict with our own.

3 Luigi Ferrajoli, *Derecho y Razón* (Madrid: Editorial Trotta, 1997), 265.

4 Ferrajoli, 266.

5 Thomas Vormbaum, *Historia Moderna del Derecho Penal Alemán* (Valencia: Tirant Lo Blanch, 2018), 204

6 Franz Von Liszt, *La idea de Fin en el Derecho Penal* (México: Universidad Nacional Autónoma de México y Universidad de Valparaíso de Chile, 1994), 112.

7 Joaquín Escriche, *Compendio de los Tratados de la Legislación Civil y Penal de Jeremías Bentham* (Madrid: Librería de la Viuda de Calleja e Hijos, 1839), 91; José Juan Moreso, *Jeremy Bentham: Luces y Sombras*, (Barcelona: Anales de la Catedra Francisco Suarez, 2013), 226.

8 Immanuel Kant, *Fundamentación de la Metafísica de las Costumbres* (San Juan: Pedro M. Rosario Barbosa, 2007), 42.

time, was considered a value: Legal Security.[9] By not giving way to any value judgment, Law seemed to be shielded from all kinds of manipulation by interpretation, ensuring that the judge could actually be *la bouche qui prononce les paroles de la loi*,[10] leaving unaffected the field of action of the legislative power.

In short, the conclusion was that due to the impact of those scientific advances on criminal law, the latter ceded them its substance and lost, at some point, its essence; or at least deviated from the direction it was headed when it was being propelled by idealism which put the person at the center of the equation.[11] In this manner, the notion of human dignity and the respect of his freedom were ignored in the pursuit of the reformation of the defendants' character, or to obtain a result useful for the rest of society[12].

Criminal Law and Neurosciences

The advances in neuroscience produced a similar impact on legal science,[13] of which the initial sign is attributed to Benjamin Libet in the 80's, who claimed that free will was an illusion.[14] This gave way to a new concept of 'update determinism' and the ancient discussion between determinism and indeterminism was reedited, with heavy repercussions today regarding culpability.

Likewise, the advances in this field by the development of the positron emission tomography (PET), the nuclear magnetic resonance (RM o fMRI) and the magnetoencephalography, intended to refute the theory that human beings act freely and voluntarily, which, as it is known, is the premise for the prevailing theses about culpability: the possibility of having acted otherwise as the foundation of reproachability. The main tenet of the concept of retributive justice resides in the idea that punishment is justified inasmuch as its enforcement compensates, or at least does not exceed, the damage caused by a guiltily committed offence.[15]

This was the origin of the so-called 'neurodeterminist postulates', which claimed that legal control of the offences should be conducted by intervening offenders' brains through the appropriate 'neurological treatments,[16] or by means of 'enhancement',[17] and not with punishments based on their reproachability. In other words, a Legal System of treatment measures, and not of punishments based on merit.[18]

---

9 Santiago Mir Puig, "Límites del Normativismo en el Derecho Penal", *Imputación Objetivas y Dogmática Penal*, (Mérida: Universidad de los Andes, 2005), 31. The Spanish author explains that, in the dogma, this had concrete consequences in the continental theory of crime. On the other hand, the action implied only the physical movement determining the causation of the result. The anti-legality was a mere description of a situation specially planned and culpability was the confirmation of a psychological connection between fact and mind. These concepts, as such, lacked all social significance and value judgement, and of reproachability, which led to what can be considered, a dehumanization of the process.

10 Montesquieu, Del espíritu de las Leyes (Buenos Aires: Libertador, 2004), 137.

11 Vormbaum, 196. In this respect, Vormbaum points out that the humanitarian aspiration of this illustration ceded in favour of a 'relentless science'.

12 In fact, the absolutization of these theories goes against the actual structure of modern democracies, in which freedom and responsibility - as the principle of culpability - configures a nuclear characteristic. See Bernardo Feijoo Sánchez, *La pena como institución jurídica*, (Buenos Aires: B de F, 2017), 156. Likewise, Eugenio Zaffaroni indicates that criminal materialism - the biological theories - configures a "dehumanized" criminal law. See Eugenio Raúl Zaffaroni, *Tratado de Derecho Penal Parte General* (Buenos Aires: Ediar, 1998), 121. Furthermore, Ferrajoli explains that corrective doctrines are not compatible with the respect of the human person. He believes that they go against the values of freedom and equality since they suppose the idea of offenders as inferior or abnormal human beings. He argues that considering criminal treatment as an absolutized idea damages the concept of human dignity and the democratic principle of respect and tolerance of human subjectivity. See Ferrajoli, 271.

13 As well as on related sciences, such as Neurophilosophy, Philosophy of Mind, and Cognitive Psychology.

14 This way, Gerhard Roth sustains that the voluntary action directed by a "conscious self" was an illusion, since it had been discovered that, as a consequence of the concatenation of the amygdala, hippocampus and ventral and dorsal node, the emotional memory of experience (which works on an unconscious level) was what defined the making of decision. This way, they would have a place in the limbic system one or two seconds before they could be perceived consciously. Wolfgang Prinz, on his part, understands free will as a social institution which does not correspond to the scientifically demonstrable reality, from a psychic point of view. Wolf Singer believes that the perceptions we experienced as objective are nothing but the result of constructive processes. This way, as well as with animal behavior, the man is completely determined, and each action is necessarily a result of the combination of the constellation that gives origin to the real stimulus with the immediately preceding brain state, determined by the genetic organization, a multitude of epigenetic factors and educational processes which affect the nervous chain and, finally, due to the immediate previous history, the dynamic neuronal interaction, see Eduardo Demetrio Crespo, "Compatibilismo Humanista: Una Propuesta De Conciliación Entre Neurociencias Y Derecho Penal" in Neurociencias y Derecho Penal, (Madrid: B de F, 2013).

15 Mercedes Pérez Manzano, "Fundamento y fines del derecho penal. una revisión a la luz de las aportaciones de la neurociencia", Indret: Revista para el Análisis del Derecho, 02/2011, https://indret.com/wp- content/themes/indret/pdf/818.pdf.

16 Ibid.

17 Reinhard Merkel, "Novedosas intervenciones del Cerebro. Mejora de la Condición Humana Mental y Límites en el Derecho Penal", Revista de Derecho Penal. Culpabilidad: Nuevas Tendencias I, No. 2, 2012 (Buenos Aires: Rubinzal Culzoni, 2013); Reinhard Merkel defines 'enhancement' as a suitable procedure to create a situation which is modified in a physiological or mental level, that cannot be considered as a healing treatment and that is perceived as a betterment to the person.

18 Bernardo Feijoo Sánchez, "Derecho Penal y Neurociencias ¿Una relación tormentosa?", Indret: Revista para el Análisis del Derecho, No. 1/2020, https://indret.com/?autor=bernardo-feijoo-sanchez

Following the aforementioned, three different positions arose with regards to the intrusion of Neuroscience in Criminal Law. The first one claims that the development of this science is not connected with Criminal Law since its object is not determined by scientific knowledge, but by normative concepts which may differ from the empirical.[19]

The second one defends the dependency of Criminal Law to neuroscientific knowledge and claims that, nowadays, a transformation is necessary to substantiate its goals.[20] The third one recognises that Criminal Law cannot overlook scientific knowledge, and points out that it cannot entail a modification of its substantiation model either. According to this position, criminalists should value scientific knowledge in attention to the goals and functions of Criminal Law.[21]

The third position, which we considered to be correct, was firstly adopted in the case *Ropper v. Simmons22* and afterwards in *Graham v. Florida*,[23] by the Supreme Court of the United States.[24]

In those cases, the Court drew on the scientific advances in the field of neuroscience to demonstrate that the culpability of an underage is less than that from an adult, and on this basis, it sustained that they could not bear the same punishment given similar events. Nevertheless, the central point of this argument is that the legal system made use of science not to support a scientific-biological concept but to give content to a normative concept of culpability; that is to say, not to describe it but to outline the legal concept considering the neuroscientific contributions without yielding the essence of the concept itself.

Criminal Law and the advent of Artificial Intelligence

It is observed that, in both experiences, the legal science had to face scientific- empiric advances of foreign sciences, forcing it to reconsider central matters of the system, the consequences of which were remarkable.

Despite the result of the first experience, the emergence of reductionist positions about human behaviour is outstanding. These positions, supported now by neurosciences, propose a Legal System, no longer of punishments based on culpability, but of treatment measures which, at some point, recalls the Marburg Program.[25] Having adopted foreign conclusions and transferring them to the legal system, as Winfried Hassemer claimed, it was a categorical mistake.[26] Notwithstanding this, it would have also been a mistake to disregard the scientific advances and diagram a legal science completely unconnected to them.

In this sense, we understand that the path taken by the third position, which proposes the use of raw data to achieve legal purposes, is the most plausible option. Our position in respect to AI is rooted to this idea:

19 Winfried Hassemer, "Neurociencias y culpabilidad en Derecho penal", Indret: Revista para el Análisis del Derecho, No. 2, 2011; Günther Jakobs, "Culpabilidad Jurídico-Penal y Libre Albedrío" in Derecho Penal de Culpabilidad y Neurociencias, (Madrid: Thomson Reuters, 2012); Bernardo Feijoo Sánchez; Peter-Alexis Albrecht, "Culpabilidad: Restricciones al poder punitivo" in Culpabilidad: Nuevas Tendencias I (Buenos Aires: Rubinzal Culzoni, 2013), 31; Stephan Stübinger "¿Persona o Paciente? Comentarios sobre el Principio de Culpabilidad en el Derecho Penal desde el punto de vista de la investigación del cerebro" in Revista de Derecho Penal. Culpabilidad: Nuevas Tendencias I, No. 2, 2012); et al.

20 Francisco Rubia, Neurociencia y Derecho penal: nuevas perspectivas en el ámbito de la culpabilidad y tratamiento jurídico-penal de la peligrosidad, comp. by Maroto Calatayud and Demetrio Crespo, (Madrid: B de F, 2013),185-190.

21 21Pérez Manzano, "Fundamento y fines del derecho penal. una revisión a la luz de las aportaciones de la neurociencia"; Eduardo Demetrio Crespo, "Compatibilismo Humanista: Una Propuesta De Conciliación Entre Neurociencias Y Derecho Penal" in Neurociencias y Derecho Penal, (Madrid: B de F, 2013); Manuel Cancio Meliá, "Psicopatía y Derecho penal: algunas consideraciones introductorias", Derecho Penal de Culpabilidad y Neurociencias (Madrid: Thomson Reuters, 2012).

22 Roper v. Simmons, (543 U.S. 551, 2005).

23 Graham v. Florida (560 U.S. 48, 2010); see also Cara H. Drinan, "Graham on the Ground" in Washington Law Review No. 87, (2012). 51.

24 In Roper v. Simmons, the court analyzed if it was permissible under the Eighth and Fourteenth Amendments to the Constitution of the United States to execute a juvenile offender who was older than 15 but younger than 18 when he committed a capital crime and, establishing the culpability, they affirmed "(...) scientific and sociological studies respondent and his amici cite tend to confirm, '[a] lack of maturity and an underdeveloped sense of responsibility are found in youth more often than in adults and are more understandable among the young. These qualities often result in impetuous and ill-considered actions and decisions.'" In Graham v. Florida, it was discussed if the Constitution permits a juvenile offender to be sentenced to life in prison without parole for a nonhomicide crime and, in that opportunity, they indicated "(...) developments in psychology and brain science continue to show fundamental differences between juvenile and adult minds. For example, parts of the brain involved in behavior control continue to mature through late adolescence (...)."

25 Feijoo Sánchez, 10.

26 Hassemer, "Neurociencias y culpabilidad en Derecho penal", 6. The ex-Judge of the German Constitutional Court explained that the categorical error derives from the violation of a principle of knowledge and science theory, this is, all sciences can only see what their instruments allow them to see and find answers where their instruments allow them to formulate questions corresponding to a categorical answer. The instruments of each science are determined by their formal object, and this is controlled by a duty-based system, paradigms, methods, and instruments. In such way, the empiric-method sciences would not be able to put to trial whether 'liberty' exists, for instance.

scientific advances should be adapted to Criminal Law[27] and not the other way around. That is to say, these should be integrated in such a way that they would not hinder the legal foundations which were built around the respect for Human Rights.[28]

It would be difficult to imagine how AI will influence this field of Law in the future, but, in the same way that it was able to make use of exact, biological and technological sciences without giving up its objective, we understand that it would be possible as well to control AI utilization by establishing legal guidelines and principles.[29]

However, due to the fact that there is no proper regulation to implement AI globally,[30] this cannot lead to a sort of unlimited permission for its implementation with no control at all. There is no doubt, or at least there should not be, that we can demand our Government to comply with a minimum implementation standard when using AI in public policies.[31] It is our understanding that such a standard can be found in the proper respect of Human Rights, demanding a reasonable implementation of technologies within the limits of the rule of law.[32]

One of the problems we face is that we do not count, at least for now, with a way of trespassing axiological values to exact value units which can be introduced inside an algorithm, nor a method to conjugate in it any reference of principles. Such problem resides, not within technological fields, but within legal science itself: Law cannot define exactly the value of legal assets and its philosophy cannot fix the axiological weight of the values and principles in mathematic terms either.[33] For such a reason, we consider that AI systems for judicial decision making are not able to provide a legal response to hard cases[34] without falling into a benthanian utilitarianism.

Following the above, for the purposes of Law, AI can only be considered as an external tool. In view of the great amount of AI systems which are being developed, not only regarding security but also jurisdiction, we understand that fixing the limits of its utilization is a necessity. As long as such limits are not legally regulated, its implementations should be examined in the concrete cases by an effective judicial control.[35]

Law counts with a series of tools to determine when a rule is akin to the legal system. As far as AI systems are concerned, whilst the goal of its government implementation is to increase effectiveness in security-related

27 We refer to Criminal Law in an objective sense: "a range of legal rules, valuations and principles which discourage and prohibit the commission of crimes and are associated to them, such as budget, punishment and / or safety measures as legal consequence". See Santiago Mir Puig, Derecho Penal, Parte General, (Barcelona: Ed. Reppertor. 2006) 45. Also, when we talk about legal rules, it should be understood in a global sense, including the criminal procedural law, punishment law, penitentiary law, juvenile criminal law and criminology. See Claus Roxin, Derecho Penal Parte General Tomo I. Fundamentos. La estructura de la teoría del delito (Madrid: Civitas, 1997), 44.

28 In such a sense, Hassemer refers to a science of Criminal Law oriented to consequences. This means that the legislator, criminal justice and prison administration are not satisfied with prosecution and compensation through offender's atonement, but their goal is to improve the offender and contain delinquency as a whole. See Winfried Hassemer, Fundamentos del Derecho Penal, (Barcelona: Bosch, 1984) 35. It should be added that, actual legitimacy of Criminal Law is verified by approval in accordance to the Constitution. See Günther Jakobs, Derecho Penal. Parte General (Madrid: Marcial Pons, 1997) 44.

29 It should be noticed that Criminal Law counts with stiffer punishments to normative injuries. This are so dangerous that civilized societies should secure them in different ways and protect they do not fall into the wrong hands, and make sure they are used carefully, equally and proportionately. See Winfried Hassemer, Crítica al Derecho Penal de Hoy, (Buenos Aires: Editorial Ad.Hoc., 1995) 19. Thereof, it appears that all ius punendi restricting principles play a fundamental role when allowing the rationalization of their use within democratic limits. See Santiago Mir Puig, Derecho Penal, Parte General, (Barcelona: Reppertor, 2006) 104.

30 In a similar way, the Special Rapporteur Philip Alston, called for the regulation of AI to ensure compliance with Human Rights and for a rethinking of the positive ways in which the digital welfare state could be a force for the achievement of vastly improved systems of social protection (Philip Alston, "Extreme poverty and Human Rights", report submitted in accordance with Human Rights Council resolution 35/19, UN - General Assembly).

31 In Dworkin words, if we cannot claim the government to provide adequate solutions to citizens' rights, we can claim them to at least try their best to do so: "(...) we can demand they take rights seriously". See Ronald Dworkin, Los Derechos en serio, (Barcelona: Ariel Derecho, 2010) 278.

32 Bernardo Feijoo Sánchez rightly points out that Constitutional Law demands a relation between legitimacy and effectiveness in such a way that an ineffective instrument or, in his case, inefficient, losses its legitimacy, taking into consideration the costs in legal terms. See Bernard Feijoo Sánchez, La pena como institución jurídica, (Buenos Aires: B de F, 2017) 92.

33 In such way, Winfried Hassemer and Francisco Muñoz Conde points out that the respect of human dignity is a legal value that does not admit any economic quantification or exchange with other values. See Winfried Hassemer, Francisco and Muñoz Conde, Introducción a la Criminología y al Derecho Penal (Valencia: Tirant Lo Blanch, 1989), 101.

34 To reach a definition related to this kind of cases it can be appeal to Dworkin's work. Ronald Dworkin, Los Derechos en serio, (Barcelona: Ariel Derecho, 2010), 146.

35 We should make clear that we refer to legal control over administrative and judicial decision, but not to legislation judicial control, defined as "strong judicial control" by Jeremy Waldron, who called into question his fundaments. See Jeremy Waldron, Contra el Gobierno de los Jueces, (Buenos Aires: Siglo Veintiuno, 2018), 61.

tasks, balancing tests[36] becomes particularly interesting. This is so due to the fact that collective goods[37] come into play against individual rights,[38] demanding a point of proportional balance between the benefits of the first ones and the affectation of the second ones. Balancing is an extremely useful legal method to solve conflict of interests, which legitimizes its interference in rights and principles.[39]

The balancing test found its origin as an argumentative technique of proportionality developed by the German Constitutional Court[40] and emigrated afterwards to other constitutional courts to finally establish itself as an instrument of excellence in the international system of Human Rights. It is a method of great utility when determining if the affectation of an individual right becomes disproportionate in the pursuit of a collective good. It imports a way of materialization from the proportionality principle.[41] This principle, which went up as one of the pillars of the modern Rules of Law[42] and also as a cornerstone of numerous theories of Criminal Law,[43] must be, however, applied cautiously since, as Günther Jakobs points out, it is too much of a formal concept,[44] which should be filled with the valuations at stake and, in that sense, allowing the entrance of the operator's subjective questions.[45]

Regarding the balancing test, the most technical development was made mostly by Robert Alexy, who outlines the "weighing formula." Alexy believed that proportionality had three sub-principles: the principles of suitability, of necessity, and of proportionality in the narrower sense.[46] The latter is the one which concerns us most since it refers exclusively to the legal possibilities. The concept of balancing arose in this sub-principle. According to this theory, in front of fundamental rights, the rule states that: the greater the degree of non-satisfaction of, or detriment to, one principle, the greater must be the importance of satisfying the other.[47]

Taking the aforementioned to our subject-matter field, this would involve determining which rights should be protected and which rights should be affected when using an efficient AI program. Three steps are thus proposed: a) to define the affectation degree of one of the principles or rights, b) to assess the importance of satisfaction of the opposite principle / right and c) to define if the importance of the opposite principle / right satisfaction justifies the restriction of the first one. An intensity scale (mild, moderate and severe) was established to measure the weight of the intervention and the degree of importance which justifies it.[48] These steps would demonstrate whether an AI program can be applied rationally and consequently, if it is applicable under ethical parameters.[49]

---

36 It should be pointed out that to analyze a collision between individual rights and collective goods it is necessary to firstly count with or enter into a theory of individual rights´ principles. This means to understand principles as optimization mandates in opposition to the rules implied in a definite mandate (as proposed by Robert Alexy - see Robert Alexy, El concepto y la Validez del Derecho, (Barcelona: Gedisa, 2004) 161); or else, the dichotomy between rights and principles (as proposed by Ronald Dworkin – see Dworkin, 32).

37 Although it is not easy to define collective goods, it is possible to state that, according to Robert Alexy, if referring to goods which structure is not distributive (i.e. it cannot be distributed among citizens) and has normative status (i.e. ordered by a legal system); and substantiated in a rule, which, according to Alexy, these rules should be substantiated in the consensus theory of truth. See Alexy, 179.

38 Understood as "political rights" whether basic or institutional, but with validity against majority decisions, according to Dworkin's theory, (see Dworkin, 37); understood as "rights as optimization mandates", according to Alexy (see Alexy, 185); or understood as "rights based on the idea of personal autonomy", i.e. a principle according to which is prescribed the maximization of people's capacity of choosing and materializing a life ideal, according to Carlos Nino (See Carlos Nino, Una teoría de la justicia para la democracia. (Buenos Aires: Siglo Veintiuno, 2013) 127).

39 Winfried Hassemer, Crítica al Derecho Penal de Hoy, 64. Hassemer considers it even as a "strong instrument".

40 See Elena Bindi, "Test de proporcionalidad en el age of balancing", UNED Revista de Derecho Político, No. 96, 05/2016. https://doi.org/10.5944/rdp.96.2016

41 Robert Alexy, Teoría de la Argumentación Jurídica. La teoría del discurso racional como teoría de la fundamentación jurídica (Lima: Palestra, 2010) 459.

42 It is the principle according to which all government intervention affecting citizens' rights should be limited. It supposes a constitutional requirement when, in Criminal Law, the state intervention affects fundamental rights. See Santiago Mir Puig, Derecho Penal, Parte General, (Barcelona: Reppertor, 2006) 127.

43 See Von Hirsch's work in the United States, Andrew Von Hirsch, Censurar y Castigar, (Madrid: Trotta, 1998) 31; or Hörnle's work in Germany, Tatjana Hörnle, Determinación de la Pena y Culpabilidad, (Buenos Aires: FD, 2003) 81.

44 Günther Jakobs, Derecho Penal. Parte General, (Madrid: Marcial Pons, 1997) 568.

45 It should be noted that its formality, although it may be dangerous since it allows entering subjective valuations, if used properly it may also be a dynamic principle since it can receive social changes and the values of each particular society.

46 Alexy, 458. In a similar sense, Hassemer calls this third sub-principle 'enforceability". However, he acknowledges that this is the final component of the principle, according to which, law can only advance if a prohibition or a mandate is enforceable. See Winfred Hassemer, "El principio de proporcionalidad como límite de las intervenciones jurídico penales" in Límites al Derecho penal. Principios operativos en la fundamentación del castigo, (Barcelona: Atelier, 2012), 195.

47 Alexy, 460. It should be clarified that for the German author, fundamental rights have the structure of principles and they constitute an optimization mandate.

48 Alexy, 462. We understand that the measurement will depend on the values and ideals of each particular society, and even though there are global agreements regarding the importance of respecting certain principles of ha rights protection, it would not always be the same in each society or particular country, due to idiosyncrasy or the specific needs of each region. We consider that the best way to achieve a correct measurement is through legal argumentation procedures such as those presented by Alexy; or the collective discussion that Carlos Nino calls 'Epistemological Constructivism'. See Carlos Nino, Una teoría de la justicia para la democracia, (Buenos Aires: Siglo Veintiuno, 2013), 89. In this sense, Alexy points out that the legitimacy of weight in Law depends on it rationality.

49 See the judicial decision of the Interamerican Court of Human Rights in the case Kimmel v. Argentina; and also, European Court of Human Rights, Belgian Linguistic Minorities, July 23, 1968.

In short, this procedure proposes to analyse if the affectation of Human Rights is justified by the degree of interference of AI systems in society.

It is reasonable to believe that, in order to guarantee the proper exercise of citizens' rights and that they are effective as much as possible, it is necessary that everyone give up a scope of exercise of that or any other right. Likewise, people should give up individual rights in order to guarantee collective goods.[50] However, this regulation should not involve an excessive affectation when the main goal is not proportional to the affectation of an individual right. In this way, an intrusion of magnitude can only be justified when an undoubtedly imperious interest is demonstrated.[51] Dworkin points out that we must acknowledge that the a government has a reason for limiting rights if it plausibly believes that a competing right is more important[52] and that the course of government is to steer to the middle, to balance the general good and personal rights, giving to each their due.[53] In Manuel Atienza's words, it is not about inventing Law but developing it.[54] This is what we propose in order to cope with the application of AI in prevention matters.

Application of the Balancing Tests

In what follows, we will point out how balancing tests can be applied to analyse the implementation of AI systems.

Firstly, we will consider PredPol.[55] This system presents itself as a predictive surveillance system, which projects on a map the places and times where specific crimes are probable to happen, allowing the efficient allocation of resources to prevent them.[56]

Nonetheless, the first thing we can note about this program is that its utilisation can lead to stigmatization of determined areas for having higher rates of crimes.[57]

Taking into consideration Alexy's steps, it is easily detectable that the purpose of the program is to apply a reinforcement of police control in certain areas, affecting the freedom of people residing there more than that of those who reside in safer areas. The reinforcement of the police control would affect unevenly the right to freedom and privacy of citizens, since they would lead, in certain areas, to more individual retentions, requisitions, etc. Imagine in this sense, and considering the triadic scale proposed by the author, that it is possible to establish that the mentioned intrusion - moderately applied and respecting personal guarantees - implies an average affectation of rights of people residing in the poorest neighborhoods. On the other hand, considering the opposite principle, it could likewise be established that there is a special interest to safeguard the competing right to safety as far as possible. Going back to the scale, imagine it is possible to establish that insecurity is an extremely serious threat. In this respect, and in view of the third step, it could be concluded that, although the uneven restriction to freedom and privacy is considerable, it is more important to guarantee safety in those areas stricken by criminal activity. Therefore, affectation is proportional to the degree of importance of its satisfaction, thus the utilization of a system like Predpol may result well-balanced.

50 Dworkin, 289. Likewise, John Rawls points out that, by limiting liberty to guarantee equal conditions to all citizens, i.e. on behalf of the common interest, the government acts on a principle that would be chosen in the "original position" (México: Fondo de Cultura Económica, 2006), 203.

51 In such way, the American Convention On Human Rights "Pact of San Jose, Costa Rica", establishes the Personal Responsibilities in the Article 32 called "Relationship between Duties and Rights": "1. Every person has responsibilities to his family, his community, and mankind. 2. The rights of each person are limited by the rights of others, by the security of all, and by the just demands of the general welfare, in a democratic society";

52 Dworkin, 288.

53 Dworkin, 294.

54 Manuel Atienza, "Constitucionalismo y Derecho Penal" in Constitución y Sistema Penal, (Madrid: Marcial Pons, 2012), 35.

55 See Predpol's website, https://www.predpol.com/. This robot was created in an investigation Project which was carried out by Los Angeles Police Department and the UCLA, tending to capitalize the information obtained from reported criminal acts.

56 This system uses three specific pieces of information (kind of felony, location and time of occurrence), which are entered into a mathematic algorithm that analyses them to predict which the most problematic areas of the city are and, based on that, organizes police patrol routes in order it strengthen prevention in the most dangerous areas. This budding algorithm works by linking certain criminal behaviours with a mathematic structure in order to determine the evolution of criminal patterns.

57 María Hernández Giménez, "Inteligencia Artificial y Derecho Penal," in Actualidad Jurídica Iberoamericana, No.10, June 2019. 817.

Secondly, we will analyze systems designed to help legal labor, among which we will focus on COMPAS[58]. This software uses algorithms to evaluate the risk of potential relapses. It is intended as a useful tool for judges in their making of legal decisions.[59]

The implementation of this system has been object of criticism among which the following stands out: a) due to the fact that algorithms are a commercial secret, it was not unveiled how the system actually works, limiting the exercise of the right to defense hence affecting the proper process;[60] b) considering that the system draws from what a programmer teaches it, exhibiting for instance, racist tendencies, they would be impacted in its function, resulting in a possible partiality;[61] and c) the system is not more precise than the predictions made by people with little or no experience in criminal justice: indeed, a linear system using only two parameters (age and number of previous convictions) will deliver almost the same results.[62]

Based on the aforementioned, it is reasonable to notice that although it is necessary to find measures to contribute to reduce error and judicial arbitrariness, we are in the presence of a tool whose suitability is under discussion and it should be properly sorted out in order to advance with the balancing test. However, for the purpose of our exercise, if it is possible to overcome the obstacle mentioned, unfavorable results would be noticed for its implementation. In fact, paying specific attention to the first critic displayed, it is easy to detect that the program implies an intrusion on the right to defense due to the fact that, as it was already mentioned, it does not reveal how the results are reached, hampering its control.[63] To this end, it could be argued that the intrusion is carried out by a 'serious' entity.[64] In respect to the opposite principle, the decisions involved tend, in general terms, to the realization of the punishment purposes, more precisely, what makes to their special- preventive type. Nevertheless, it should be noticed that this procedure is not outlined as essential for all purposes and is merely constituted as another tool to the intervening judge. Due to this, it could be established that the interest in ratifying it would fluctuate between mild and moderate. In this sense, it is inferred that the implementation of this system would not justify a restriction of such magnitude to the right to defense; hence, since it is not possible to verify the existence of proportionality, it could be concluded that it is not advisable to apply a system like COMPAS, as it was herein presented.[65]

In spite of the criticism mentioned in the previous paragraph, it should be noticed that it may subside in connection to PSA,[66] an algorithm developed by the Laura and John Arnold Foundation, which objective is to streamline the use of pretrial detention. Basically, this system tries to determine on a 1 to 6 scale how probable it is that an accused person incurred in the perpetration of new crimes and how probable it is that he

58 (Correctional Offender Management Profiling for Alternative Sanctions).

59 The conclusions arose from the answer to 137 items about the particular circumstances of each offender, including all the information connected to their criminal record. The exact number was extracted from the investigation work carried out by Julia Dressel and Hany Farid, "The accuracy, fairness, and limits of predicting recidivism," Sci. Adv. 2018;4: eaao5580 17 January 2018;and the article by Ed Young, "A Popular Algorithm Is No Better at Predicting Crimes Than Random People," in The Atlantic, January 17, 2018, (https://www.theatlantic.com/technology/archive/2018/01/equivant-compas-algorithm/550646/). See also Maria Hernández Gimenez, "Inteligencia Artificial y Derecho Penal," in Actualidad Jurídica Iberoamericana, No.10, June 2019. 817.

60 See State of Wisconsin v. Eric L. Loomis, (2015AP157-CR, 2016). This matter was put into the United States Justice: Supreme Court of Wisconsin' consideration.

61 Hernández Gimenez, "Inteligencia Artificial y Derecho Penal", 824-825. To support what is exposed thereat, see Julia Angwin, and others, "Machine Bias", in ProPublica, May 23, 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

62 Julia Dressel and Hany Farid, "The accuracy, fairness, and limits of predicting recidivism", Science Advances, January 17, 2018.

63 A joint UNICRI-INTERPOL report on AI and Robotics for Law Enforcement, establishes that to respect citizens' fundamental rights and avoid potential liability, their use in law enforcement should be characterized by the following features: fairness, accountability, transparency and capability of being explained. When explaining the concept of transparency, it was stipulated that the path taken by the system to arrive at a certain conclusion or decision must not be a "black box," and while taking about the capability of being explained, they assured that the decisions and actions of a system must be comprehensible to human users (UNICRI- INTERPOL, Artificial Intelligence and Robotics for Law Enforcement, report launched on April 2, 2019 at the High-Level Meeting: Artificial Intelligence and Robotics-Reshaping the Future of Crime, Terrorism and Security). In a similar way, the Special Rapporteur Philip Alston, stipulated that the public need to be able to understand and evaluate the policies that are buried deep within the algorithms (Philip Alston, "Extreme poverty and Human Rights," report submitted in accordance with Human Rights Council resolution 35/19, UN - General Assembly).

64 See the case State of Wisconsin v. Eric L. Loomis to find certain arguments tending to mitigate this intrusion, such as i) that COMPAS do not predict the specific probability of an offender recidivism, but the results are obtained by comparing the situation of that individual with a group of similar cases; ii) that in the report issued by COMPAS there are specified some of the elements weighted with the possibility that the same may be controverted. However, it is true that the judge and the parties do not know with exactitude all the parameters considered nor how the system really works. This may be aggravated if, indeed, is possible to ascertain the rest of the deficiencies attributed.

65 It should be reminded that the carried out weight is intimately linked to the values and ideals of each particular society, summarized in the different legal systems in force, which could actually tip the balance to the admissibility, or not, of the system under investigation. It should be noted that at the time of providing the decision in the case "Loomis," the Wisconsin Court of Justice begun it resolution by specially referring to the intention of developing mechanisms to consolidate public safety, strengthening the dominance of the principles at stake. Through the said pronouncement, it was ratified the possibility that the judge could use this instrument, even if with certain limits and under certain warnings.

66 Public Safety Assessment. See Pubic Safety Assessment's website, https://www.psapretrial.org/about.

will not attend to citations required by the justice. To obtain those results, this program uses nine parameters including the current age of the person, his/her previous convictions, his/her outstanding sentence, non-appearances to court prior to trial, and on the other hand, it leaves aside all matters related to race, sex, job positions and education levels.[67] When publishing the totality of the balancing factors to the obtainment of results, this system excels due to the fact that it allows controverting them. Ergo, the affectation of the right in question could be diminished, altering eventually the final result of the equation proposed by Alexy, and the possibility of justifying its implementation.

It should be noted that the examples given do not constitute a definitive analysis but are just an illustrative exercise of how a balancing test should be carried out. In order to be able to conduct a task of sheer scale it would be necessary to collect statistical data of greater accuracy in order to analyse them in depth, which exceeds the purpose of the present paper.

In addition to the systems cited, the District Court of The Hague has recently ruled in the so called 'SyRI Case'[68] in which the judges stated that the right to privacy prevailed over the hunt against alleged benefits fraudsters. SyRI[69] was developed by the Dutch Ministry of Social Affairs in 2014 to find out those who are most likely to commit fraud against the social system in order to receive government benefits. It is a risk profiling application. The court found that the Dutch government failed to make a balance between the right to privacy and the public interest in detecting welfare fraud, and thus, the use of SyRI was disproportionate to the aim it sought to achieve.[70]

The SyRi case revealed that it is actually possible to control AI technology with the condition that we use Law properly, placing the Human Rights in the centre of the equation. Through the correct balancing between the affected rights and the benefits of AI systems it can be determined whether AI is being used as an instrument to the service of humanity or, by contrast, as a weapon against us.

Conclusion

The arrival of systems that collect, process, analyse and interpret rapidly large amounts of information is a reality which undoubtedly carries important benefits, albeit with certain risks, among which we should emphasize the following: a) by implementing them, certain fundamental Human Rights result disproportionately affected; b) they may be potentially used by criminal or terrorist groups; and c) in a more apocalyptic level, a reckless development of them might culminate in a point of no return which may jeopardize human survival. To avoid any of the mentioned outcomes, we understand that it is necessary to count with pioneering regulations and an effective judicial control in order to allow the responsible development of these new technologies.

In view of the above, by adjusting not only to our reality but also to the objective of the present paper, what we propose briefly is to try to capitalize past experiences and develop a method to, at least, be able to fight against the syndicated risk as a starting point.

By analysing how the empiric sciences impacted the Law during the 19th Century, we tried to pose the idea that the scientific advances cannot determine Law so as to compromise its essence. On the contrary, we understand that Law, and more specifically, Criminal Law, should take advantage and make use of all the contributions that other sciences have to offer while, at the same time, preserving its constitutional structure.

Having clarified this idea, we should be extremely careful with the implementation of AI-rigged systems that could collide with citizens' rights, with the only excuse of increasing productivity and efficiency. What is more, we understand that each of them, in cases of dispute, should pass a judicial filter of proportionality in which

67 Hernández Gimenez, "Inteligencia Artificial y Derecho penal", 826-827.

68 NJCM cs/ De Staat der Nederlanden. In the case, also, the Special Repporteur Philip Alston submitted as an amicus curiae, concluding as well that SyRI posed significant potential threats to Human Rights, in particular for the poorest in society.

69 System Risk Indicator

70 For the application of the balancing test, the following factors were considered: a) that the data processing within SyRi affected a considerable part of the population; b) that it was aimed to disadvantaged groups of citizens (understanding this as being in violation of the prohibition of discrimination of Art. 14 and the right to respect for private life of Art. 8, both from the European Convention of Human Rights); c) that there was a cross-checking of data taken from multiple sources; d) that the purposes of the program and the category of information used were broad and vague; e) that risks reports give way to important sanctions; and f) that the results were of little use. Due to the aforementioned, it was concluded that the deployment needed to the implementation of SyRi was disproportionate.

their relation with rights or with the principles in conflict are analysed, determining the degree of affectation of one party and the necessity of satisfaction of the other, to finally establish if the end actually justifies the interference.

In consequence, although there is a lack of regulation to provide certain safety, we consider that, through the proper utilisation of judicial tools, it is possible to lead and control the use of AI systems within a trail marked by a strong legal frame. It is our understanding that this is the only way this technology can stand as a pillar for the evolution of humankind.

References

Hirsch, Andrew Von, Censurar y Castigar, (Madrid: Trotta, 1998).

Feijoo Sánchez, Bernardo, La pena como institución jurídica, (Buenos Aires: B de F, 2017).

Feijoo Sánchez, Bernardo, "Derecho Penal y Neurociencias ¿Una relación tormentosa?", Indret: Revista para el Análisis del Derecho, No. 1/2020, https://indret.com/?autor=bernardo-feijoo-sanchez.

Drinan, Cara H., "Graham on the Ground" in Washington Law Review No. 87, (2012).

Nino, Carlos, Una teoría de la justicia para la democracia. (Buenos Aires: Siglo Veintiuno, 201.

Muller, Catelijne, "European Economic and Social Committee. 526th EESC plenary session of 31 May and 1 June 2017" in Official Journal of the European Union, 08/2017: C 288/1–288/9, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52016IE5369&from=EN.

Roxin, Claus, Derecho Penal Parte General Tomo I. Fundamentos. La estructura de la teoría del delito (Madrid: Civitas, 1997)

Young, Ed, "A Popular Algorithm Is No Better at Predicting Crimes Than Random People," in The Atlantic, January 17, 2018, (https://www.theatlantic.com/technology/archive/2018/01/equivant-compas-algorithm/550646/).

Crespo, Eduardo Demetrio, "Compatibilismo Humanista: Una Propuesta De Conciliación Entre Neurociencias Y Derecho Penal" in Neurociencias y Derecho Penal, (Madrid: B de F, 2013).

Bindi, Elena, "Test de proporcionalidad en el age of balancing", UNED Revista de Derecho Político, No. 96, 05/2016. https://doi.org/10.5944/rdp.96.2016

Zaffaroni, Eugenio Raúl, Tratado de Derecho Penal Parte General (Buenos Aires: Ediar, 1998).

Rubia, Francisco, Neurociencia y Derecho penal: nuevas perspectivas en el ámbito de la culpabilidad y tratamiento jurídico-penal de la peligrosidad, comp. by Maroto Calatayud and Demetrio Crespo, (Madrid: B de F, 2013).

Liszt, Franz Von, La idea de Fin en el Derecho Penal (México: Universidad Nacional Autónoma de México y Universidad de Valparaíso de Chile, 1994).

Jakobs, Günther, "Culpabilidad Jurídico-Penal y Libre Albedrío" in Derecho Penal de Culpabilidad y Neurociencias, (Madrid: Thomson Reuters, 2012)

Jakobs, Günther, Derecho Penal. Parte General (Madrid: Marcial Pons, 1997)

Kant, Immanuel, Fundamentación de la Metafísica de las Costumbres (San Juan: Pedro M. Rosario Barbosa, 2007).

Waldron, Jeremy, Contra el Gobierno de los Jueces, (Buenos Aires: Siglo Veintiuno, 2018).

Escriche, Joaquín, Compendio de los Tratados de la Legislación Civil y Penal de Jeremías Bentham (Madrid: Librería de la Viuda de Calleja e Hijos, 1839).

Rawls, John, Teoría de la Justicia (México: Fondo de Cultura Económica, 2006).

Moreso, José Juan, Jeremy Bentham: Luces y Sombras, (Barcelona: Anales de la Catedra Francisco Suarez, 2013).

Angwin, Julia, and others, "Machine Bias", in ProPublica, May 23, 2016.

Dressel, Julia and Farid, Hany, "The accuracy, fairness, and limits of predicting recidivism," Sci. Adv. 2018;4: eaao5580 17 January 2018.

Ferrajoli, Luigi, Derecho y Razón (Madrid: Editorial Trotta, 1997).

Atienza, Manuel, "Constitucionalismo y Derecho Penal" in Constitución y Sistema Penal, (Madrid: Marcial Pons, 2012).

Cancio Meliá, Manuel, "Psicopatía y Derecho penal: algunas consideraciones introductorias", Derecho Penal de Culpabilidad y Neurociencias (Madrid: Thomson Reuters, 2012).

Hernández Giménez, María, "Inteligencia Artificial y Derecho Penal," in Actualidad Jurídica Iberoamericana, No.10, June 2019.

Pérez Manzano, Mercedes "Fundamento y fines del derecho penal. una revisión a la luz de las aportaciones de la neurociencia", Indret: Revista para el Análisis del Derecho, 02/2011, https://indret.com/wp- content/themes/indret/pdf/818.pdf.

Montesquieu, Del espíritu de las Leyes (Buenos Aires: Libertador, 2004).

Albrecht, Peter-Alexis, "Culpabilidad: Restricciones al poder punitivo" in Culpabilidad: Nuevas Tendencias I (Buenos Aires: Rubinzal Culzoni, 2013).

Alston, Philip, "Extreme poverty and Human Rights," report submitted in accordance with Human Rights Council resolution 35/19, UN - General Assembly.

Merkel Reinhard, "Novedosas intervenciones del Cerebro. Mejora de la Condición Humana Mental y Límites en el Derecho Penal", Revista de Derecho Penal. Culpabilidad: Nuevas Tendencias I, No. 2, 2012 (Buenos Aires: Rubinzal Culzoni, 2013).

Alexy, Robert, El concepto y la Validez del Derecho, (Barcelona: Gedisa, 2004).

Alexy, Robert, Teoría de la Argumentación Jurídica. La teoría del discurso racional como teoría de la fundamentación jurídica (Lima: Palestra, 2010).

Dworkin, Ronald, Los Derechos en serio, (Barcelona: Ariel Derecho, 2010).

Mir Puig, Santiago, "Límites del Normativismo en el Derecho Penal", Imputación Objetivas y Dogmática Penal, (Mérida: Universidad de los Andes, 2005).

Mir Puig, Santiago, Derecho Penal, Parte General, (Barcelona: Ed. Reppertor. 2006).

Stübinger, Stephan "¿Persona o Paciente? Comentarios sobre el Principio de Culpabilidad en el Derecho Penal desde el punto de vista de la investigación del cerebro" in Revista de Derecho Penal. Culpabilidad: Nuevas Tendencias I, No. 2, 2012).

Hawking, Stephen, Breves respuestas a las grandes preguntas (Buenos Aires, Crítica, 2019).

Hörnle, Tatjana, Determinación de la Pena y Culpabilidad, (Buenos Aires: FD, 2003).

Vormbaum, Thomas, Historia Moderna del Derecho Penal Alemán (Valencia: Tirant Lo Blanch, 2018).

UNICRI- INTERPOL, Artificial Intelligence and Robotics for Law Enforcement, report launched on April 2, 2019 at the High-Level Meeting: Artificial Intelligence and Robotics-Reshaping the Future of Crime, Terrorism and Security.

Hassemer, Winfred, "El principio de proporcionalidad como límite de las intervenciones jurídico penales" in Límites al Derecho penal. Principios operativos en la fundamentación del castigo, (Barcelona: Atelier, 2012).

Hassemer, Winfried, "Neurociencias y culpabilidad en Derecho penal", Indret: Revista para el Análisis del Derecho, No. 2, 2011.

Hassemer, Winfried, Crítica al Derecho Penal de Hoy, (Buenos Aires: Editorial Ad.Hoc., 1995).

Hassemer, Winfried and Muñoz Conde, Francisco, Introducción a la Criminología y al Derecho Penal (Valencia: Tirant Lo Blanch, 1989).

Hassemer, Winfried, Fundamentos del Derecho Penal, (Barcelona: Bosch, 1984).

2. FAIRNESS, TRUST AND FACIAL RECOGNITION TECHNOLOGY IN POLICING

Emily Johnson, Antoni Napieralski, Ziga Skorjanc*

Abstract

The global proliferation and use of facial recognition technologies by law enforcement has justifiably triggered mixed feelings from the public to whom the technology surveils. This contribution argues that the lack of trust results from an absence of regulation. As demonstrated in this paper, the trust is inherently connected to the European Union (EU) data protection fairness principle as the essence of fairness is trust. To compensate for the lack of trust, we suggest the utilisation of the fairness principle as a policy guideline in the regulation of facial recognition technology. This contribution presents the current ambivalent attitude of the public towards police use of facial recognition technologies globally. Following on, the principle of 'fairness' is deconstructed and then applied as a potential aid in fostering trust in facial recognition technologies and new technologies generally.

Keywords: Trust, facial recognition, principles, data protection.

Introduction

The deployment of facial recognition technology (FRT) is on the rise globally and is expected to rise in market value from $3.2 billion in 2019 to $7 billion in 2024.[1] With this seemingly inevitable increase in FRT comes the inescapable concern of the misuse of facial data. Concerns are unsurprisingly associated with the unique and sensitive nature of face data along with the ease of use paired with the potential to misuse the data.

Unsurprisingly retaliations against FRT have ensued. Anti-facial recognition makeup, camouflage, hairstyles,[2] masks[3] and glasses[4] have all materialised as a reaction to the threat of public exertion of FRT. There understandably exist reservations about the legal parameters (or lack thereof) of the use of FRT. In regard to the use of AI-powered technologies, it has been noted that "[a]ny tendency to put blind faith in what in effect remains largely untrusted technology can lead to misleading and sometimes dangerous conclusions."[5] The same premise applies to FRT. With FRT, as a technology carrying an increased risk of mass surveillance, trust stems from the reliability and accuracy of the technology itself as well as the responsible, fair and transparent deployment of it.[6]

1 * Authors are Research Assistants at the Department of Innovation and Digitalisation in Law, University of Vienna, Austria.

'Facial Recognition Market by Software & Services - 2024| MarketsandMarkets'.

2 'CV Dazzle: Camouflage from Face Detection'.

3 Jip van Leeuwenstein, 'Surveillance Exclusion'.

4 Kikuchi, 'Privacy Visor Thwarts Facial-Recognition Tech'.

5 Hurlburt, 'How Much to Trust Artificial Intelligence?'

6 Bradford et al., 'Live Facial Recognition'.

As stated by Jelinek: "Without the guarantee that the protection of personal data is respected, there will be no trust in the technology."[7] Following this approach, as this contribution argues, to foster trust in technologies as controversial as FRT, regulation must be present and enforceable in favour of protecting citizens' fundamental rights and freedoms. Currently there is little regulation in the EU or internationally focusing explicitly on FRT. Data protection rules and principles in the EU do however provide a basis for the appropriate direction of future legislation. Nevertheless, it is not our intention to analyse the challenges of regulation enforcement. Rather, this chapter is focused on delivering guidelines that would foster design of fair regulation of FRT.

The following contribution examines the lack of societal trust[8] presented by FRTs and how this lack of trust could be mitigated by implementing some of the longstanding principles of data protection. This contribution will not simply assess the relevant data protection principles as the black-letter law, but rather as broader policy guidelines for regulating new and potentially invasive technologies. To start, this contribution presents the current level of trust and attitudes towards FRT in ranging international settings. Following on from this, this contribution examines how an overarching law that derives its contents from fundamental principles in data protection and privacy could bolster trust in data protection. To assess this point, this contribution analyses instances where entrenched legal and social principles have serviced as a source of law in the past, and then the subsequent international and global impact associated with this form of regulating. This contribution then focuses more specifically on the principle of lawfulness, fairness and transparency and also where and why this principle could help improve trust levels in FRTs. This contribution ultimately argues that regulation can be a source of fairness and trust and a lack of regulation significantly contributes to a lack of trust for FRT.

What about Trust?

With FRT being increasingly used throughout the world, the following section will examine citizens' attitudes towards FRT in a collection of locations.

A recent European Commission White Paper on AI, excellence and trust in Europe asserts that the deployment of facial recognition in public places threatens people's dignity through possible interferences with the right to respect for private life and protection of personal data. In the context of law enforcement, a strict application of the necessity and proportionality principles must be adhered to, as must the principle of authorisation by EU or national law paired with the appropriate safeguards.[9]

The Ada Lovelace Institute, an independent research body working to ensure that data and AI work for people in society, conducted the first national survey on the public opinion of FRT in the UK.[10] The study concluded that "[t]here is no unconditional support for police to deploy facial recognition technology"[11] to the extent that "55% of people think the government should limit police use of facial recognition to specific circumstances".[12] One third of research participants also stated that they "feel uncomfortable being presented with a scenario of police use of facial recognition technology".[13] Reasons given for the discomfort towards police use of FRT ranged from concerns about privacy, normalisation of surveillance technologies, lack of options to opt out or to consent and general lack of trust in police use of FRT ethically.[14] These public opinions are not unfounded. An independent review of the MET Police use of FRT found that there was an inadequate legal basis for the use of the technology and a failure to satisfy the necessity requirement, and therefore a potential violation under fundamental rights law.[15]

When examining the USA, a 2019 study by the Pew Research Center[16] found that the majority of Americans surveyed "trust law enforcement to use facial recognition responsibly, but the public is less trusting

7 'EDPB on Twitter'.
8 'British Public Want Restrictions on the Use of Facial Recognition Technology'; 'Beyond Face Value'.
9 'White Paper on Artificial Intelligence'; Charter of Fundamental Rights of the European Union.
10 'Beyond Face Value'.
11 Ibid., 2.
12 Ibid., 10.
13 Ibid., 11.
14 Ibid.
15 Fussey and Murray, 'Independent Report on the London Metropolitan Police Service's Trial of Live Facial Recognition Technology'.
16 Smith, 'More Than Half of U.S. Adults Trust Law Enforcement to Use Facial Recognition Responsibly'.

of advertisers and technology companies."[17] 59% of those surveyed voted that it was acceptable for law enforcement to use facial recognition technology when assessing security threats in public spaces. In comparison, only 15% said that it was acceptable for advertisers to use the technology in public spaces to monitor responses to public ad displays. Interestingly, the study also found disparities among demographic groups with a comparatively smaller portion of young adults trusting facial recognition by law enforcement in public spaces. While this study demonstrates general support and trust for law enforcement use of FRT, 40% of those surveyed do not trust law enforcement to use FRT responsibly.

In Germany, plans by the Interior Minister, Horst Seehofer, have materialised with the aim of using FRT widely at airports and railway stations as a move to amend the Federal Police Act to improve police technical capabilities and extend their responsibilities.[18] Responses by activist groups included a blanket ban on the FRT in public spaces by the state.[19]

In the wake of COVID-19, Chinese companies have begun releasing enhanced FRT that can "detect elevated temperatures in a crowd or flag citizens not wearing a face mask."[20] This facial recognition information is then fed to apps in order to alert users in the area of the personal health information of others to establish whether they have been within the proximity of infected patients. While there is no verifiable overview of the attitudes of the Chinese citizens when it comes to their general feelings of trust, individual citizens have expressed their concerns that the government is finding more reasons to surveil their citizens in public.[21] Public acts in retaliation against FRT also demonstrates a lack of trust and even resent towards FRT by the government. In Hong Kong, protestors destroyed facial recognition towers in protest believing them host Chinese FRT.[22]

While portions of the societies discussed to harbour trust for police employment of FRT, the societal trust remains incomplete and the concerns numerous. Given that FRT is a surveillance technology, it has been a viewed as inherently being a source of mistrust and an invasion of privacy. As a means of enforcing trust and sufficiently regulating those behind the FRT, this paper looks to the principle of fairness in data protection so as to provide some contribution to the amendment of the concerns and their sources and the ultimate encouragement of trust.

Legal Principles: Response to the Lack of Trust

As the challenges posed by the lack of societal trust in FRT are global in character, it is necessary to take an international perspective when seeking a possible solution. This contribution argues that the fundamental principles of data protection law have the potential of becoming global policy-guidelines for regulating new technologies. While each of the six principles of data protection are undoubtedly essential in protecting the rights and interests of the data subject, this paper argues that the principle of 'fairness' as set out in Article 5(1)(a) GDPR[23] and Article 4(1)(a) Directive 2016/680[24], may in particular have a broader significance than simply regulating the processing of personal data in the EU.

As this paper argues, the principle of fairness contributes to building trust between the controller and the data subjects, by obliging the (future) controllers to design and carry out the processing with the interests of their own accountability and responsibility to data subjects. If this principle is employed as a part of a policy-guideline (as opposed to a legal norm), it could bear significant effects globally in favour of better use of new technologies and thus more public trust. However, first, one needs to establish what the exact meaning of the fairness principle is. In order to do so, the first step will be distinguishing it from other data protection principles.

17 Ibid.
18 Grüll, 'Germany's Plans for Automatic Facial Recognition Meet Fierce Criticism'.
19 Ibid.
20 Kuo, '"The New Normal"'.
21 Ibid.
22 Doffman, 'Hong Kong Exposes Both Sides Of China's Relentless Facial Recognition Machine'.
23 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) (Text with EEA relevance).
24 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.

Distinguishing fairness from other data protection principles

The fairness principle is commonly referred to as part of a triad of principles. Article 5(1)(a) GDPR places the principles of lawfulness, fairness and transparency in one shared unit. This combined approach creates doubts as to what extent the fairness principle is separate from its conjunctive principles. However, as argued by the scholarship: "a systematic and teleological analysis of Article 5(1)(a) GDPR dictates that this provision is to be viewed as presenting three distinct but overlapping principles."[25] This, in our view, is a convincing interpretation and can be supported by conclusions of the following analysis.

Clifford and Ausloos argue that the principles of fairness have always been closely linked.[26] Going one step further, Kranenborg points out that transparency is an element of fair processing.[27] Those arguments can be traced back to the wording of Article 5(1)(a) GDPR where the fairness principle is accompanied by the lawfulness and transparency principles.

However, it has been pointed out in the scholarship that seeing transparency duties of the controller as the core of the fairness principle, results from the lack of a clear transparency principle in the Directive 95/46/EC.[28] Therefore, the Court of Justice of the European Union (hereinafter CJEU) needed to interpret transparency duties of the controller from the fairness principle.[29] Nevertheless, in the GDPR, transparency has found its way to become a standalone principle (or at least entitled to a separate mention in the legislative text).[30] As such, fairness can (and should) in turn be treated as more than just the requirement of informing the data subject about the modalities of data processing affecting them. The decisions of the lawmakers to introduce both transparency and fairness in the legislative text opposes the interpretation of fairness as transparency.

Meaning of the fairness principle

Having resolved the most confusing overlap between the fairness and transparency principles, this section scrutinises the meaning of the fairness principle. Although it is clear that fairness has been applied and interpreted throughout a plethora of research areas (ranging from political or legal philosophy up to management studies),[31] this contribution follows the interpretations of fairness available in the data protection scholarship.

While it has been stated that "fairness supposes a moral compass, a sense of the just society",[32] the exact definition and meaning of the fairness principle remains one of the most enigmatic in European data protection law. Within the EU the principle can be found in the Charter of Fundamental Rights of the European Union (hereinafter Charter), in the GDPR and in the Directive 2016/680[33] (hereinafter 'Law Enforcement Directive' or 'LED'). However, the exact meaning of the fairness principle (partially due to the broad meaning of the word 'fair') is, however, open to different interpretations.

To begin with, it is important to note that 'fairness', as an autonomous concept of EU law, cannot be interpreted by a reference to the meaning attributed to fairness in national legal systems.[34] Rather, for the sake of a unified application throughout the EU, it is necessary to interpret the term in relation to the overarching system of EU law. Therefore, the first step should be to discuss the literal meaning of the term.

25 Clifford and Ausloos, 'Data Protection and the Role of Fairness', 159.
26 Ibid., 138.
27 Kranenborg, 'Article 8 - Protection of Personal Data', 254.
28 Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data.
29 Roßnagel, 'Artikel 5', para. 45.
30 Clifford and Ausloos, 'Data Protection and the Role of Fairness'.
31 Rawls, 'Justice as Fairness'; Hart, 'Are There Any Natural Rights?'; Klosko, 'The Principle of Fairness and Political Obligation'; Phillips, 'Stakeholder Theory and A Principle of Fairness'; Van Buren, 'If Fairness Is the Problem, Is Consent the Solution?
32 Franck, *Fairness in International Law and Institutions*, 7–8.
33 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.
34 Kramer, 'Art. 5', para. 12; Roßnagel, 'Artikel 5', para. 46; Herbst, 'Art. 5 DS-GVO', para. 13.

As defined by the Collins English Dictionary "[f]airness is the quality of being reasonable, right and just."[35] The multi-component literal definition of fairness has been further developed by the data protection scholarship. Following the comparative linguistic analysis of the term 'fairness' and exploring broader national and cultural interpretations in different linguistic versions of the GDPR, Malgieri identified three key meanings to which the various translations of 'fairness' refer: loyal, equitable and correct.[36] One can note a significant discrepancy between the meaning of 'fairness' as defined by the English dictionary and as used in the GDPR and translated throughout Europe. The only overlap between the two is the aspect of being right/equitable. Or in simple terms: of doing the right thing. The concept of what is fair/right/equitable is however prone to change in the course of time.[37] As such, 'fairness' is evidently a subjective concept, but one made up of similar and overlapping interpretations resulting from the translation and interpretation of language and no doubt cultural and legal developments often dependent on each national environment.

Following Malgieri, the essence of the fairness principle is the "substantial mitigation of unfair imbalances that create situations of 'vulnerability'".[38] Fairness is not so much centred upon meeting the procedural requirements of the GDPR (such as transparency or lawfulness), but rather touches upon the core of the power relation between the controller and the data subject.[39] A similar point is made by Clifford & Ausloos, who argue that the fairness principle ought to "re-balance the asymmetric data subject-controller relationship."[40] This argument suggests that fairness acts to help "prevent unfair imbalances between data subjects and data controllers."[41] In the sense that fairness assists in maintaining balance between the data subject and controller, fairness has the role of an overarching yet standalone principle in the GDPR; it is a lens through which data processing activities ought to be assessed.[42]

This contribution builds upon a distinction made by Clifford & Ausloos between explicit and implicit meaning of the fairness principle in data protection.[43] As they argue, the fairness principle implicitly protects the data subjects from controllers' abuse.[44] This correlates with the explicit meaning of the fairness principle which aims at preventing any deception of data subjects as to the "nature and purposes of the processing operations."[45] Additionally, the observations made by Malgieri allow for a departure from the narrow reading of the data protection legislation thus expanding it to the broader challenges of technology regulation.

In following Herbst, Roßnagel or Malgieri this contribution argues that mere compliance with the GDPR and other legal provisions might not always suffice to prevent a data processing operation from being qualified as being in breach of the fairness principle.[46] Rather, it is the general (im)balance of power between the data subject and the controller that requires thorough assessment in light of the fairness principle.[47] Should the data subject be disadvantaged by data processing in any way through the lack of the balance of power (as deemed necessary by the GDPR), even if no other norm is infringed, the processing shall be qualified as a breach of the fairness principle and thus unlawful.[48]

Based on the aforementioned arguments, we want to go one step further and employ the fairness principle as a policy-regulating standard on the global law enforcement scene. As rightly pointed out by Roßnagel, unfair conduct is simply a conduct in violation of the data subject's trust.[49] Therefore, we seek to introduce the fairness principle as a globally recognised method of mitigating the lack of trust in policing. In doing this, we argue that the presence of the fairness principle (as understood on the grounds of the GDPR) in regulation encourages trust. By minimising the (im)balance of power between the controller and the data subject, obedience to the fairness principle allows to mitigate risks arising from data processing as controversial as the use of FRT. Its cross-sectional nature allows to assess data processing from a broader perspective and ensures that also other principles of data protection are being followed.

35 'Fairness Definition and Meaning | Collins English Dictionary'.
36 Malgieri, 'The Concept of Fairness in the GDPR', 27–30.
37 Bygrave, *Data Protection Law*, 58.
38 Malgieri, 'The Concept of Fairness in the GDPR', 2.
39 Ibid.
40 Clifford and Ausloos, 'Data Protection and the Role of Fairness', 159.
41 Malgieri, 'The Concept of Fairness in the GDPR', 2.
42 Clifford and Ausloos, 'Data Protection and the Role of Fairness', 159.
43 Ibid., 138.
44 Ibid., 140.
45 Ibid.; see also Bygrave, *Data Protection Law*, 58.
46 Herbst, 'Art. 5 DS-GVO', para. 47.
47 Ibid.
48 Ibid.
49 Roßnagel, 'Artikel 5', para. 47.

One last piece of the analysis that is necessary in this context is the global conditions of the fairness principle as it stands now. Thus, the next section analyses the place fairness has in some key global and national instruments and approaches to technology regulation.

Fairness as a global standard

In international law, fairness can be viewed as "the rubric under which [...] tension is discursively managed."[50] An example of this international standard of fairness can be seen in the expression of the UN Human Rights Council Annual Report on 'The right to privacy in the digital age' which states that fairness is a principle component making up the minimum standards for the processing of personal data.[51] A further example of fairness as a global standard is Article 5 of the Council of Europe Convention 108 which states that "personal data undergoing automatic processing shall be: a) obtained and processed fairly and lawfully."[52] Moreover, regarding the modernization of Convention 108, Article 5 is expanded upon regarding the "[l]egitimacy of data processing and quality of data" where it states that in processing personal data there must be a fair balance between the interests of those concerned at all stages of processing.[53] The insertion of the requirement that there must be a fair balance here suggests that there is an increasing acknowledgement of the necessity of the inclusion and exercise of the fairness principle in data protection at an international level.

The Organisation for Economic Cooperation and Development (OECD) asserts that on the basis of the 'collection limitation principle', there should be limits to the collection of personal data, requiring that data be obtained by lawful and fair means.[54] The OECD does not however elaborate on the meaning of 'fair' in the context of data protection.

Taking a more national view, within the UK, the ICO highlights three requirements when assessing fairness of data processing:

1. How may the processing affect the individuals concerned, and can any adverse impacts be justified?
2. All processing is in line with the data subject's expectations. Any processing which does not meet expectations must be justifiable.
3. Data subjects must not be deceived or misled when data is collected and processed.[55]

These criteria emphasise the effect on the data subject, their expectations and trust.

The Law Society of Scotland expresses that "[i]n order to process personal data fairly, the processing must be in line with the data subject's expectations."[56] This statement returns to the notion of fairness and transparency. The data subject can form expectations and harbour trust if they have transparency of the intentions and application of FRT. If the use of FRT is then employed fairly in line with the expectations of the data subject, then trust for the technology may ensue.

Conclusions

Fairness is a key principle which, if exercised in the context of FRT, offers to mitigate risks associated with accuracy, validity, security, bias, discrimination, privacy and transparency – all of which contribute to the overall (im)balance of power between the controller and the data subject.

Public distrust for FRT is present. While this distrust is undoubtedly due to the nature and characteristics of FRT, which is often seen as invasive and clandestine, a significant element of this distrust is arguably due

50 Franck, *Fairness in International Law and Institutions*, 7.
51 'The Right to Privacy in the Digital Age. Report of the United Nations High Commissioner for Human Rights'.
52 Council of Europe, Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data.
53 Council of Europe, Convention 108+ for the Protection of Individuals with regard to Automatic Processing of Personal Data.
54 OECD, Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data.
55 ICO, 'Principle (a)'.
56 'Fair and Transparent Processing'.

to the lack of regulation and the subsequent regulated application of FRT. From analysing the principles of 'fairness' in data protection and applying it to the use of FRT, we find an extension of the quality of law test in the sense that for the 'fair' deployment of FRT and processing of the resulting biometric data, those being surveilled may not be necessarily required to be informed about single cases and investigations. This may be practically and operationally impossible for law enforcement, but rather the public should be aware of the technology in the first instance. In this case, there should be an awareness of the types of technology, who is using it, where it is being used, how the data is being collected, processed and stored. Further, the rules about sharing the data with third parties, or using it for further purposes should also be clear and transparent. Ultimately, the rules on the deployment of FRT should be clearly stipulated. In applying the fairness principle in this context, data subjects must also retain a means of challenging FRT and/or being able to rely on a supervisory authority. This approach contributes to fairness and simultaneously lawfulness and transparency.

An outcome that would certainly be welcome would be clear regulation resulting in accountability of the controllers for the processing with the use of FRT; in consequence this would lead to higher levels of trust on the part of the public. From this perspective, this paper merely marks the first step in fostering public trust in new technologies by law enforcement. Future research would need to examine a test for the assessment of fairness (or lack thereof) anytime new and innovative technologies aimed at assisting law enforcement are being deployed.

References

'Beyond Face Value: Public Attitudes to Facial Recognition Technology'. Accessed 14 March 2020. https://www.adalovelaceinstitute.org/beyond-face-value-public-attitudes-to-facial-recognition-technology/.

Bradford, Ben, Julia Yesberg, Jonathan Jackson, and Paul Dawson. 'Live Facial Recognition: Trust and Legitimacy as Predictors of Public Support for Police Use of New Technology'. Preprint. SocArXiv, 10 January 2020. https://doi.org/10.31235/osf.io/n3pwa.

Nuffield Foundation. 'British Public Want Restrictions on the Use of Facial Recognition Technology', 9 September 2019. https://www.nuffieldfoundation.org/news/british-public-want-restrictions-on-the-use-of-facial-recognition-technology.

Bygrave, Lee A. *Data Protection Law: Approaching Its Rationale, Logic, and Limits*. Information Law Series 10. The Hague ; New York: Kluwer Law International, 2002.

Charter of Fundamental Rights of the European Union, 326 OJ C § (2012).

Clifford, Damian, and Jef Ausloos. 'Data Protection and the Role of Fairness'. *Yearbook of European Law* 37 (1 January 2018): 130–87. https://doi.org/10.1093/yel/yey004.

Council of Europe. Convention 108+ for the Protection of Individuals with regard to Automatic Processing of Personal Data (2018).

———. Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, Pub. L. No. European Treaty Series – No. 108 (1981).

'CV Dazzle: Camouflage from Face Detection'. Accessed 14 March 2020. https://cvdazzle.com/.

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, Pub. L. No. OJ L 119 p. 89–131. Accessed 12 August 2019. https://eur-lex.europa.eu/eli/dir/2016/680/oj.

Doffman, Zak. 'Hong Kong Exposes Both Sides Of China's Relentless Facial Recognition Machine'. Forbes. Accessed 14 March 2020. https://www.forbes.com/sites/zakdoffman/2019/08/26/hong-kong-exposes-both-sides-of-chinas-relentless-facial-recognition-machine/.

Twitter. 'EDPB on Twitter: „EDPB Chair to @EP_Justice : "Whenever personal data is processed in the context of fighting COVID-19, data protection rules are indispensable. Without the guarantee that the protection of personal data is respected, there will be no trust in the technology." https://t.co/f7HXX0TpMz" / Twitter'. Accessed 16 May 2020. https://twitter.com/EU_EDPB/status/1258428969965281280.

'Facial Recognition Market by Software & Services - 2024| MarketsandMarkets'. Accessed 14 March 2020. https://www.marketsandmarkets.com/Market-Reports/facial-recognition-market-995.html.

Law Society of Scotland. 'Fair and Transparent Processing'. Accessed 14 March 2020. https://www.lawscot.org.uk/members/business-support/gdpr-general-data-protection-regulation/gdpr-guide/create-a-record-of-data-processing/fair-and-transparent-processing/.

'Fairness Definition and Meaning | Collins English Dictionary'. Accessed 8 March 2020. https://www.collinsdictionary.com/dictionary/english/fairness.

Franck, Thomas M. *Fairness in International Law and Institutions*. Reprinted. Oxford: Oxford Univ. Press, 2002.

Fussey, Peter, and Daragh Murray. 'Independent Report on the London Metropolitan Police Service's Trial of Live Facial Recognition Technology', 2019. http://repository.essex.ac.uk/24946/1/London-Met-Police-Trial-of-Facial-Recognition-Tech-Report-2.pdf.

Grüll, Philipp. 'Germany's Plans for Automatic Facial Recognition Meet Fierce Criticism'. *Www.Euractiv.Com* (blog), 10 January 2020. https://www.euractiv.com/section/data-protection/news/german-ministers-plan-to-expand-automatic-facial-recognition-meets-fierce-criticism/.

Hart, H. L. A. 'Are There Any Natural Rights?' *The Philosophical Review* 64, no. 2 (1955): 175–91. https://doi.org/10.2307/2182586.

Herbst, Tobias. 'Art. 5 DS-GVO'. In *Datenschutz-Grundverordnung/BDSG. Kommentar*, edited by Jürgen Kühling and Benedikt Buchner, 2nd ed., 2018.

Hurlburt, George. 'How Much to Trust Artificial Intelligence?' *IT Professional* 19, no. 4 (2017): 7–11. https://doi.org/10.1109/MITP.2017.3051326.

ICO. 'Principle (a): Lawfulness, Fairness and Transparency', 6 November 2019. https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/principles/lawfulness-fairness-and-transparency/.

Jip van Leeuwenstein. 'Surveillance Exclusion'. Accessed 14 March 2020. http://www.jipvanleeuwenstein.nl/.

Kikuchi, Daisuke. 'Privacy Visor Thwarts Facial-Recognition Tech'. The Japan Times, 13 May 2016. https://www.japantimes.co.jp/news/2016/05/13/national/privacy-glasses-thwart-face-recognition-tech/.

Klosko, George. 'The Principle of Fairness and Political Obligation'. *Ethics* 97, no. 2 (1 January 1987): 353–62. https://doi.org/10.1086/292843.

Kramer, Philipp. 'Art. 5'. In *Datenschutz-Grundverordnung. Bundesdatenschutzgesetz Und Nebengesetze. Kommentar*, edited by Martin Eßer, Philipp Kramer, and Kai von Lewinski, 6th ed., 2018.

Kranenborg, Herke. 'Article 8 - Protection of Personal Data'. In *The EU Charter of Fundamental Rights: A Commentary*, edited by Steve Peers, Tamara Hervey, Jeff Kenner, and Angela Ward, 01 ed., 223–66, 2014.

Kuo, Lily. '"The New Normal": China's Excessive Coronavirus Public Monitoring Could Be Here to Stay'. *The Guardian*, 9 March 2020, sec. World news. https://www.theguardian.com/world/2020/mar/09/the-new-normal-chinas-excessive-coronavirus-public-monitoring-could-be-here-to-stay.

Malgieri, Gianclaudio. 'The Concept of Fairness in the GDPR'. *Proceedings of FAT* '20, January 27–30, 2020*, 2020. https://doi.org/10.1145/3351095.3372868.

OECD. Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data (2013).

Phillips, Robert A. 'Stakeholder Theory and A Principle of Fairness'. *Business Ethics Quarterly* 7, no. 1 (January 1997): 51–66. https://doi.org/10.2307/3857232.

Rawls, John. 'Justice as Fairness'. *The Philosophical Review* 67, no. 2 (1958): 164–94. https://doi.org/10.2307/2182612.

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) (Text with EEA relevance), Pub. L. No. OJ L 119, pp. 1–88. Accessed 12 August 2019. https://eur-lex.europa.eu/eli/reg/2016/679/oj?eliuri=eli%3Areg%3A2016%3A679%3Aoj.

Roßnagel, Alexander. 'Artikel 5'. In *Datenschutzrecht. DSGVO Mit BDSG*, edited by Spiros Simitis, Gerrit Hornung, and Indra Spiecker, 363–99. Nomos, 2019.

Smith, Aaron. 'More Than Half of U.S. Adults Trust Law Enforcement to Use Facial Recognition Responsibly'. Pew Research Center, n.d. https://www.pewresearch.org/internet/wp-content/uploads/sites/9/2019/09/09.05.19.facial_recognition_FULLREPORT_update.pdf.

'The Right to Privacy in the Digital Age. Report of the United Nations High Commissioner for Human Rights'. Human Rights Council, 2018. https://daccess-ods.un.org/TMP/1219208.61303806.html.

Van Buren, Harry J. 'If Fairness Is the Problem, Is Consent the Solution?Integrating Isct and Stakeholder Theory.' *Academy of Management Proceedings* 1999, no. 1 (1 August 1999): C1–6. https://doi.org/10.5465/apbpp.1999.27630644.

European Commission. 'White Paper on Artificial Intelligence: A European Approach to Excellence and Trust'. Text. Accessed 14 March 2020. https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en.

3. PURPOSE LIMITATION BY DESIGN AS A COUNTER TO FUNCTION CREEP AND SYSTEM INSECURITY IN POLICE AI

Ivo Emanuilov*, Stefano Fantin**, Thomas Marquenie***, Plixavra Vogiatzolgou****

Abstract

AI's dual nature makes it both a threat and a means to protect human rights and information technology systems. Amongst others, issues pertaining to the opacity and inclusion of potential biases in algorithmic processes as well as the inherent security vulnerabilities of such applications, unveil a tension between such technological pitfalls and the aptness of current regulatory frameworks. As a consequence, normative concepts might need to be reconsidered as to support the development of fair AI. This paper reflects on the importance of the purpose limitation principle and its role in the design phase, to mitigate the adverse impact of AI on human rights and the security of information systems.

To define, elaborate, and 'manufacture' the purpose for which AI is deployed is critical for mitigating the intrusive impact on human rights. However, the inevitable uncertainty in the formulation of these objectives may lead to scenarios where machines do what we ask them to do, but not necessarily what we intend. Moreover, the continuous development of a system's capabilities may allow for uses far beyond the scope of its originally envisaged deployment and purpose.

In an AI context, the deployment of AI beyond its originally specified, explicit and legitimate purposes can lead to function creep as well as exacerbate security incidents. For example, AI systems intended for specific crime prevention goals might gradually be repurposed for unwarranted surveillance activities not originally considered. Furthermore, the lack of a defined purpose in combination with the inherent security vulnerabilities of AI technology draw into question the suitability of using machine learning tools in complex information technology systems.

In data protection law, the principle of purpose limitation requires the purposes for which data is processed to be specified, and subsequent use limited thereto (OECD, 1981). This paper seeks to determine whether this principle can address the consequences of function creep by exploring the use cases of predictive policing and information systems security.

It is argued that, although this core principle can improve the security of AI systems and their better alignment with human rights, it currently often fails to do so. We propose that a more incisive assessment of the envisioned purposes should take place during the design phase to improve the security of AI systems and their better alignment with human rights.

Keywords: Artificial intelligence, purpose limitation, data protection, predictive policing, cybersecurity

Introduction

Machine, or artificial, intelligence (AI) is often defined as machines' capability to act in a way that helps *them* achieve *their* objectives.[1] Essentially, these optimisation machines embed objectives and desiderata set out by their designers and users. However, with time, AI, whether it is the system itself or its intended use, may exceed the initial purposes envisioned during the design process. In the strive for optimisation of societal security, the misalignment of objectives and the deployment of AI beyond its originally specified, explicit and legitimate purposes may create a function creep, potentially threatening human rights.[2]

As early as 1960, Norbert Wiener voiced the concern that "[i]f we use, to achieve our purposes, a mechanical agency with whose operation we cannot interfere effectively (...) we had better be quite sure that the purpose put into the machine is the purpose which we really desire."[3] To define, elaborate, and 'manufacture' the purpose for which AI is deployed is, therefore, critical for mitigating the intrusive impact on human rights. However, the inevitable uncertainty in the formulation of these objectives may lead to scenarios where machines do what we ask them to do, but not necessarily what we intend.[4] Moreover, the continuous development of a system's capabilities may allow for uses far beyond the scope of its originally envisaged deployment and purpose. For example, AI systems intended for specific crime prevention goals might gradually be repurposed for unwarranted surveillance activities not considered initially. Furthermore, the lack of a defined purpose coupled with the inherent security vulnerabilities of AI question the suitability of using data-driven tools in complex information technology systems.

Against this backdrop, the paper argues that the current interpretation of the legal principle of purpose limitation, whose origins lie in data protection law, falls short of being able to address the ensuing risks of adverse impact on human rights. We suggest that a more incisive assessment of the envisioned purposes should take place during the design phase to improve the security of AI systems and their better alignment with human rights.

The paper is structured in three main parts. The first section provides a genealogical overview of the principle of purpose limitation in data protection law and suggests a different approach towards mitigating the negative impact of function creep. The second and third sections analyse the problem of function creep in the context of two specific cases: predictive policing and information security in the law enforcement sector.

The function of purpose in AI

Function creep is colloquially referred to as the expansion of the intended use of technology to a different use, bringing with it a series of unintended and uncontrolled consequences.[5] A recent definition suggested by Koops describes it as "denoting an imperceptibly transformative and therewith contestable change in a data-processing system's proper activity."[6] In the context of AI, we conceptualize function creep as pertaining to three main groups of cases. The first concerns the problem of technology performing tasks it has been

1 Parts of this work have been performed under the Cybersecurity Initiative Flanders – Strategic Research Program (CIF), and the H2020 786629 project MAGNETO, which has received funding from the European Union's Horizon 2020 Program

* Doctoral Researchers at the KU Leuven Centre for IT and IP Law. All authors contributed equally to the drafting and research of this paper.

Stuart Russell, *Human Compatible: Artificial Intelligence and the Problem of Control* (Penguin Publishing Group, 2019), 11.

2 Such as, for example, the 'radicalisation' of the Microsoft Tay chatbot and the adverse effects of nudging click-through content selection algorithms. Another example is the controversial deployment of risk assessment tools in criminal justice, such as the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) systems used in the US. See 'Criminal Law - Sentencing Guidelines - Wisconsin Supreme Court Requires Warning before Use of Algorithmic Risk Assessments in Sentencing - State v. Loomis 881 N.W.2d 749 (Wis. 2016) Recent Cases', *Harvard Law Review* 130, no. 5 (2016–2017): 1530–37.

3 Norbert Wiener, "Some Moral and Technical Consequences of Automation," *Science* 131, no. 3410 (5 June 1960): 1358, https://doi.org/10/ftpxdb.

4 Russell, Human Compatible, 12.

5 The problem of function creep has been studied extensively in the field of surveillance studies, data protection and privacy law. See, inter alia, Langdon Winner, *Autonomous Technology: Technics-out-of-Control as a Theme in Political Thought* (Cambridge, Mass.: MIT Press, 1977), 28 and David Lyon, *Surveillance Studies: An Overview* (Polity, 2007), 201.

6 Bert-Jaap Koops, "The Concept of Function Creep," *Law, Innovation and Technology* 13, no. 1, Published ahead of print, 03 March 2020, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3547903.

designed to perform but whose outcome does not reflect the intended purpose of its designers.[7] For example, an AI system may be given a specific objective to perform an optimisation task, but the way this purpose is specified may trigger an optimisation process which, ultimately, fails to achieve the system designer's intended purpose. The second refers to the problem of deploying a technology intended for a particular purpose to a different use or context, which may lead to unintended consequences.[8] The third concerns the scenario where it is the user, equipped with new capabilities, that goes on to use the technology for a new purpose. This is the case where the user's objectives are intertwined with the machine's optimisation objective. In other words, the system's goal of maximising expected utility by offering "hypotheses about the future, based on all its past experience"[9] nudges the user towards deploying the technology for a different purpose.

Arguably, each of these forms of function creep in the context of AI may adversely affect the exercising and protection of the full spectrum of human rights.[10] In order to counter these impacts, policymakers have turned their eyes to legal tools, chief among them being the so-called principle of purpose limitation. Its operation in practice, however, has been plagued by the concept's elusiveness and contextual dependency.

The purpose limitation principle, more specifically, is rooted in the Organisation for Economic Co-operation and Development's guidelines on privacy and in the Convention 108, and is considered to comprise two sub-principles, the purpose specification and the compatible use limitation or compatibility principles.[11] Similarly stipulated in EU law, the purpose limitation principle guarantees that personal data must be processed for predefined specified purposes and not further processed in a manner that is incompatible with those purposes, unless under exceptional circumstances.[12] The principle, embedded in theories of informational self-determination, allows for the individual to obtain control over their personal data,[13] albeit individual control in the law enforcement environment has been arguably contested as unrealistic.[14] The purpose limitation principle is said to offer a balanced approach between the reasonable expectations of individuals, by enhancing trust and legal certainty, and certain competing interests, by acknowledging the pragmatic need for further processing in specific occasions.[15]

On the one hand, the purpose specification principle focuses on the clear delineation of the purposes for which personal data are initially collected. The compatibility principle, on the other hand, aims at preventing further processing of personal data for purposes that are incompatible to the purpose for which the data were originally collected. Compatibility may be assessed through a set of factors including the link between purposes, the context, the nature of data, the possible consequences and the existence of safeguards.[16] Exemptions from the compatibility test are foreseen *to safeguard [inter alia] national security, public security*

7 This case concerns an inherent problem with the optimisation performance of a system. For example, this may be the result of over optimisation in facial recognition software, leading to unintentional racial discrimination. See Fumio Shimpo, "The Principal Japanese AI and Robot Strategy toward Establishing Basic Principles", in *Research Handbook on the Law of Artificial Intelligence* (Edward Elgar Publishing, 2018), 116, https://www.elgaronline.com/view/edcoll/9781786439048/9781786439048.00015.xml.

8 This concerns the problem where decision-makers deploy a technology to a use case it was never intended to solve.

9 Russell, Human Compatible, 37.

10 Lorna McGregor, Daragh Murray, and Vivian Ng, "International Human Rights Law as a Framework for Algorithmic Accountability'," *International & Comparative Law Quarterly* 68, no. 2 (2019): 315–316.

11 OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data, 1981 (updated in 2013), art 9, https://www.oecd.org/internet/ieconomy/oecdguidelinesontheprotectionofprivacyandtransborderflowsofpersonaldata.htm (accessed on 18 Sep. 19); Convention 108, art 5.

12 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), (GDPR), OJ L 119, 4.5.2016, 1–88, art 5.1(b); 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, OJ L 119, 4.5.2016, p. 89–131 (DPLED), art 4.1(b); Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, OJ L 281, 23.11.1995, p. 31–50 (DPD) art 6.1(b).

13 Liana Colonna, "Data Mining and Its Paradoxical Relationship to the Purpose Limitation Principle," in *Reloading Data Protection* Multidisciplinary Insights and Contemporary Challenges, eds. Serge Gutwirth, Ronald Leenes, Paul De Hert (Springer Netherlands, Dordrecht 2014), 299.

14 Catherine Jasserand, "Subsequent Use of GDPR Data for a Law Enforcement Purpose: The Forgotten Principle of Purpose Limitation," *European Data Protection Law Review* 4, no. 2 (2018): 152.

15 Article 29 Data Protection Working Party (WP29), "Opinion 03/2013 on purpose limitation", WP 203, 2 April 2013.

16 GDPR, art 6(4), following Article 29 Data Protection Working Party, Opinion 03/2013. In addition, further processing for historical, statistical or scientific purposes is by law considered not to be incompatible GDPR, art 5.1(b).

and the prevention of criminal offences.[17] In addition, the purpose limitation principle may be restricted overall, insofar as the restriction *respects the essence of the fundamental rights and freedoms and is a necessary and proportionate measure in a democratic society to safeguard* the same objectives.[18]

Purpose limitation is almost identically defined in both the General Data Protection Regulation (GDPR) and the accompanying Directive (EU) 2016/680 (DPLED),[19] which applies to the processing of personal data by competent authorities for the purposes of prevention, investigation, detection, and prosecution of crime. Nevertheless, compatibility is no further explained in the latter, nor are exceptions formulated in the same detailed manner. In particular, the DPLED is focused on the distinction between 'law enforcement purposes' and 'non-law enforcement purposes' rather than articulating how purposes within law enforcement may be different and incompatible. To that end, it has been pointed out that every individual purpose of processing should be detailed, as 'law enforcement *per se*, shall not be considered as one specified, explicit and legitimate purpose,'[20] and two law enforcement purposes should not be *de facto* considered compatible because they belong in the same field.[21] Due to the legal nature of the DPLED, however, the application of the purpose limitation principle will fluctuate depending on national implementing laws.

Finally, the function of purpose within EU human rights case law is similarly elusive. In cases of surveillance, the Court of Justice of the EU, for instance, tends to assimilate the presence of a general purpose of security to the presence of an objective of general interest,[22] despite the enshrinement of purpose specification principle within the Charter of Fundamental Rights of the EU (CFREU), as part of the right to data protection.[23] However, this approach renders the purpose limitation principle, in essence, useless; the examination of the existence of an objective of general interest would have been performed regardless, in accordance with the CFREU, during the assessment of any non-absolute right.[24]

The purpose limitation principle is moreover often overlooked at an operational level without concrete measures taken to enforce it. Contrary to the spirit of the principle described above, the mere existence of a law enforcement objective is often equated to compliance with the purpose limitation principle. A seemingly narrower approach is followed, where each individual part of the principle, *i.e.* purpose specification, compatibility, and its restriction under the condition of necessity and proportionality, are not fully taken into consideration. The lack of such examination facilitates this imperceptible re-purposing of an AI system, which creates room for function creep.

We suggest that a conscious effort to integrate a more in-depth analysis of purpose limitation is necessary at the design stage. A by-design approach could enforce and impose a mindful, transparent and comprehensive assessment of the original purpose, the potential compatible further purposes and the restricted use for incompatible purposes only within the limits of suitability, absence of less intrusive alternative means and strict proportionality[25]. This approach would thus address the function creep and current drawbacks of

17 GDPR, art 6(4). In other words, such law provides for a new legal basis, rendering a compatibility assessment void. This provision had been criticised by the WP29 and the European Data Protection Supervisor (EDPS), for ignoring the fact that establishing a legal basis and respecting the purpose limitation principle consist of two separate and cumulative requirements. While the respective opinions referred to the proposed version of Article 6(4) Draft GDPR, this line of criticism still stands, as the current version, while substantially modified, is narrower in scope without however fully addressing the concern raised. Article 29 Data Protection Working Party, Opinion 03/2013; European Data Protection Supervisor (EDPS), "Opinion of the European Data Protection Supervisor on the data protection reform package", Brussels, 7 March 2012, https://edps.europa.eu/sites/edp/files/publication/12-03-07_edps_reform_package_en.pdf.

18 GDPR, art 23. While the respective provision on restrictions in DPD and in the Draft GDPR explicitly named data quality principles, comprising *inter alia* the purpose limitation principle, to be included within the scope of restriction, the wording of the final text provides for this different and vague structure. It is thus unclear to what extent purpose limitation is considered as corresponding to 'rights or obligations' or is set aside as a fundamental principle. See also Catherine Jasserand, Subsequent Use of GDPR Data for a Law Enforcement Purpose: The Forgotten Principle of Purpose Limitation.

19 GDPR, art 5.1(b) and DPLED art 4.1(b).

20 Article 29 Data Protection Working Party (WP29), 'Opinion 03/2015 on the draft directive on the protection of individuals with regard to the processing of personal data by competent authorities for the purposes of prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and the free movement of such data', WP 233, 01 December 2015.

21 European Data Protection Supervisor (EDPS) 'Opinion on the Data Protection Reform Package, 12 March 2012.

22 See for example Court of Justice of the European Union, C-293/12 Digital Rights Ireland, C-203/15 and C-698/15 Tele2 Sverige / Watson, Opinion 1/15, C-207/16 Ministerio Fiscal.

23 Charter of Fundamental Rights of the European Union (CFREU), OJ C 202, 7.6.2016, 391–407, article 8.

24 CFREU, art 52(1).

25 The principle of proportionality in human rights law entails that the restricting measure must be suitable and appropriate to meet the objective, the least onerous in relation to that objective, and proportionate stricto sensu, i.e. achieving a fair balance. See for example: Jonas Christoffersen, *Fair balance: Proportionality, subsidiarity and the primarity in the European Convention on Human Rights* (Martinus Nijhoff Publishers, 2009).

purpose limitation in the law enforcement context. Moreover, purpose limitation could also serve as a conduit to better understand the acceptable level of AI deployment by focusing on the security risks therein.

System end-users and manufacturers should, therefore, engage in an open dialogue regarding the purpose of the envisioned tools. Prior to deployment, clear limits and restrictions on system processes and functionalities should be included at the software level to prevent the applications from gradually being used for different and broader purposes. The paper will seek to better illustrate this through the lenses of predictive policing practices and police information systems security vulnerabilities.

Predictive Policing

Concept

In general terms, the concept of predictive policing refers to the use of analytical techniques to assess a broad variety of data in order to anticipate potential occurrences of crime.[26] By using statistical models, predictive policing tools can identify the locations, persons and circumstances most likely to be involved in criminal acts.[27] This information can be used to guide police interventions and support the ability of law enforcement agencies to assess crime data for the purpose of acting in a proactive and targeted manner. To this end, predictive policing tools generally fall into two main categories. Predictive mapping or place-oriented techniques aim to anticipate the locations in which crime is most likely to take place at a given time and under certain circumstances.[28] These techniques can identify so-called 'crime hotspots' in order to effectively deploy police resources by analysing police records, crime trends and other variables determined to be relevant to the occurrence of criminal acts. Similarly, predictive identification or person-oriented applications seek to determine which individuals are at higher odds of committing crimes or being victimized thereby.[29] By conducting risk assessments and establishing the circumstances that increase the likelihood of a person being involved in a crime, law enforcement agencies can identify high-risk individuals so as to intervene proactively and steer them away from crimes.

Beyond the use of these techniques by law enforcement agencies and public authorities, predictive policing may also entail the involvement of private entities. A part of the fight against financial crime, money laundering and terrorist financing, for instance, is being outsourced to banks, which have to abide by their EU Anti-Money Laundering (AML) obligations.[30] In trying to comply with the latter, banking institutions often resort to predictive analytics on transactional activities in order to identify potential future criminal behaviour and, if possible, prevent the manifestation of crime.[31] However, while predictive analytics employed by law enforcement authorities feed on mostly criminal data, predictive analytics performed by banking institutions rely primarily on data generated in the course of regular commercial and financial transactions. Still, the underlying logic remains the same, and the algorithms employed must be trained to identify potentially criminal intentions behind transactions.

Furthermore, while the technologies behind predictive policing carry great promise and are often presented as a revolution in the fight against crime[32], the effectiveness and impact of the practice remain contested and

26 Beth Pearsall, "Predictive Policing: The Future of Law Enforcement?," *National Institute of Justice Journal* 266, no. 1 (2010).

27 Walter L. Perry, Brian McInnis, Carter C. Price, Susan C. Smith, John S. Hollywood, *Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations* (RAND Corporation, 2013).

28 *PHRP Expert Meeting On Predictive Policing* (Amnesty International Police and Human Rights Programme, May 2019)

29 Fieke Jansen, *Data Driven Policing in the Context of Europe* (Data Justice Lab Working Paper, May 2018).

30 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, OJ L 156, 19.6.2018, 43–74. For more information on the anti-money laundering and counter terrorist financing policies and legislation, see the official European Commission website https://ec.europa.eu/info/business-economy-euro/banking-and-finance/financial-supervision-and-risk-management/anti-money-laundering-and-counter-terrorist-financing_en.

31 See for example EY Global, "How data analytics is leading the fight against financial crime," EY Global (blog), 6 December 2019. https://www.ey.com/en_be/advisory/how-data-analytics-is-leading-the-fight-against-financial-crime .

32 Joel Rubin, "Stopping crime before it starts," *Los Angeles Times,* August 21, 2010; Zach Friend, "Predictive Policing: Using Technology to Reduce Crime," *FBI Law Enforcement Bulletin*, April 9, 2012.

controversial. Despite the growing deployment of predictive policing tools around the world,[33] the empirical evidence is yet to conclusively validate the claims of consistent and significant reductions in crime.[34] Similarly, serious concerns are frequently raised about the potentially negative impact on human rights, police accountability, and the fairness of law enforcement interventions.[35] It is in the context of this controversy and uncertainty that this article aims to assess the risks of function creep in predictive policing tools that can exacerbate these issues and might necessitate a different approach to purpose limitation to improve the accountability and reliability of these applications.

Function creep and the limits of purpose limitation

Purpose limitation, as it currently stands, does not appear to be an effective safeguard in the context of predictive policing. As these applications are gradually applied for ever broader purposes and the expanding datasets to which the police have access are utilized for any and all relevant analytical processes,[36] purpose limitation seems unable to address the risk of function creep. This is due to the broad mandate law enforcement agencies have to fight crime, the far-reaching grounds for the collection and use of personal data,[37]the subsequent growing amounts of data they can access and store, and the big data capabilities of the tools they use. Moreover, multiple factors, such as the rise of terrorist threats and the digitisation of crime, have led to a growing overlapping of competencies between law enforcement and intelligence services, which enlarges the scope of data and authorities with access to it.[38]

In the context of predictive policing, this manifestation of function creep primarily concerns the risk of analytical tools being used for increasingly broad purposes that blur the lines with mass surveillance practices. While early iterations of predictive policing tools were limited to making modest improvements to existing intelligence-led practices that allowed for the more accurate identification of high-risk crime areas,[39] technological advancements and the growing availability of big data have enabled predictive systems to analyse larger datasets, consider more variables and draw increasingly far-reaching inferences about locations and people.[40] Additionally, the threshold of inclusion in law enforcement databases is often low and the scope of the data analysed is frequently expanded to include external information[41] such as utility bills, delivery orders and license plate recognition footage.[42] As such, the expansive nature of predictive policing methods poses the risk that these tools and the data they use are gradually used and re-purposed in manners that closely resemble mass surveillance, without proper safeguards or regards to the principle of purpose limitation.

In the United States, tools like Operation LASER are used to score individuals on the basis of their circumstances, personal lives and interactions with the criminal justice system in order to target specific profiles for further surveillance,[43] while software like PredPol provides a map of crime predictions to steer officers towards continuously monitoring high-risk communities and neighbourhoods.[44] Similarly, European police agencies

33 Odhran James McCarthy, "AI & Global Governance: Turning the Tide on Crime with Predictive Policing," *AI & Global Governance Articles & Insights* (United Nations University Centre for Policy Research, February 26, 2019), https://cpr.unu.edu/ai-global-governance-turning-the-tide-on-crime-with-predictive-policing.html.

34 Albert Meijer and Martijn Wessels, "Predictive Policing: Review of Benefits and Drawbacks," *International Journal of Public Administration* 42, no. 12 (2019).

35 Hannah Couchman and Alessandra P. Lemos, *Policing by Machine: Predictive policing and the threat to our rights* (Liberty Human Rights, 2019).

36 Sarah Brayne, "Big Data Surveillance: The Case of Policing," *American Sociological Review* 82, no. 5, 2017.

37 Joris van Hoboken, "From collection to use in privacy regulation? A forward-looking comparison of European US frameworks for personal data processing," in *Exploring the Boundaries of Big Data*, eds. Bart Van Der Sloot, Dennis Broeders and Erik Schrijvers (The Netherlands Scientific Council for Government Policy, 2016).

38 John Vervaele, Surveillance and Criminal Investigation: Blurring of Thresholds and Boundaries in the Criminal Justice System?; Ales Zavrsnik, "Blurring the Line between Law Enforcement and Intelligence: Sharpening the Gaze of Surveillance?," *Journal of Contemporary European Research* 9, no 1 (2013): 52.

39 Rutger Rienks, *Predictive Policing: Taking a Chance for a Safer Future* (Politieacademie Lectoraat Intelligence, 2015).

40 Wojciech Filipkowski, "Predictive Policing using the latest technological advancements," ASC Annual Meeting Presentation, San Francisco, November 15, 2019.

41 Predictive analytical tools even go as far as advertising their capabilities of fusing police records with web data, social media information and other external sources of intelligence relating to a person's activities, history and location. See, for example, SINTELIX, "Big Data & Analytics In Law Enforcement: Predictive Policing", SINTELIX (blog), March 30, 2018, https://sintelix.com/blog/big-data-analytics-law-enforcement-policing/.

42 Sarah Brayne, Big Data Surveillance: The Case of Policing.

43 Craig D. Uchida and Marc L. Swatt, "Operation LASER and the Effectiveness of Hotspot Patrol: A Panel Analysis," *Police Quarterly* 16, no. 3, 2013.

44 Kristian Lum and William Isaac, "To predict and serve?," *Significance Journal* 10, no. 5, 2016.

such as those in London[45] and the West Midlands[46] have begun deploying equally advanced risk assessment programs used to identify high-risk individuals and areas in order to continuously monitor their likelihood of being involved in a crime on the basis of expansive datasets and broad statistical inferences.[47]

Finally, in the case of privatised predictive policing, for instance when employed by banking institutions for the prevention of financial crime, purpose limitation is already restricted from the moment of its employment. Personal data are being further processed for the purpose of crime prevention, which is different than and incompatible with the original one, *i.e.,* financial transactions. The existence of a legal basis, being the AML obligations, allows for no further consideration of necessity and proportionality. Any assessment of the purpose limitation principle, therefore, becomes in this instance obsolete.

In light of the above, it appears that the principle of purpose limitation in its current state is unable to address the issue of function creep. Embedding, however, considerations on purpose specification and compatibility of purposes already at the design phase, could signify a crucial step towards a more efficient application of the principle and the shielding of predictive policing systems against the risk of function creep. More specifically, all actors involved, both from the developers and from the LEAs and intelligence side, would be required to discuss the pragmatic data needs for the AI systems to identify crime trends and to develop individual risk assessments. This implies a better understanding of how predictive analytics function, and a clearer demarcation of types of data needed for the purpose that each specific AI system seeks to achieve. Depending on the predictive analytics employed, restrictions should be embedded in the system on the basis of, *inter alia*, explicit types of data, data origin, separately defined for crime and for commercial data, and type of crime the system aims to prevent. Where issues of compatibility or of necessity and proportionality would arise, the system could be taught to pause any action until a proper assessment with additional checks and balances is performed.

Such exercise would, first and foremost, enhance foreseeability and transparency of predictive policing analytics and the specific purposes for which they are employed. It would further reduce function creep risks by not allowing neither the AI system nor the user to expand the scope of purposes for which the system is employed. Necessitating more in-depth discussions on how to apply the purpose limitation principle, in advance of designing a predictive policing AI system, would render it more efficiently and practically enforceable.

Purpose limitation in support of information security

The implementation of AI in existing information systems amplifies the threat landscape with novel security vulnerabilities featuring a higher degree of uncertainty.[48] Unlike traditional cyber-security vulnerabilities, where flaws in computer codes are often the result of human errors, like faulty programming or intentional backdoors, AI systems embed vulnerabilities that are inherently dependent on the current state-of-the-art technology.[49] As a result, the old-fashioned process of *finding-and-patching* simply may not work until the technological research renders the whole application more robust.[50]

The use of AI-based predictive applications for policing is a use-case where system robustness and security are crucial to enable the adherence of the investigative tool with ethical and fundamental legal principles.[51]

45 Mayor of London Office for Policing and Crime, *Review of the Metropolitan Police Service Gangs Matrix* (Report, 2018).

46 Alexander Babuta and Marion Oswald, *Data Analytics and Algorithmic Bias in Policing* (Royal United Services Institute Briefing Paper, 2019).

47 Wim Hardyns and Anneleen Rummens, "Predictive Policing as a New Tool for Law Enforcement? Recent Developments and Challenges," *European Journal on Criminal Policy and Research* 24, no. 3, 2018.

48 Rik Ferguson, "Autonomous Cyber Weapons - The Future of Crime?", *Forbes*, September 10, 2019, https://www.forbes.com/sites/rikferguson1/2019/09/10/autonomous-cyber-weapons-the-future-of-crime/#36b0ba415b1a; Greg Allen and Taniel Chan, *Artificial Intelligence and National Security* (Belfer Center for Science and International Affairs, Harvard Kennedy School, 2017) 19 and 26; Miles Brundage et al., *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation* (Malicious AI Report, 2018).

49 For example, as Herpig puts it: "*Vulnerabilities in conventional software and hardware are complemented by machine learning specific ones. One example is the training data which can be manipulated by attackers to compromise the machine learning model. This is an attack vector that does not exist in conventional software as it does not leverage training data to learn*". Sven Herpig, *Securing Artificial Intelligence* (Stiftung Neue Verantwortung, 2019), 2.

50 Marcus Comiter, *Attacking Artificial Intelligence: AI's Security Vulnerability and What Policymakers Can Do About It* (Belfer Center for Science and International Affairs, Harvard Kennedy School, 2019), 3.

51 See European Commission, *On Artificial Intelligence - A European approach to excellence and trust* (white paper, February 19, 2020 - COM(2020) 65 final) 10, 18, 21-22. See also, Rec. 38, 51 and Art. 11, 29 of Directive (EU) 2016/680.

System security is strictly intertwined with the protection of individual rights. For this reason, and given the exponential expansion of the AI-related security threat landscape, technical and organizational measures are to be sought to reduce security risks when deploying such technologies.[52] In this respect, we believe the purpose limitation principle could be borrowed from the data protection field to support the adoption of information security risk-assessment processes aimed at mapping -and by so, reducing- the attack surface. The adoption of such a principle at the design phase needs further elaboration as to what extent such an approach could prove viable for an overall evaluation on the deployment of predictive tools.

Amongst the most recent regulatory proposals to tackle the uncertainties of AI, enhancing transparency of purposes to reduce risks is a suggestion backed by many authoritative sources. In 2019, the OECD – quoting Weinberger's 2018 work[53] – emphasized that one way to achieve this goal is by looking at the purpose for the use of AI against, *inter alia*, safety considerations: "This would require a declaration of what an AI system is optimised for, with the understanding that optimisations are imperfect, entail trade-offs and should be constrained by "critical constraints", such as safety and fairness."[54]

Against this backdrop, the idea of a mandatory 'suitability test'[55] has received a degree of attention. It is within such a process that we see purpose limitation as supporting information security. Suitability tests could be summarized as an assessment to be undertaken prior to the adoption of a certain AI system, whereby the involved actors are asked to evaluate security risks and uncertainties presented by the concrete deployment of AI. In other words, such an approach implies a risk-based practice in which stakeholders would ask themselves, '*How suitable is this system to achieve the declared goal, and how would I weigh this goal against the security threats of implementing AI?*'. According to Comiter, such an assessment shall take into account five parameters (added value, ease of attack, damage, opportunity cost, alternatives),[56] and the result of this exercise shall provide a documented (thus accountable) evaluation of the appropriate level of AI deployment. This is particularly important since security breaches of AI applications may lead to serious cascading effects on pre-existing interconnected information systems. Provided that security risks related to AI are often inherent to the specific use for which AI is deployed, we believe that understanding the *value* parameter (*i.e.*, the justification of the social benefit in the use of AI) could be helped by embedding an assessment phase inspired by the purpose limitation principle. This entails that the exact aim for which AI will be deployed should be declared and evaluated at the design point due to the specificity of such risks, an exercise that would result beneficial also for ex-post transparency and auditability.[57] This latter point is an element to take into serious consideration when evaluating the use of AI in a law enforcement context, whereby, given the asymmetric power relation between citizens and governments,[58] investigative methods are expected to undergo several levels of scrutiny, including public accountability, independent oversight, and forensic admissibility.

As we mentioned, such a principle would be supportive of the suitability test only insofar as it is adopted and understood as follows. Differently from data protection law, purpose limitation within the value suitability test should not only look at (personal) data, rather at the *whole predictive architecture*. This is because of the inherent (and unknown) security risks that do not solely pertain to datasets. Thus, the approach towards purpose limitation shall be intended comprehensively, i.e., considering the overall AI infrastructure. This means that we would need to re-model the two basic elements of the purpose limitation principle. Adapting it to an overall AI-architecture security approach within the suitability test would mean re-thinking its basic requirements. Purpose specification would read as *the system shall be designed and used for specified, explicit, and legitimate use*, whereas the compatible use would recite *the system shall be designed in a way that it is solely compatible with the declared use*. Should the assessment reveal incompatibility, then further evaluation of necessity and proportionality of the deployment should take place as a next step.

Secondly, the adoption of such a principle within the suitability test implies undertaking the overall assessment already in different moments, which would operationally take place in two different stages of the technology

52 See Michael Horowitz et al., *Artificial Intelligence and International Security* (Center for New American Security, 2018), 13.
53 David Weinberger, "Optimization over Explanation - Maximizing the benefits of machine learning without sacrificing its intelligence", *Berkman Klein Center* on *Medium*, January 28, 2018), https://medium.com/berkman-klein-center/optimization-over-explanation-41ecb135763d
54 Organisation for Economic Co-operation and Development (OECD), *Artificial Intelligence in Society* (report, June 11, 2019), 93.
55 Marcus Comiter, Attacking Artificial Intelligence: AI's Security Vulnerability and What Policymakers Can Do About It, 56.
56 Marcus Comiter, Attacking Artificial Intelligence: AI's Security Vulnerability and What Policymakers Can Do About It, 57.
57 OECD, Artificial Intelligence in Society, 95: '*In managing risk, there appears to be broad-based agreement that high-stakes' contexts require higher degrees of transparency and accountability, particularly where life and liberty are at stake'*.
58 OECD, Artificial Intelligence in Society, 64.

life cycle.[59] At the discovery phase, this exercise suggests a close collaboration between AI developer, decision-maker, owner and end-user,[60] towards the conclusion of a participatory risk assessment exercise. In it, the purposes for deploying AI are contextualized and assessed against both existing security threat landscapes and novel uncertainties that such a deployment brings along. The use of the purpose limitation principle may be done in an agile fashion or in a more formalistic way. The latter approach could imply turning purpose limitation (or suitability testing) into a requirement placed in contractual arrangements, procurement obligations, service agreements or other regulatory solutions.

At the live phase, should the owner of the system need a re-purposing of the AI application, declaring the purpose for deployment and confronting it with the potential security threats would have to begin again. This evaluation will have to take into account the new risks that the re-purposing brings forward for the overall information system. In doing so, adopting this principle within the value suitability test means making it a continuous exercise, thereby covering all stages of the potential attack surface, too.[61]

The added value of this recommendation is twofold. From an organizational and regulatory perspective, it would render the decision-making process more accountable and collaborative, since it would be documentable and based on multi-stakeholder participation. From a security perspective, deploying such a mandatory exercise through the application of the purpose limitation principle would increase awareness amongst all actors involved in the process regarding the level of risks behind the access to such applications.

Conclusion

This paper argues that purpose limitation in its current application might be failing to redress function creep risks born by emerging AI systems in the area of law enforcement. Specification of a legitimate purpose is often confined to a mere reference to a legal basis, i.e. laws and policy measures for crime prevention, while an assessment of compatibility, and in the case of incompatibility, an examination of necessity and proportionality seem absent. The overarching tendency to optimise the entire crime prevention system in order to mitigate emerging security threats through the use of AI and predictive analytics awards purpose limitation a secondary role. This is likely to have adverse effects on human rights as the risk of function creep threatens to gradually expand police processing of personal data and result in law enforcement practices that closely resemble broad surveillance without adequate safeguards in place. As such, we contend that the effective enforcement of this principle, already at the design phase, could safeguard human rights against the adverse effects of function creep. Furthermore, an alternative interpretation of purpose limitation can support a better evaluation of the information security of AI tools, specifically within the so-called suitability tests. In order to succeed in such an exercise, we suggest the adoption of purpose limitation at the design phase as an instrumental methodology for balancing legitimate law enforcement objectives with inherent information security risks.

59 Comiter writes, "Carefully considering and weighing [machine learning] security implications has to be done before machine learning reaches an adoption rate that would render it virtually impossible to secure a posteriori", Marcus Comiter, Attacking Artificial Intelligence: AI's Security Vulnerability and What Policymakers Can Do About It.

60 For an overview on AI and procurement in the governmental sectors, see Congressional Research Service, *Artificial Intelligence and National Security* (report, 2019), or Andrew Ilachinski, "AI, Robots, and Swarms, Issues, Questions, and Recommended Studies," (Center for Naval Analysis 2017).

61 Such as training environment, deployment environment, outside world. See Marcus Comiter, Attacking Artificial Intelligence: AI's Security Vulnerability and What Policymakers Can Do About It, 14.

References

Albert Meijer and Martijn Wessels, "Predictive Policing: Review of Benefits and Drawbacks," *International Journal of Public Administration* 42, no. 12 (2019).

Ales Zavrsnik, "Blurring the Line between Law Enforcement and Intelligence: Sharpening the Gaze of Surveillance?," *Journal of Contemporary European Research* 9, no 1 (2013): 52.

Alexander Babuta and Marion Oswald, *Data Analytics and Algorithmic Bias in Policing* (Royal United Services Institute Briefing Paper, 2019).

Andrew Ilachinski, "AI, Robots, and Swarms, Issues, Questions, and Recommended Studies," (Center for Naval Analysis 2017).

Article 29 Data Protection Working Party (WP29), 'Opinion 03/2015 on the draft directive on the protection of individuals with regard to the processing of personal data by competent authorities for the purposes of prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and the free movement of such data', WP 233, 01 December 2015.

Article 29 Data Protection Working Party (WP29), "Opinion 03/2013 on purpose limitation", WP 203, 2 April 2013.

Bert-Jaap Koops, "The Concept of Function Creep," *Law, Innovation and Technology* 13, no. 1, Published ahead of print, 03 March 2020, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3547903.

Beth Pearsall, "Predictive Policing: The Future of Law Enforcement?," *National Institute of Justice Journal* 266, no. 1 (2010).

Catherine Jasserand, "Subsequent Use of GDPR Data for a Law Enforcement Purpose: The Forgotten Principle of Purpose Limitation," *European Data Protection Law Review* 4, no. 2 (2018): 152.

Charter of Fundamental Rights of the European Union (CFREU), OJ C 202, 7.6.2016, 391–407.

Congressional Research Service, *Artificial Intelligence and National Security* (report, 2019).

Council of Europe, Convention for the protection of individuals with regard to automatic processing of personal data (Convention 108).

Court of Justice of the European Union, C-203/15 and C-698/15 Tele2 Sverige / Watson.

Court of Justice of the European Union, C-207/16 Ministerio Fiscal.

Court of Justice of the European Union, C-293/12 Digital Rights Ireland.

Court of Justice of the European Union, Opinion 1/15.

Craig D. Uchida and Marc L. Swatt, "Operation LASER and the Effectiveness of Hotspot Patrol: A Panel Analysis," *Police Quarterly* 16, no. 3, 2013.

'Criminal Law - Sentencing Guidelines - Wisconsin Supreme Court Requires Warning before Use of Algorithmic Risk Assessments in Sentencing - State v. Loomis 881 N.W.2d 749 (Wis. 2016) Recent Cases', *Harvard Law Review* 130, no. 5 (2016–2017): 1530–37.

David Lyon, *Surveillance Studies: An Overview* (Polity, 2007), 201.

David Weinberger, "Optimization over Explanation - Maximizing the benefits of machine learning without sacrificing its intelligence", *Berkman Klein Center* on *Medium*, January 28, 2018), https://medium.com/berkman-klein-center/optimization-over-explanation-41ecb135763d.

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, OJ L 119, 4.5.2016, p. 89–131.

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, OJ L 156, 19.6.2018, 43–74.

Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, OJ L 281, 23.11.1995, p. 31–50 (DPD)

European Commission, https://ec.europa.eu/info/business-economy-euro/banking-and-finance/financial-supervision-and-risk-management/anti-money-laundering-and-counter-terrorist-financing_en.

European Commission, *On Artificial Intelligence - A European approach to excellence and trust* (white paper, February 19, 2020 - COM(2020) 65 final) 10, 18, 21-22. See also, Rec. 38, 51 and Art. 11, 29 of Directive (EU) 2016/680.

European Data Protection Supervisor (EDPS) 'Opinion on the Data Protection Reform Package, 12 March 2012.

European Data Protection Supervisor (EDPS), "Opinion of the European Data Protection Supervisor on the data protection reform package", Brussels, 7 March 2012, https://edps.europa.eu/sites/edp/files/publication/12-03-07_edps_reform_package_en.pdf.

EY Global, "How data analytics is leading the fight against financial crime," EY Global (blog), 6 December 2019. https://www.ey.com/en_be/advisory/how-data-analytics-is-leading-the-fight-against-financial-crime .

Fieke Jansen, *Data Driven Policing in the Context of Europe* (Data Justice Lab Working Paper, May 2018).

Fumio Shimpo, "The Principal Japanese AI and Robot Strategy toward Establishing Basic Principles", in *Research Handbook on the Law of Artificial Intelligence* (Edward Elgar Publishing, 2018), 116, https://www.elgaronline.com/view/edcoll/9781786439048/9781786439048.00015.xml.

Greg Allen and Taniel Chan, *Artificial Intelligence and National Security* (Belfer Center for Science and International Affairs, Harvard Kennedy School, 2017).

Hannah Couchman and Alessandra P. Lemos, *Policing by Machine: Predictive policing and the threat to our rights* (Liberty Human Rights, 2019).

Joel Rubin, "Stopping crime before it starts," *Los Angeles Times,* August 21, 2010.

John Vervaele, Surveillance and Criminal Investigation: Blurring of Thresholds and Boundaries in the Criminal Justice System?.

Jonas Christoffersen, *Fair balance: Proportionality, subsidiarity and the primarity in the European Convention on Human Rights* (Martinus Nijhoff Publishers, 2009).

Joris van Hoboken, "From collection to use in privacy regulation? A forward-looking comparison of European US frameworks for personal data processing," in *Exploring the Boundaries of Big Data*, eds. Bart Van Der Sloot, Dennis Broeders and Erik Schrijvers (The Netherlands Scientific Council for Government Policy, 2016).

Kristian Lum and William Isaac, "To predict and serve?," *Significance Journal* 10, no. 5, 2016.

Langdon Winner, *Autonomous Technology: Technics-out-of-Control as a Theme in Political Thought* (Cambridge, Mass.: MIT Press, 1977), 28.

Liana Colonna, "Data Mining and Its Paradoxical Relationship to the Purpose Limitation Principle," in *Reloading Data Protection* Multidisciplinary Insights and Contemporary Challenges, eds. Serge Gutwirth, Ronald Leenes, Paul De Hert (Springer Netherlands, Dordrecht 2014), 299.

Lorna McGregor, Daragh Murray, and Vivian Ng, "International Human Rights Law as a Framework for Algorithmic Accountability'," *International & Comparative Law Quarterly* 68, no. 2 (2019): 315–316.

Marcus Comiter, *Attacking Artificial Intelligence: AI's Security Vulnerability and What Policymakers Can Do About It* (Belfer Center for Science and International Affairs, Harvard Kennedy School, 2019).

Mayor of London Office for Policing and Crime, *Review of the Metropolitan Police Service Gangs Matrix* (Report, 2018).

Michael Horowitz et al., *Artificial Intelligence and International Security* (Center for New American Security, 2018).

Miles Brundage et al., *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation* (Malicious AI Report, 2018).

Norbert Wiener, "Some Moral and Technical Consequences of Automation," *Science* 131, no. 3410 (5 June 1960): 1358, https://doi.org/10/ftpxdb.

Odhran James McCarthy, "AI & Global Governance: Turning the Tide on Crime with Predictive Policing," *AI & Global Governance Articles & Insights* (United Nations University Centre for Policy Research, February 26, 2019), https://cpr.unu.edu/ai-global-governance-turning-the-tide-on-crime-with-predictive-policing.html.

OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data, 1981 (updated in 2013), https://www.oecd.org/internet/ieconomy/oecdguidelinesontheprotectionofprivacyandtransborderflowsofpersonaldata.htm .

Organisation for Economic Co-operation and Development (OECD), *Artificial Intelligence in Society* (report, June 11, 2019) .

PHRP Expert Meeting On Predictive Policing (Amnesty International Police and Human Rights Programme, May 2019).

Recent Developments and Challenges," *European Journal on Criminal Policy and Research* 24, no. 3, 2018.

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), (GDPR), OJ L 119, 4.5.2016, 1–88.

Rik Ferguson, "Autonomous Cyber Weapons - The Future of Crime?", *Forbes*, September 10, 2019, https://www.forbes.com/sites/rikferguson1/2019/09/10/autonomous-cyber-weapons-the-future-of-crime/#36b0ba415b1a.

Rutger Rienks, *Predictive Policing: Taking a Chance for a Safer Future* (Politieacademie Lectoraat Intelligence, 2015).

Sarah Brayne, "Big Data Surveillance: The Case of Policing," *American Sociological Review* 82, no. 5, 2017.

SINTELIX, "Big Data & Analytics In Law Enforcement: Predictive Policing", SINTELIX (blog), March 30, 2018, https://sintelix.com/blog/big-data-analytics-law-enforcement-policing/.

Stuart Russell, *Human Compatible: Artificial Intelligence and the Problem of Control* (Penguin Publishing Group, 2019).

Sven Herpig, *Securing Artificial Intelligence* (Stiftung Neue Verantwortung, 2019).

Walter L. Perry, Brian McInnis, Carter C. Price, Susan C. Smith, John S. Hollywood, *Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations* (RAND Corporation, 2013).

Wim Hardyns and Anneleen Rummens, "Predictive Policing as a New Tool for Law Enforcement?

Wojciech Filipkowski, "Predictive Policing using the latest technological advancements," ASC Annual Meeting Presentation, San Francisco, November 15, 2019.

Zach Friend, "Predictive Policing: Using Technology to Reduce Crime," *FBI Law Enforcement Bulletin*, April 9, 2012.

4. FROM EVIDENCE TO PROOF: SOCIAL NETWORK ANALYSIS IN ITALIAN CRIMINAL COURTS OF JUSTICE

Dr. Roberto Musotto*

Abstract

Social network analysis has changed the way scholars and analysts look at human relationships and it has also helped understand why, when, and how specific behavioural choices and resultant legally relevant interactions are made. Legal documents such as laws, policies and court cases are increasingly used as the starting point for such analysis, yet, its practical application in the judicial procedure is sparse or non-existent. This raises the following questions: is it possible to use social network analysis as an investigative and judicial tool to implement the production of evidence in court? More importantly, how would it account for helping public prosecutors, judges, juries and defenders in presenting evidence during a trial? In which ways could the analysis of a network contribute to the creation of a legal proof?

In this paper, arguments are presented for understanding how this implementation is possible and answering the above questions. There are opportunities but also potential pitfalls to introducing network analysis as a means by which proof can be created. The paper will draw from the context of the Italian criminal courts of justice and its criminal procedure laws in order to understand how social network analysis would fit and behave, especially in the pursuit of serious crimes.

Social network analysis in criminal courts of justice can be successfully adopted not just as an investigative tool, but also as a cost-cutting filter in preliminary hearings and as an efficient threshold for sentencing, enabling judges, prosecution and defence to give and asses evidence. These advantages, however, still present potential bias such as the limits of the criminal network and the subjective way the network could have been built.

Keywords: Social network analysis, criminal procedure, legal proceedings, evidence, proof, criminal justice.

Introduction

Social Network Analysis (SNA) is an extensive approach to methodological techniques that allows analysts and researchers to explore segments of patterns that recur, forming in social interactions.[1] These patterns that form between individuals and amongst groups contribute to explaining the decisions and choices that actors take in their everyday lives because their rational actions follow a flux of influences, driven by the need to maximize wealth, utility or happiness. Decisions and interactions can be promoted or constrained through the application of the law. A contract, for example, promotes the interaction of parties with competing interests, while a criminal law – whose primary aim is deterrence – can prevent that same interaction. Criminal law aims to discourage members of society from committing specific actions that are considered to be illegal according to the criminal law of their country. Should such actions be committed, criminal procedure laws ensure the fairness, efficiency, and effectiveness of the whole process, guaranteeing that, from a practical perspective, if a crime has been committed, a judicial body will deliver a judgement.

The relevance of SNA has been established for a long time[2] in the social sciences and, specifically, in a criminology context, mainly to describe the topology of the network[3] or differential features that a group has and the influence of the group over individual choices.[4] In this context, academic research has been looking at social interactions between individuals committing criminal actions in terms of a networked community.[5]

Notwithstanding that these interactions have been the subject of academic research, no attention at all has been paid to the practical and procedural application of these techniques inside courts of justice. This paper shows how and where Social Network Analysis techniques can be successfully transferred to courts of justice. Judicial documents are commonly used as the starting point for SNA analysis, proving to be helpful for understanding how networks operate, but the actors involved in the production of legal and judicial records rarely, if ever, use such a methodology to drive their investigation[6]. Reasons for this could be: a) lack of time during investigations and trial; b) issues in managing costs of the trial, and c) lack of skills and knowledge required to run such analysis. This exploratory article looks at judicial documents as an endpoint of an analysis done through social networks.[7] It deals with the problem of transferring social network analysis techniques into judicial courts of justice and how this could be practically done in the context of the Italian criminal procedural system. This article points out that Social Network Analysis could be applied well over its primary nature as an investigative tool, but also, despite its potential bias, be a companion in the creation and evaluation of evidence in preliminary hearings, trials and sentencing.

---

1 The author thanks Hanna Ahlström, Adam Hardy, Aleksandra Irnazarow, Rebecca Marples, Corinne McAlary, Martin Nøkleberg, Brian Nussbaum, Eszter Párkányi, Yanna Papadodimitraki, Maria Grazia Porcedda, Camilo Tamayo Gomez, Jakob Walberg, David Wall and Ken Yin for their comments and suggestions. A first version of this article was presented at the 2nd Annual Workshop on Law and Social Science Methods, University of Oslo, Norway in October 2018. Interviews were taken at events under Chatham House Rule. The work has been supported by the Cyber Security Research Centre Limited whose activities are partially funded by the Australian Government's Cooperative Research Centres Programme.
* Cyber Security Cooperative Research Centre, School of Business and Law, Edith Cowan University, Perth, Australia. Email: r.musotto@ecu.edu.au
Malcolm K. Sparrow, "The application of network analysis to criminal intelligence: An assessment of the prospects." *Social networks* 13, no. 3 (1991): 251-274. https://doi.org/10.1016/0378-8733(91)90008-H.

2 Moreno's research on inmates from Sing Sing prison and the Hudson reform school are the first examples of social interaction analysis between inmates and criminal behaviour see Jacob L. Moreno, Application of the group method to classification. New York: National Committee on Prisons and Prison Labor, 1932; Jacob L. Moreno, The first book on group psychotherapy. Psychodrama and Group Psychotherapy Monographs 1, (Beacon House, 1957). For a complete review of criminological theories and their interaction with social relations see Alex R. Piquero, ed. The handbook of criminological theory (John Wiley & Sons, 2015).

3 I.e. the arrangement of the network: Lawrence M. Besaw et al., "Graphic display of network topology." U.S. Patent 5,276,789 filed May 14 1990 and issued January 4 1994; Gueorgi Kossinets, and Duncan J. Watts, "Empirical analysis of an evolving social network." *Science* 311, no. 5757 (2006): 88-90. And how its shape changes under different constraints: Rasmus Petersen Rosenqvist, Christopher J. Rhodes, and Uffe Kock Wiil, "Node removal in criminal networks." In *2011 European Intelligence and Security Informatics Conference*, ed. IEEE Staff, 360-365. IEEE, (2011).

4 John Scott, and Peter J. Carrington, *The SAGE handbook of social network analysis* (London: SAGE, 2011), chap. 17, 236-255, https://methods-sagepub-com.ezproxy.ecu.edu.au/book/the-sage-handbook-of-social-network-analysis.

5 Yong Lu et al., "Social network analysis of a criminal hacker community." *Journal of Computer Information Systems* 51, no. 2 (2010): 31-41.

6 Author, interview with public prosecutor of Netherlands, September 2018.

7 De facto it aims at changing the tack of current literature in criminal networks: from an *ex post* to an *ex ante* analysis.

As a case study on a single country, its results can be generalised into the majority of civil law legal systems, with the due exception of countries that adopt a purely inquisitorial legal system.[8] It aims at reversing the direction of some of the current research in social network analysis and criminology: many researchers look at how it is possible to employ judicial documents for their analysis[9] in order to predict or prevent the behaviour. While analysts and academics start their research from a judicial document[10], this paper argues that law practitioners would benefit from such analysis to create a legally relevant document.

Ideally, this article expands from the idea of 'Robot Judges',[11] or 'AI Judges',[12] into the realm of the other law professionals involved in litigation and dispute resolution. The advantage of employing techniques from social network analysis inside a judicial system is that it allows to assess whether the behaviours and patterns arising from SNA are legally relevant or, perhaps, not. In the civil sector, neural networks[13] have been explored already in alternative dispute resolution in construction claims and have been implemented in certain jurisdictions[14] for small motoring claims. These two types of claims have some similarities and are prone to a more automated process because: a) claims are homogeneous (i.e. they aim at obtaining the same effect) b) there is a limited number of arguments that can be presented in court as a defence, c) there is a lower risk of wrongdoing by courts d) there is an interest in the parts of solving the claim quickly. In criminal courts of justice, the implementation of network analysis is mainly aimed at helping to find nuances in court cases and not at overcoming the role and function of dispute resolution lawyers.[15]

Two pressing issues will be addressed: firstly, how social network analysis can take its place within a legal system and, secondly, how it can contribute to proving links and connections beyond any reasonable doubt, turning them from investigative evidence into judicial proof.

By having SNA techniques implemented in a judicial context, there is an opportunity to save time, as the analysis of patterns could be partially automated, therefore facilitating the work of prosecution. Saving time would consequentially reduce the cost required to facilitate the investigation, prosecution, and trial as it would make the whole system more efficient. To benefit from these advantages, however, the current lack of knowledge in the legal profession regarding this area must be overcome through education, allowing practitioners to understand both positive elements and potential pitfalls.

This paper is structured as follows. The next section looks at the complexity of social network analysis in crime research. Section three presents how the criminal legal system in the Italian courts of justice is structured, followed by a description of possible interpolations of network techniques according to the current criminal procedure code and system. The fourth section reviews and discusses the implications of these results.

---

8 Distinctions must be made for those countries adopting adversarial systems of dispute resolution (where the same considerations apply) and inquisitorial common law systems (where results of this article might not apply), in terms of the diverse standard that is necessary to transform evidence collected from investigations or during the trial into a judicial proof. In inquisitorial legal systems the court investigates the case and collects the evidence that then is presented in court to a defendant. Therefore, while SNA would still be a useful tool for the investigation, it might not be needed in a preliminary hearing because the selection and evaluation of evidence would have happened at an earlier stage without the involvement of the defence. The case study of this article reflects the experience of an adversarial criminal law system.

9 Leslie Ball, "Automating social network analysis: A power tool for counter-terrorism." *Security Journal* 29, no. 2 (2016): 147-168; Byungun Yoon and Yongtae Park, "A text-mining-based patent network: Analytical tool for high-technology trend." *The Journal of High Technology Management Research* 15, no. 1 (2004): 37-50.

10 Leslie Ball, "Automating social network analysis: A power tool for counter-terrorism." *Security Journal* 29, no. 2 (2016): 147-168; Byungun Yoon, and Yongtae Park. "A text-mining-based patent network: Analytical tool for high-technology trend." *The Journal of High Technology Management Research* 15, no. 1 (2004): 37-50. See, among all, the way the criminal network has been built in Berlusconi et al. (2016), Ficarra et al (2019) and Lu et al. (2010): different judicial documents have been sourced first and then coded to draw assumptions about the group. Giulia Berlusconi et al., "Link prediction in criminal networks: A tool for criminal intelligence analysis." *PloS one* 11, no. 4 (2016), doi: https://doi.org/10.1371/journal.pone.0154244; Annamaria Ficara et al., "Social Network Analysis of Sicilian Mafia Interconnections." *International Conference on Complex Networks and Their Applications* 8. In *Social Network Analysis of Sicilian Mafia Interconnections*, (Springer, Cham, 2019), 440-450; Yong Lu et al., "Social network analysis of a criminal hacker community." *Journal of Computer Information Systems* 51, no. 2 (2010): 31-41.

11 Aviva Rutkin, "The judge is a robot." *New Scientist* 222, no. 2973 (2014): 24.

12 AI judges are those assisted or enabled by technology in their decision-making tasks. A robot judge is a computer that draws patterns and makes decisions without a human interaction (Sourdin, 2018). Tania Sourdin, "Judge v. Robot: Artificial Intelligence and Judicial Decision-Making." *UNSWLJ* The 41, no. 4 (2018): 1114.

13 N. B Chaphalkar, and Sayali S. Sandbhor, "Application of neural networks in resolution of disputes for escalation clause using neuro-solutions." *KSCE Journal of Civil Engineering* 19, no. 1 (2015): 10-16.

14 In the UK a chatbot has been created to dispute parking fines (Laurie and Potton, 2017). Laurie Points, and Ed Potton, "Artificial intelligence and automation in the UK." Briefing Paper: 8152 (2017); while Estonia is creating an 'AI Judge' for small claims (Fabian, 2020). Seda Fabian, "Artificial Intelligence and the Law: Will Judges Run on Punch Cards." *Common L. Rev.* 16 (2020): 4.

15 There are also important questions to be raised about rights at stake, as discussed *infra*.

Social Network Analysis in crime and legal research

Several studies have drawn from judicial documents and legal sources in order to apply social network techniques to map, describe and even predict criminal networks and community behaviour.[16] In such a context, a community is not the sum of independent choices made by a group of individuals, but, following Simmel,[17] it is the result of their interactions that constantly influence and shape relationships. In criminal networks, there is a portion of the community in which the majority is in favour of those who wish to break the rules, rather than those individuals who respect them.[18] The intensity and frequency of these relationships create criminal behaviour, which is noticeable through dynamics that can emerge during an investigation or a trial, as people's actions are recorded within a specific context[19] and time frame.

SNA in its simplest form, link analysis,[20] has been already deployed as an aid in investigations and trials, specifically in money laundering,[21] and gang cases.[22] In the US, COPLINK software, for example,[23] acts as a database system where different entities can be connected through the results of text analysis. COPLINK provides link analysis that is currently admitted in trials.[24] It is a softer version of SNA, mainly employed as a visualisation tool. What is still missing, however, is a more elaborate version of SNA that provides measurable evidence for the involvement of each individual or the standard set by the criminal conduct.

Furthermore, judicial documents are one of the various resources for identifying and building social networks. Campana and Varese (2012),[25] among others, agree that the access to these kinds of documents allows researchers to gather data from wiretaps and investigations with a certain degree of validity and reliability as long as there is: a) no self-censorship within the group that is investigated; b) there is widespread group coverage with c) a large sample of conversations. Once SNA is applied to the same legal documents employed by legal operators, it would help achieve a better understanding of the group dynamics that are investigated. On the other hand, there are methodological problems concerning the limits of the network[26] and the problem of missing information in covert networks,[27] which translate into unclear or imprecise dynamics within the group,[28] since the investigation could have focused on just one part of the whole network[29] or set a threshold that is too high or too low for measuring the involvement of the single individual into the illicit group.

In a broader context, consideration should be given to understanding why network dynamics taken from these resources cannot be more precise: it is because there is a trade-off between people's rights[30] and the need to have thorough investigations and trials. On the one hand, there is the interest on the part of the state

16 Sparrow, "analysis to criminal intelligence." 251-274; Mangai Natarajan, "Understanding the structure of a large heroin distribution network: A quantitative analysis of qualitative data." *Journal of Quantitative Criminology* 22, no. 2 (2006): 171-192; Edward R. Kleemans, and Christianne J. De Poot, "Criminal careers in organized crime and social opportunity structure." *European Journal of Criminology* 5, no. 1 (2008): 69-98; Francesco Calderoni, "The structure of drug trafficking mafias: the 'Ndrangheta and cocaine." *Crime, law and social change* 58, no. 3 (2012): 321-349; Morselli, Carlo, 2009. *Inside criminal networks* 8 (New York: Springer, 2009); David A. Bright, Caitlin E. Hughes and Jenny Chalmers, "Illuminating dark networks: A social network analysis of an Australian drug trafficking syndicate." *Crime, law and social change* 57, no. 2 (2012): 151-176; Roberto Musotto, "Social and spatial network analysis of organised crime." PhD dissertation, Università degli Studi di Messina (2016).

17 Georg Simmel, "The problem of sociology." *American Journal of Sociology* 15, no. 3 (1909): 289-320

18 Edwin H. Sutherland, Donald R. Cressey, and David F. Luckenbill, *Principles of criminology* (Lanham: AltaMira Press, 1992).

19 Gueorgi Kossinets, "Effects of missing data in social networks." *Social networks* 28, no. 3 (2006): 247-268.

20 Where different objects are connected and evaluated in a network according to one of their property.

21 Andrea F. Colladon and Elisa Remondi, "Using social network analysis to prevent money laundering." *Expert Systems with Applications* 67 (2017): 49-58.

22 Davis, Roger H. "Social network analysis: An aid in conspiracy investigations." *FBI L. Enforcement Bull* 50, (1981): 11; Paul AC. Duijn, and Peter PHM. Klerks, "Social network analysis applied to criminal networks: recent developments in Dutch law enforcement." In *Networks and network analysis for defence and security*, (2014): 121-159.

23 Jennifer Schroeder, Tucson Police Dept, and United States of America. "COPLINK: Database integration and access for a law enforcement intranet, final report." *Washington, DC: US Department of Justice* (2001).

24 Paul Jen-Hwa Hu et al., "Law enforcement officers' acceptance of advanced e-government technology: A survey study of COPLINK Mobile." *Electronic Commerce Research and Applications* 10, no. 1 (2011): 6-16.

25 Paolo Campana, and Federico Varese, "Listening to the wire: criteria and techniques for the quantitative analysis of phone intercepts." *Trends in organized crime* 15, no. 1 (2012): 13-30.

26 I.e. who can be part of the network.

27 Giulia Berlusconi, "Do all the pieces matter? Assessing the reliability of law enforcement data sources for the network analysis of wire taps." *Global Crime* 14, no.1 (2013): 61-81; Antonio Scaglione, *Reti mafiose. Cosa Nostra e Camorra: organizzazioni criminali a confronto*. FrancoAngeli, 2011.

28 Valdis E. Krebs, "Mapping networks of terrorist cells." *Connections* 24, no. 3 (2002): 43-52.

29 Renee C. Van der Hulst, "Introduction to Social Network Analysis (SNA) as an investigative tool." *Trends in Organized Crime* 12, no. 2 (2009): 101-121.

30 Especially human rights.

to punish (*ius puniendi*)[31] the perpetrator of a crime and, on the other, the right to respect one's private life, without interference by a public authority.[32] Through a careful blend of the two,[33] it is possible to reach an understanding of the network while avoiding any unnecessary restriction of rights without just cause. The next section presents the case study, describing the Criminal Justice system in Italy and discusses where network techniques can coexist within this legal framework.

Network analysis and the Italian Criminal Justice system

Italian Criminal Justice is an adversarial system of litigation,[34] where a public prosecutor (*Pubblico Ministero*) and a defendant collect and present evidence to a judge who decides which can be elevated as proof in the sentencing of an alleged crime. The procedure[35] - in its ordinary structure - is shared in four distinct phases: investigation, preliminary hearing, trial, and execution.

During the investigation, circumstantial evidence is collected by police investigating a crime. All evidence is collected according to a specific procedure that is listed in the criminal procedure code,[36] but other techniques can be employed if expressly authorised by the public prosecutor.[37] At the end of an investigation,[38] the public prosecutor, if it is believed that a crime has been committed, asks police to take all the people that have been investigated[39] into custody. The motivation and evidence are collected in the arrest warrant.[40]

There are several reasons why it is propitious to use network analysis in the arrest warrant; not just for researchers, but also for legal operators, in that: 1) it is the document that closes the investigation, so no further evidence will be collected. By being the last document in the investigation, 2) it is also the most exhaustive document that sources out evidence. Against this backdrop, 3) SNA contributes in giving a glance of the bigger picture straight away - showing who is more important in the network - out of lengthy documents and 4) any conduct and interaction is documented in the arrest warrant objectively and devoid of legal relevance.[41] These advantages are discussed *infra*.

In Italy, there is no evidence of employing this SNA in investigations,[42] yet where the use of social network analysis tools is widespread in other police forces such as in Virginia (USA), Australian Victoria and New South Wales, these advantages are clearly visible, as is already outlined in the literature.[43]

---

31 According to Kelsen (1967), this is the power of the State and thus, by acting in the interest of other people's subjective rights it expresses and defines the range of behaviours that are authorised. Hans Kelsen, *Pure Theory of Law*. Berkeley: University of California Press, 1967.

32 This is a right that is present in multiple constitutions and conventions, under different formulations. Among all, article 8, second section of the European Convention on Human Rights (ECHR) that states: "there shall be no interference by a public authority with the exercise of this right except such as is in accordance with the law and is necessary in a democratic society in the interests of national security, public safety or the economic well-being of the country, for the prevention of disorder or crime, for the protection of health or morals, or for the protection of the rights and freedoms of others."

33 Whose discussion is beyond the scope of this article.

34 Until 1988 it used to be a mixed inquisitory system, where wide powers were given to police and investigating judges in order to gather evidence to bring into court (Garofoli, 2008). Vincenzo Garofoli, *Diritto processuale penale*. (Giuffrè Editore, 2008).

35 According to the specific crime investigated and the nature of the investigated people, there are more complex or shortened procedures in place.

36 These are namely inspection, perquisition, seizure of evidence and wiretapping. Wiretaps are the most relevant for building and investigating a network. However, all of them, as well as the way evidence is collected (e.g. by seizing a list of contacts or agenda) can contribute in the analysis of a specific group. For example, in Sicilian Mafia trials the use of paper documents (called *pizzini*) was extremely useful in building links and transactions between different affiliated members and their victims (Musotto, 2016). Roberto Musotto, "Social and spatial network analysis of organised crime." PhD dissertation, Università degli Studi di Messina, (2016).

37 Over time, the supreme court authorised the use of more sophisticated investigative tools, such as the use of Trojan (Cass. n. 26889 28th of April 2016) and spyware (by law, this time: n.103/2017). Their use must be justified by the gravity and nature of the crime investigated. Therefore, it has been authorised for Mafia-related and human trafficking trials but refused for lower offence crimes.

38 Which can last maximum up to two years for extremely complex cases (art. 407 of the criminal procedure code.

39 According to the Italian criminal procedure code, there is a difference between arrest and custody, which is highlighted here for the sake of clarity. The first is the power to guarantee criminals to justice and to block them from committing more illicit actions. It can be executed only when the offence is in flagrante (a person is committing an offence) or in quasi flagrante (a person just committed an offence). It is executed by police or, sometimes, by private citizens and it cannot be issued. The second is a preventive detention that is executed in all the other situations, that do not allow for an arrest to take place, and it is issued by the *Pubblico Ministero* (Public Prosecutor).

40 *Provvedimento di fermo* in Italian.

41 As legal relevance of the conducts and interactions will be evaluated afterwards by a judge.

42 Except for hierarchy-type listing of Mafia/large criminal groups (Falcone and Turone, 2015). Giovanni Falcone, and Giuliano Turone. "Tecniche di indagine in materia di mafia." *Rivista di Studi e Ricerche sulla criminalità organizzata* 1, no. 1 (2015): 116-153.

43 Morgan Burcher, and Chad Whelan. "Social network analysis as a tool for criminal intelligence: Understanding its potential from the perspectives of intelligence analysts." *Trends in organized crime* 21, no. 3 (2018): 278-294; Jennifer Johnson et al., "Social network analysis: A systematic approach for investigating." *FBI Law Enforcement Bulletin,* (2013): 350. https://leb.fbi.gov/articles/featured-articles/social-network-analysis-a-systematic-approach-for-investigating.

The arrest warrant discloses for the very first time the fact that there is an ongoing investigation. The juridical nature of the documents is unique.[44] Usually, the arrest warrant is one of the very first documents issued in any criminal process in Italy by the Public Prosecutor and, most of the time, the last document of every investigation, together with the notification of its end. Ideally, every investigation leads to an arrest warrant if there is enough evidence to support a prosecution. The order of custody can be issued by the Public Prosecutor under a series of alternate conditions: 1) there is an unequivocal weight of evidence against the person under investigation; 2) there is a risk the person may try to evade justice because of the weight of evidence against him/her; 3) the investigated crime may be punished with life in prison or a jail term with at least two years of jail time as a minimum and at least six years as maximum if it concerns terrorism or democratic subversion (as Mafia or terrorism crimes do).

However, the most significant feature of the warrant is that the Public Prosecutor is legally obliged to present all evidence against and in favour of the person that is currently under investigation. This means that technically the Prosecutor does not have any real power to select evidence,[45] because this will be done during the trial in an adversarial way where the defence can bring other evidence to reframe or refute the alleged legally relevant conduct. All evidence must be presented to the deciding judge by the prosecutor so that the judge can evaluate the position of the person under investigation[46] in light of the evidence. What the Public Prosecutor could show at this stage through Social Network Analysis is: a) whether there is an appearance of a structure or not;[47] b) who holds a more important role in the network;[48] c) how bonded the group is.[49]

At the preliminary hearing, all evidence is presented to analyse the result of the investigation and for the judge to decide if there is enough evidence to proceed with a trial against the person accused. This is where SNA can be used to look at the bigger picture, while also being important for the single accused person. The judge has to measure the probability that evidence collected is exactly linking to the accused person.[50] In such a context, an analysis of the dynamic influence of nodes and links,[51] as well as an understanding of centrality or homophily measures[52] would be helpful as a persuasive argument in order to assess the involvement of the single individual in the organisation and its relevance as criminal conduct, because it would allow the creation of a quantifiable measure against the standard conduct forbidden by the law.

---

44 And part of the literature (Scaglione, 2011, Ficara et al., 2019, Musotto, 2016, Wall and Musotto, 2019) successfully deployed arrest warrants as a starting point for social network analysis. Paolo Tonini, *Manuale di procedura penale* (Giuffrè editore, 2012); Antonio Scaglione, *Reti mafiose. Cosa Nostra e Camorra: organizzazioni criminali a confronto* (FrancoAngeli, 2011); Annamaria Ficara et al., "Social Network Analysis of Sicilian Mafia Interconnections." In *International Conference on Complex Networks and Their Applications* (Springer, Cham, 2019), 440-450; Roberto Musotto, "Social and spatial network analysis of organised crime." PhD dissertation, Università degli Studi di Messina, (2016); Roberto Musotto, and David S. Wall, "Are Booter Services (Stressors) indicative of a new form of organised crime group online?" *Journal of Digital Forensics* 1, no.1 (2019): 41-50.

45 This is so the only proof he is actually allowed to discard are the ones that are redundant (e.g. doubles) or not relevant (e.g. recordings of an alleged mafioso not engaged in criminal or suspicious conduct) to the investigation, as prescribed by Law number 332 from 1995 and Law number 47 from 2015. This task of collecting all the proofs that are in favour and against the investigated person is extremely difficult to achieve in many cases because, after all, the Public Prosecutor must prove the foundation of the alleged accusations. In terms of social network analysis this can nonetheless create problems of fuzzy boundaries in defining the criminal network. E.g. two people are sitting in a car while one of them starts talking about incriminating facts, such as racketing a business or meeting other suspects. The car is wiretapped and the listener is not known as an offender. Investigators and Public Prosecutor, while listening to this conversation, have to make a choice: should the listener be included in the criminal network or not? If yes, they might suspect that he would probably know more or share more information in the future. This might result in the criminal network expanding to people that might not be part of it, thus causing an unreasonable compression of their rights. If not, part of the covert network might not be revealed.

46 The peculiarity of this source is that the amount of available data for each node (person connected) and link (interaction) reduces the risks of missing data inside the investigated network, assuming the investigation has been done thoroughly. Even if the entire list of links inside the organisation may not be evident to investigators, all the observable information about every link is gathered together with the help of wiretap records, audio and video collections, family trees and documentations. The information collected should be therefore the least biased possible. Afterwards, during the trial the judge will create a smaller network of criminally punishable nodes and links. The fact that the arrest warrant is more complete than the other judicial acts in the entire process and investigation is the reason why the network analysis would be helpful for both researchers and investigators.

47 This is assessed through an analysis of the type of *connections* of every single individual investigated.

48 This is done through an analysis of *centrality* measures as they tell who is more important in that specific network.

49 This would show the presence of inner groups within a group and it could be done through *clique* analysis or local *cluster* coefficient. The first studies the social circle and how it is connected compared to the rest of the group. The second tells how likely it is that some individuals (or objects) are connected and likely to become a clique (in a simplified example, if A knows B that knows C and D, it is likely that C and D know each other or A).

50 For example, the preliminary hearing judge could consider SNA centrality measures and set a numeric threshold in order to decide who to prosecute or not, while also being able to require further investigation or dismiss the case for those individuals who do not meet such requirements.

51 Stephen P. Borgatti et al., "Network analysis in the social sciences." *Science* 323, no. 5916 (2009): 892-895.

52 Martin Nøkleberg, "Examining the how of plural policing: Moving from normative debate to empirical enquiry." *The British Journal of Criminology*, (2020). doi: https://doi.org/10.1093/bjc/azz080

If the trial proceeds, the judge will gather evidence from the public prosecutor and the defence, and it is they who decide on the weight of proof against the defendant, as any proof in favour of the defence might act in counterbalance to the allegations. The judge is the body that makes a selection between evidence, allowing it to be considered as proof.[53] He is helped in this task by means of bringing evidence into trial.[54]

The trial ends with a verdict that is the result of this selective activity, which is another point where social network analysis techniques would fit because in all those offences, where an organisation[55] and a group is involved, the network has to be 'virtually reconstructed'[56] by the judge and every proof evaluated in the light of its criminal relevance. What results is not the actual network, but the one relevant to criminal laws. So, the network that is displayed by the judge is *per se* limited to a smaller fraction of the list of nodes and links that was first presented to the judgement.[57] SNA here would be able to automate this procedure once the threshold is set by the deciding judge. The latter network is the result of what is criminally relevant, i.e., what it is worth to be punished according to criminal laws.[58] In fact, what matters for the judicial body is that all the available evidence about the organisation of crime and a criminal organisation is set up for judgement.

Once the judge crystallises the network, and the verdict is emitted, the process of reselecting nodes and links can be repeated in the two following appeals. The main difference from the first trial in terms of network analysis is that it is not possible to bring new evidence to trial and therefore create a new proof.[59] In the appeal there are two possible network outcomes: the size and connections are confirmed or some of the accused can be acquitted by the court with the result of reducing the network to a smaller fraction. In the final phase, when all the means of appeal have been exhausted or expired, sentences must be executed, aiming at re-educating the offender in the light of his social rehabilitation.[60]

Conclusion

The aim of this paper was to explore whether, where, and how social network analysis would fit in a court system. By employing the current organisation of Italian criminal courts of justice, it allows for the identification of multiple areas where future practice and research should focus: 1) as a persuasive tool at the end of investigations; 2) as a means to turn evidence into proof during the trial and 3) as a tool for judges in selecting the criminal relevant network.

Social Network Analysis in court serves two purposes: 1) to systematise and canalise the amount of intelligence collected during the investigation over a group of individuals,[61] and, at the same time, 2) to simplify and automate the activity of public prosecutors and judges into the creation of proof during a trial, given rigid

53 Or the jury in specific crimes and trials (e.g. crimes against humanity).

54 Which are listed from article 194 until 243 of the Italian criminal procedure code.

55 Even a minimal one.

56 The judge re-creates the structure of the network on the basis of the evidence in front of him.

57 Giulia Berlusconi et al., "Link prediction in criminal networks: A tool for criminal intelligence analysis." *PloS one* 11, no. 4 (2016). doi: https://doi.org/10.1371/journal.pone.0154244.

58 For researchers, this is why the sentence is less apt to be analysed or to reconstruct the network than the arrest warrant. In this latter document, proofs are displayed in a raw format and have not yet been filtered through the grinding procedure of the judgement, which is a contingent process. Consequently, the arrest warrant has interesting features in order to gather quality data for network analysis concerning criminal networks. This is so, because, despite the fact it is a secondary source that might have incorrect attributions in the way that nodes and ties are shaped and created (Campana, 2016), the way the warrant is written might reduce the issues of missing data. Paolo Campana, "Explaining criminal networks: Strategies and potential pitfalls." *Methodological Innovations* 9, (2016).

59 Unless they were discovered after the end of the first trial as prescribed by the article 603 of the criminal procedure code.

60 This is another area where SNA would be able to help. Inmates keep socialising during their time in prison and their network influence can result in positive or negative implications for recidivism (Berg and Huebner, 2011; Duwe and Clark, 2013). Mark T. Berg, and Beth M. Huebner. "Reentry and the ties that bind: An examination of social ties, employment, and recidivism." *Justice quarterly* 28, no. 2 (2011): 382-410; Grant Duwe, and Valerie Clark. "Blessed be the social tie that binds: The effects of prison visitation on offender recidivism." *Criminal Justice Policy Review* 24, no. 3 (2013): 271-296. and expanding the criminal network (Tremblay, 2017). Pierre Tremblay, "Searching for suitable co-offenders." In *Routine activity and rational choice* ed. (R. V. G. Clarke and Marcus Felson, Routledge, 2017), 17-36. So SNA could be employed to predict (Gravel and Tita, 2015) who might be at a higher risk by looking at connections and interactions. Jason Gravel, and George E. Tita. "With great methods come great responsibilities: social network analysis in the implementation and evaluation of gang programs." *Criminology & Pub. Policy* 14, no.3 (2015): 559-527.

61 Author, interview with Officer from Valencia Police, February 2018.

procedural timeframes,[62] and the notoriously excessive length of judicial proceedings.[63] The advantage of using SNA in courts is that it saves time and therefore reduces costs in trial management. It also clarifies the bigger investigative picture while being able to measure the involvement of an individual in a quantifiable matter.

The disadvantages are that the analysis might be biased as the full extent of the information is not available; there is an issue of boundaries (whom to include in the network); and it is possible to foresee a lack of willingness to take advantage of SNA because of the paucity of digital skills in the sector. All these biases need to be further investigated with practical and qualitative research before its implementation. In order to facilitate the use of SNA techniques, educating law experts is crucial for its successful application into investigative methods and practice, as is already happening in some legal contexts around Europe.[64]

References

Ball, Leslie. "Automating social network analysis: A power tool for counter-terrorism." *Security Journal* 29, no. 2 (2016): 147-168.

Berg, Mark T., and Beth M. Huebner. "Reentry and the ties that bind: An examination of social ties, employment, and recidivism." *Justice quarterly* 28, no. 2 (2011): 382-410.

Berlusconi, Giulia. "Do all the pieces matter? Assessing the reliability of law enforcement data sources for the network analysis of wire taps." *Global Crime* 14, no.1 (2013): 61-81.

Berlusconi, Giulia, Francesco Calderoni, Nicola Parolini, Marco Verani, and Carlo Piccardi. "Link prediction in criminal networks: A tool for criminal intelligence analysis." *PloS one* 11, no. 4 (2016). doi:https://doi.org/10.1371/journal.pone.0154244.

Besaw, Lawrence M., Jeff C. Wu, Cho Y. Chang, Darren D. Smith, and Mark J. Kean. "Graphic display of network topology." U.S. Patent 5,276,789, filed May 14 1990 and issued January 4 1994.

Borgatti, Stephen P., Ajay Mehra, Daniel J. Brass, and Giuseppe Labianca. "Network analysis in the social sciences." *Science* 323, no. 5916 (2009): 892-895.

Bright, David A., Caitlin E. Hughes, and Jenny Chalmers. "Illuminating dark networks: A social network analysis of an Australian drug trafficking syndicate." *Crime, law and social change* 57, no. 2 (2012): 151-176.

Burcher, Morgan, and Chad Whelan. "Social network analysis as a tool for criminal intelligence: Understanding its potential from the perspectives of intelligence analysts." *Trends in organized crime* 21, no. 3 (2018): 278-294.

Calderoni, Francesco. "The structure of drug trafficking mafias: the 'Ndrangheta and cocaine." *Crime, law and social change* 58, no. 3 (2012): 321-349.

Campana, Paolo. "Explaining criminal networks: Strategies and potential pitfalls." *Methodological Innovations* 9, (2016).

Campana, Paolo, and Federico Varese. "Listening to the wire: criteria and techniques for the quantitative analysis of phone intercepts." *Trends in organized crime* 15, no. 1 (2012): 13-30.

Carrington, Peter J. (2011) "Crime and social network analysis." *The SAGE handbook of social network analysis*: 236-255.

Chaphalkar, N. B., and Sayali S. Sandbhor. "Application of neural networks in resolution of disputes for escalation clause using neuro-solutions." *KSCE Journal of Civil Engineering* 19, no. 1(2015): 10-16.

Colladon, Andrea Fronzetti, and Elisa Remondi. "Using social network analysis to prevent money laundering." *Expert Systems with Applications* 67 (2017): 49-58.

62 Author, interview with Public Prosecutor of Netherlands, September 2018.

63 This is true for most European countries. Excessive length of judicial proceedings is the main ground reason of applications in front of the European Court of Human Rights for violation of the article 6 of its treaty (CEPEJ, 2016). Council of Europe, European Commission for the Efficiency of Justice (CEPEJ). "*Report on European Judicial Systems, Efficiency and Quality of Justice.*" CEPEJ Studies no. 23 (2016).

64 Public prosecutors and judges in the Netherlands and in Spain have just begun to get training on social network analysis after their qualification.

Council of Europe, European Commission for the Efficiency of Justice (CEPEJ). "*Report on European Judicial Systems, Efficiency and Quality of Justice.*" CEPEJ Studies no. 23, (2016).

Davis, Roger H. "Social network analysis: An aid in conspiracy investigations." *FBI L. Enforcement Bull* 50, (1981): 11.

Duijn, Paul AC, and Peter PHM Klerks. "Social network analysis applied to criminal networks: recent developments in Dutch law enforcement." In *Networks and network analysis for defence and security*, pp. 121-159. Springer, Cham, 2014.

Duwe, Grant, and Valerie Clark. "Blessed be the social tie that binds: The effects of prison visitation on offender recidivism." *Criminal Justice Policy Review* 24, no. 3 (2013): 271-296.

Fabian, Seda. "Artificial Intelligence and the Law: Will Judges Run on Punch Cards." *Common L. Rev.* 16 (2020): 4.

Falcone, Giovanni, and Giuliano Turone. "Tecniche di indagine in materia di mafia." *Rivista di Studi e Ricerche sulla criminalità organizzata* 1, no. 1 (2015): 116-153.

Ficara, Annamaria, Lucia Cavallaro, Pasquale De Meo, Giacomo Fiumara, Salvatore Catanese, Ovidiu Bagdasar, and Antonio Liotta. "Social Network Analysis of Sicilian Mafia Interconnections." In *International Conference on Complex Networks and Their Applications*. Springer, Cham, 2019, 440-450.

Garofoli, Vincenzo. *Diritto processuale penale*. Giuffrè Editore, (2008).

Gravel, Jason, and George E. Tita. "With great methods come great responsibilities: social network analysis in the implementation and evaluation of gang programs." *Criminology & Pub. Policy* 14, no. 3 (2015): 559-527.

Hu, Paul Jen-Hwa, Hsinchun Chen, Han-fen Hu, Cathy Larson, and Cynthia Butierez. "Law enforcement officers' acceptance of advanced e-government technology: A survey study of COPLINK Mobile." *Electronic Commerce Research and Applications* 10, no. 1 (2011): 6-16.

Johnson, Jennifer, John David Reitzel, B. Norwood, D. McCoy, B. Cumming, and R. Tate. "Social network analysis: A systematic approach for investigating." *FBI Law Enforcement Bulletin,* (2013): 350. https://leb.fbi.gov/articles/featured-articles/social-network-analysis-a-systematic-approach-for-investigating.

Kelsen, Hans. *Pure Theory of Law*. Berkeley: University of California Press, 1967.

Kleemans, Edward R., and Christianne J. De Poot. "Criminal careers in organized crime and social opportunity structure." *European Journal of Criminology* 5, no. 1 (2008): 69-98.

Kossinets, Gueorgi. "Effects of missing data in social networks." *Social networks* 28, no. 3 (2006): 247-268.

Kossinets, Gueorgi, and Duncan J. Watts. (2006). "Empirical analysis of an evolving social network." *Science* 311, no. 5757, 88-90.

Krebs, Valdis E. "Mapping networks of terrorist cells. "*Connections* 24, no. 3 (2002): 43-52.

Lu, Yong, Xin Luo, Michael Polgar, and Yuanyuan Cao. "Social network analysis of a criminal hacker community. "*Journal of Computer Information Systems* 51, no. 2 (2010): 31-41.

Moreno, Jacob Levy. *The first book on group psychotherapy*. Psychodrama and Group Psychotherapy Monographs 1, Beacon House, 1957.

Moreno, Jacob Levy. *Application of the group method to classification*. New York: National Committee on Prisons and Prison Labor, 1932.

Morselli, Carlo. 2009. *Inside criminal networks* (Vol. 8). New York: Springer.

Musotto, Roberto. "Social and spatial network analysis of organised crime." PhD dissertation, Università degli Studi di Messina, 2016.

Musotto, Roberto, and David S Wall. "Are Booter Services (Stressors) indicative of a new form of organised crime group online?" *Journal of Digital Forensics* 1, no.1 (2019): 41-50.

Natarajan, Mangai. "Understanding the structure of a large heroin distribution network: A quantitative analysis of qualitative data." *Journal of Quantitative Criminology* 22, no. 2 (2006): 171-192.

Nøkleberg, Martin. "Examining the how of plural policing: Moving from normative debate to empirical enquiry." *The British Journal of Criminology*, (2020). doi: https://doi.org/10.1093/bjc/azz080

Petersen, Rasmus Rosenqvist, Christopher J. Rhodes, and Uffe Kock Wiil. "Node removal in criminal networks." In *2011 European Intelligence and Security Informatics Conference*, ed. IEEE Staff, 360-365. IEEE, (2011).

Piquero, Alex. R. ed. *The handbook of criminological theory*. John Wiley & Sons, 2015.

Points, Laurie, and Potton, Ed. "Artificial intelligence and automation in the UK." Briefing Paper: 8152 (2017).

Rutkin, Aviva. "The judge is a robot." New Scientist 222:2973, no. 24 (2014).

Scaglione, Antonio. *Reti mafiose. Cosa Nostra e Camorra: organizzazioni criminali a confronto*. FrancoAngeli, 2011.

Schroeder, Jennifer, (2001). Tucson Police Dept, and United States of America. "COPLINK: Database integration and access for a law enforcement intranet, final report." *Washington, DC: US Department of Justice*

Simmel, Georg. "The problem of sociology." *American Journal of Sociology* 15 no.3, (1909): 289-320.

Sourdin, Tania. (2018). "Judge v. Robot: Artificial Intelligence and Judicial Decision-Making." *UNSWLJ* 41, no. 4 (2018): 1114.

Sparrow, Malcolm. K. "The application of network analysis to criminal intelligence: An assessment of the prospects." *Social networks 13* no. 3 (1991): 251-274.

Sutherland, Edwin H., Donald R. Cressey, and David F. Luckenbill. *Principles of criminology*. Lanham: AltaMira Press, 1992.

Tonini, Paolo. *Manuale di procedura penale*. Giuffrè editore, 2012.

Tremblay, Pierre. "Searching for suitable co-offenders." In *Routine activity and rational choice* ed. R. V. G. Clarke and Marcus Felson, Routledge, 2017. 17-36.

Van der Hulst, Renee C. "Introduction to Social Network Analysis (SNA) as an investigative tool." *Trends in Organized Crime* 12, no. 2 (2009): 101-121.

Yongtae Park and Yoon, Byungun. "A text-mining-based patent network: Analytical tool for high-technology trend."*The Journal of High Technology Management Research*15, no. 1 (2004): 37-50.

5. ARTIFICIAL INTELLIGENCE IN HEALTHCARE: RISK ASSESSMENT AND CRIMINAL LAW

Federico Carmelo La Vattiata*

Abstract

The advances in the field of artificial intelligence (AI) are changing the nature of medical care. They involve both the sectors of diagnostics and therapeutics. Medical literature has widely analysed the advantages and the risks of AI. Researchers have found that early diagnoses are essential in order to avert the decline of patients' health status. This can be achieved through improving the analysis procedures on healthcare data by means of AI techniques. However, in order to guarantee the security of AI medical devices, their clinical evaluation is crucial.

This article aims at conceptualising and solving the questions related to the application of AI in healthcare from the point of view of criminal law. The traditional criminal law categories will be investigated, so as to understand whether it is possible to consider deaths and injuries occurring in the context of medical care as criminal offences to prevent and prosecute, when AI techniques are used. The study will be carried out in a comparative perspective. In conclusion, this will allow to propose a new AI-risk assessment paradigm, based on the integration of criminal law, civil law, and administrative law measures, so as to guarantee an equilibrium between the fundamental rights of the accused (a fair trial) and of the victims (a compensation for damages they have suffered).

Keywords: AI, healthcare, risk, criminal law.

Introduction

The advances in the field of artificial intelligence are changing the nature of medical care. AI refers to "systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals. AI-based systems can be purely software-based, acting in the virtual world [...] or AI can be embedded in hardware devices."[1] In other words, an *intelligent system* consists of a set of algorithms that can use data, to solve (more or less complex) problems in different contexts.

One should clarify a fundamental technique in the field of AI, i.e., *machine learning* (ML). The latter is based on complex mathematical techniques, since the related knowledge-representations involve the theory of probability and the statistics. The *artificial neural nets* (ANNs) are crucial elements. They are neural computational systems that are inspired by the functioning of the human brain, i.e., *biological neural nets* (BNNs). In particular, there are two similarities between them. Firstly, the building blocks of both nets are highly interconnected computational "tools". Secondly, ANNs consist in computing networks that are distributed in parallel and function like the varying synaptic strengths of the biological neurons: there are

1 * Federico Carmelo La Vattiata is a PhD Student in Comparative and European Studies (Criminal Law) at the University of Trento, Italy.
EU Commission. *Artificial intelligence for Europe*, 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, 2 (April 25, 2018).

many *input signals* to neurons, and the impact of each input is affected by the *weight* given to it, namely the adaptive coefficient within the network that determine the intensity of the input signal. In conclusion, "the output signal of a neuron is produced by the summation block, corresponding roughly to the biological cell body, which adds all of the weighted inputs algebraically."[2]

There are three kinds of ML:

1. the *supervised learning*, where the programmer *trains* the system by defining a set of expected results in relation to a certain in-put range, and by constantly evaluating the achievements of the objectives. The system then formulates a hypothesis. Every time it makes a mistake, the hypothesis is reviewed;

2. the *unsupervised learning*, where the user provides for neither expected results nor error-reports;

3. the *reinforcement learning*, where the system is led by a sort of *reward-punishment* mechanism, i.e., feedback messages about what has been done well or badly. In complex situations, the success or the failure of the system is reported after many decisions, and a sort of procedure for *assignment of credits* identifies the decisions that likely lead to success.

Another important technique is the so-called *deep learning* (DL), a subset of ML: it is a technique that allows the machines to better identify patterns through a more complicated use of neural nets. Because of this characteristic, it is possible to recognise data's patterns at various hierarchical levels. In other words, DL can identify a multilevel representation of knowledge.[3]

In brief, the ML algorithms use statistics and mathematics to find a pattern and correlations in big data.[4] The more a system processes data, the more it improves. Hence, the *detention* of big data is important. Very few persons and entities hold this new kind of *wealth*. This involves the emersion of new power, which is particularly difficult to control and limit by virtue of the traditional means. For this reason, new *checks and balances* are needed, so as to balance two interests: *a)* to promote the advantageous developments of AI; and *b)* to prevent abuses that can cause threats to individual rights.[5]

In general, there are four classification criteria of risks and opportunities resulting from the use of AI. When the AI is well used, it allows the human *self-regulation* and *agency*, as well as the improvement of societal *potential* and *cohesion*. Instead, when the AI is misused or overused, it reduces the *competences* of humans, it removes their *responsibilities*, and it undervalues their skills of *self-control* and *self-determination.*[6]

Ethics Guidelines for Trustworthy Artificial Intelligence

In 2018 the European Commission set up an independent High-Level Expert Group on Artificial Intelligence (AI HLEG) that published a document, entitled "Ethics Guidelines for Trustworthy Artificial Intelligence", with the aim to promote trustworthy AI.

According to the document, "trustworthiness is a prerequisite for people and societies to develop, deploy and use AI systems."[7] Trustworthy AI should be: *a) lawful*; *b) ethical*; and *c) robust*.

The document provides a framework for achieving trustworthy AI based on fundamental rights, that are enshrined in the Charter of Fundamental Rights of the European Union, and relevant international human

2 Young-Seuk Park and Sovan Lek, "Artificial Neural Networks: Multilayer Perceptron for Ecological Modelling," *Developments in Environmental Modelling*, 28 (2016): 124, https://doi.org/10.1016/B978-0-444-63623-2.00007-4.

3 Margaret A. Boden, *Artificial Intelligence. A Very Short Introduction* (Oxford: Oxford University Press, 2018), Italian trans. *L'intelligenza artificiale* (Bologna: Il Mulino, 2019), 46.

4 The definition of "big data" is: very large sets of data that are produced by people using the internet, and that can only be stored, understood, and used with the help of special tools and methods.
See https://dictionary.cambridge.org/it/dizionario/inglese/big-data.

5 Carlo Casonato, "Potenzialità e sfide dell'intelligenza artificiale," *BioLaw Journal*, 1 (2019): 178.

6 Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Cristoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, Burkhard Schafer, Peggy Valcke and Effy Vayena, "AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations," *Minds and machines* 28, 4 (2018): 689-707, https://doi.org/10.1007/s11023-018-9482-5.

7 EU High-Level Expert Group on AI, *Ethic Guidelines for Trustworthy AI*, 4 (April 8, 2019), https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

rights law, such as the European Convention on Human Rights. Furthermore, the document sets three series of key guidance. First, it "identifies the ethical principles and their correlated values that must be respected in the development, deployment and use of AI systems": *a) respect for human autonomy, prevention of harm, fairness and explicability*; *b)* the need for attention to situations involving more vulnerable groups, and to situations which are characterised by asymmetries of power or information; and *c)* acknowledgment of the risks that AI involves (including impacts which may be difficult to anticipate, identify or measure), and adoption of adequate measures to mitigate them. Secondly, guidance on how trustworthy AI can be realised is provided, and for this purpose seven key requirements that AI systems should meet are listed: *i)* human agency and oversight; *ii)* technical robustness and safety; *iii)* privacy and data governance; *iv)* transparency; *v)* diversity; non-discrimination and fairness; *vi)* environmental and societal well-being; and *vii)* accountability." Finally, the document provides an assessment list that will need to be tailored to the specific use case of the AI system.

As we will see in the following paragraph, AI can be widely applied in medicine, and its potential contributions to this field seem limitless. However, ethical challenges can arise, given that *intelligent* systems have a tremendous capability to threaten patients' preferences, safety and privacy. Benefits and risks due to AI application in healthcare need to be balanced carefully.[8] Furthermore, criminal law issues can arise with respect to the area of medical malpractice, such as black-box problems concerning the etiological link between the use of AI and damages, as well as difficulties in identifying precautionary rules whose culpable violation justifies convictions for gross negligence or recklessness.

AI in Healthcare

Experts started to debate the topic of difficult decisions in complex clinical situations assisted by computers in 1959.[9]

AI in healthcare involves both the sectors of diagnostics and therapeutics.[10] Applications are diverse, from mobile apps that make a diagnosis to surgical robots.[11] The advertising hyperbole of AI medical devices, however, "has led to skepticism and misunderstanding of what is and is not possible" with ML.[12] One should investigate the reasons for the doubts of credibility that affect the adoption of clinical decision support systems (CDSS). There are complexities that limit the ability to move ahead quickly. They reflect the ones of clinical practice: *a)* black boxes are unacceptable, since CDSS require transparency; *b)* time is a scarce resource, given that CDSS should be "efficient in terms of time requirements and must blend into the workflow of the busy clinical environment"; *c)* "complexity and lack of usability thwart use", as CDSS "should be intuitive and simple to learn and use"; *d)* "relevance and insight are essential", in fact "CDSS should reflect an understanding of the pertinent domain and the kinds of questions with which clinicians are likely to want assistance"; *e)* delivery of knowledge and information must be respectful, and CDSS "should offer advice in a way that recognizes the expertise of the user, making it clear that it is designed to inform and assist but not to replace a clinician"; and finally *f)* scientific foundation must be strong, as CDSS "should have rigorous, peer-reviewed scientific evidence establishing its safety, validity, reproducibility, usability, and reliability."[13]

8 Michael J. Rigby, "Ethical Dimensions of Using Artificial Intelligence in Health Care," *AMA J Ethics* 21, 2 (2019): 121-124, https://journalofethics.ama-assn.org/article/ethical-dimensions-using-artificial-intelligence-health-care/2019-02.

9 Robert S. Ledley and Lee B. Lusted, "Reasoning foundations of medical diagnosis; symbolic logic, probability, and value theory aid our understanding of how physicians reason," *Science* 130, 3366 (July 3, 1959); Edward H. Shortliffe and Martin J. Sepúlveda, "Clinical Decision Support in the Era of Artificial Intelligence," *JAMA* 320, 21 (December 4, 2018): 2199.

10 Within the *genus* "therapeutics", three *species* can be distinguished: *a)* medical therapy; *b)* surgical therapy; *c)* mixed therapy (e.g., some neoplasia needs to be treated by means of: first, an oncological intervention, so as to reduce the *critical mass*; secondly, a surgical intervention to remove it; and finally, another oncological intervention in order to prevent the risk of metastasis).

11 However, recent studies have compared the level of diagnostic efficiency of humans and AI systems. The outcome was the humans' *victory*. The spread between the samples is directly proportional to the rarity of diseases: the more a disease is atypical, the more humans take advantage of their personal skills, since the clinical question cannot be simply solved by means of statistics. Instead, when a disease is common, AI systems can use their (potentially boundless) skills in data processing. See Hannah L. Semigran, David M. Levine, Shantanu Nundy and Ateev Mehrotra, "Comparison of Physician and Computer Diagnostic Accuracy," *JAMA Internal Medicine* 176, 12 (2016): 1860-1861, https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/2565684.

12 Suchi Saria, Atul Butte and Aziz Sheikh, "Better medicine through machine learning: What's real, and what's artificial?," *PloS Med* 15, 12 (2018): 1, https://doi.org/10.1371/journal.pmed.1002721.

13 Edward H. Shortliffe and Martin J. Sepúlveda, *Ibidem*.

Although human physicians cannot be replaced by AI systems in the foreseeable future, AI could play a key-role in assisting physicians to make better clinical decisions. In some cases, *intelligent* systems could even replace human judgement in certain fields of healthcare.

As a matter of fact, a large volume of healthcare data can be computed efficiently by AI algorithms, in order to assist clinical practice. *Intelligent* systems can be equipped with learning and self-correcting skills to improve their accuracy, and support physicians in reducing diagnostic and therapeutic errors. Moreover, AI systems can be used so as to extract information from a large patient population and assist in making real-time inferences for a health risk alert and outcome prediction.[14]

AI medical devices can be classified into two categories: *a)* ML techniques that analyse healthcare data (e.g., imaging, genetic and EP data) in order to cluster patients' characteristics or infer the probability of the outcome of certain diseases; *b)* natural language processing (NLP) methods that process unstructured data (e.g., clinical notes and medical journals) in order to develop structured medical data. In particular, DL is very performing in the interpretation of data in the form of images, by virtue of the complexity of factors that it can take into account.

Medical literature has widely discussed the advantages of AI, mainly regarding three disease types: cardiovascular disease, cancer and nervous system disease. These are the leading causes of death. Thus, early diagnoses are essential in order to avert the decline of patients' health status. This can be achieved through improving the analysis procedures on healthcare data by means of AI techniques.

Having said this, several challenges still need to be faced. In the first place, current regulations lack standards to assess the AI systems' safety and efficacy. Furthermore, scientists deal with problems of data-exchange. From this point of view, one should consider that *intelligent* systems need to be constantly trained by clinical data: the first training can be based on historical datasets; however, the following steps require a continuation of the information. This is a crucial issue for the development and improvement of the system.[15]

Criminal Law Questions Raised by the Use of AI in Healthcare

Many questions in the field of criminal law arise from the development of software as a medical device (SaMD). In brief, we need to clarify:

1. whether AI systems must be considered as *agents*, in as much as they can be *legal persons*, or mere *instruments*, through which humans might commit crimes;

2. how AI crimes (AIC) – i.e. crimes involving an AI system – can be performed in the field of medical devices, namely whether they are crimes typically based on specific conduct or requiring a specific event to occur;[16] also, in the latter case, how we can solve the questions concerning the etiological link between the agent's conduct and the event, and the varieties of fault; and

3. finally, whether a new fairer model may be theorised.

AI Entities as Legal Persons

Gabriel Hallevy theorised possible forms of AI systems' criminal liability based on the attribution to them of a legal personality. He postulated three models:

1. the *perpetration-through-another* model, where the AI system is considered as an *innocent agent*, a mere *instrument* used by the actual perpetrator (*principal*), i.e. the programmer or the user;

14 Fei Jiang, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen and Yongjun Wang, "Artificial intelligence in healthcare: past, present and future," *Stroke and Vascular Neurology* 2, 4 (2017): 230, http://dx.doi.org/10.1136/svn-2017-000101.

15 Fei Jiang *et al.*, *Ivi*, 241.

16 Thomas C. King, Nikita Aggarwal, Mariarosaria Taddeo and Luciano Floridi, "Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions," *Sci. Eng. Ethics* 26, (2020): 90-91, https://doi.org/10.1007/s11948-018-00081-0.

2. the *natural-probable-consequence* model, where programmers/users may be held criminally liable for a crime committed via AI and occurring as a natural and probable consequence of their intentional or negligent behaviour;

3. the *direct liability* model, where it is assumed that AI is endowed with *mens rea*, and therefore compatible with a burden of liability for conduct materially executed by itself.

In the last model, Hallevy hypothesises a possible application of punishment constructed according to a principle of equivalence between machine and man. It would be a matter of eliminating the software, intending to neutralise the system, or deactivating it for a pre-established length of time in order to encourage its re-education. Yet, Hallevy's third model cannot be accepted. It is based on vitiated arguments.

First, AI systems are not actually *intelligent*. In law, speculation without scientific evidence cannot constitute a valid reference. As Luciano Floridi, a professor at the Internet Oxford Institute, argues, the best definition of AI is still the one provided by John McCarthy in 1955: the AI problem "is taken to be that of making a machine behave in ways that would be called intelligent if a human were so behaving".[17] Thus, we can call such a behaviour *intelligent* in as much as a human behaves in that way, but it does not mean that the machine is intelligent.[18]

Hallevy argues that the objections against the criminal liability of AI entities (above all concerning the requirement of *mens rea*) are based on arguments that are similar to the ones relating to the liability of corporations. Then, there would be "no substantial legal difference between the idea of criminal liability imposed on corporations and on AI entities." Yet, there is a substantial difference: "under the [...] 'superior agent' rule, corporate criminal liability [...] is limited to situations in which the conduct is performed or participated in by the board of directors or a high managerial agent."[19] Instead, Hallevy's third model "does not assume any dependence of the AI entity on a specific programmer or user."[20]

Finally, criminal responsibility is based on the two crucial concepts of *wrongdoing* and *attribution,*[21] which presuppose two requisites:

1. a *human* (and *voluntarily taken*) act or omission[22] (also known as *actus reus*);[23]

2. the *mens rea* (varieties of fault), i.e., the act/omission needs to be *covered* by a guilt mind.

AI Entities as Instruments: Responsibility of the Producers and the Users

We have clarified that AI systems are not *subjects*. Instead, they have to be considered as *instruments*, through which the actor can commit a crime. In the field of AI medical devices, we can distinguish two types of offences: the ones committed by the *producers* and the ones committed by the *users*.

---

17 John McCarthy, Marvin L. Minsky, Nathaniel Rochester and Claude E. Shannon, "A proposal for the Dartmouth Summer Research Project on Artificial Intelligence; August 31, 1955," *AI Magazine* 27, 4 (2006): 12, https://doi.org/10.1609/aimag.v27i4.1904.

18 Luciano Floridi, "Digital's Cleaving Power and Its Consequences," *Philos. Technol.* 30, (2017): 123-129, https://doi.org/10.1007/s13347-017-0259-1.
The author argues that if one affirmed that the machine is *intelligent*, it would be "a fallacy that smacks of superstition (compare: the river reaches the lake by following the best possible path, removing obstacles in its way; if this had been done by someone, we would have considered that behaviour intelligent; so the river's behaviour is intelligent)."

19 Wayne R. LaFave and Austin W. Scott Jr., *Substantive Criminal Law*, Volume 1 (St. Paul – Minnesota: West Publishing Co., 1986), 360.

20 Gabriel Hallevy, "The Criminal Liability of Artificial Intelligence Entities - from Science Fiction to Legal Social Control," *Akron Intellectual Property Journal* 4, 2, Article 1 (2010): 186, http://ideaexchange.uakron.edu/akronintellectualproperty/vol4/iss2/1.

21 An act is *wrongful* when it satisfies the definition of an offence and is unjustified. The definition of an offence is composed by a set of (objective and subjective) elements that constitutes the so-called *incriminating case* against the charged person. It is provided for in a *basic* norm, i.e., a norm, directed to the citizens, which *prohibits* particular acts or *requires* particular acts to be performed. So, they impose *duties of compliance* on individuals. Therefore, the concept of *wrongful conduct* is formal, being defined by the incompatibility of the act with respect to the norms of the legal system. Instead, the concept of *wrongdoing* is substantial: to say that the perpetrator is a *wrongdoer* (or that he/she engaged in wrongdoing) is to pass judgment on the intrinsic quality of his/her deeds. The concept of *attribution* can be referred to in two dimensions:
objective attribution is the term used in Germany to qualify the general process of holding individuals accountable for the occurrence of harm or the acts of other persons; and
subjective attribution, instead, refers to the distinct question of the *criteria for holding persons accountable* for their deeds.
With particular reference to *subjective* attribution, it is based on norms other than *basic* ones, directed (not to the citizens but) to the judge. These rules do not generate exceptions to the basic norms, they only *excuse* their violation.
In this regard, see *amplius* the fundamental work by George P. Fletcher, *Rethinking Criminal Law* (Oxford/New York: Oxford University Press, 2000), 454-491.

22 George P. Fletcher, *Ivi*, 475.

23 Thomas C. King *et al.*, *Ibidem*.

Producers can be found liable for having committed crimes either within the pattern of *manifest criminality*, or within the pattern of *harmful consequences*. In the first case, "the significance of the criminal act is that it manifests criminality and unnerves the community's sense of security."[24] In other words, what is merely punished is that the actor puts in *danger* a legal interest (e.g., life, or public health). These offences are forms of *direct* liability that can consist of either *active* or *omissive* conduct (the breach of a specific statutory duty to act), and are characterised by considerable anticipation of the threshold to trigger criminal law protection. In this respect, we may generally distinguish:

1. the responsibility for the *type* of production, when the perpetrator develops certain types of AI devices *absolutely prohibited* by the law, since they involve a high risk of causing harmful events, that cannot be (substantially) reduced through the respect of some precautionary measures; and

2. the responsibility for the *modality* of production, that relates to areas of production (*allowed* by the law) characterised by a lower risk-coefficient, since the available nomological skills permit to set precautionary measures in order to prevent/reduce the risk of harmful events.[25] Thus, producers can be held accountable for having produced AI devices without complying with the aforementioned precautionary rules.

Furthermore, they can be found guilty of a crime within the pattern of *harmful consequences*, e.g., manslaughter, in as much as, within the productive cycle of an AI device, they have not respected the precautionary rules mentioned above, and, because of this violation, they have caused the *actual occurrence of a harm* to the legal interest protected by the law (individuals' life), such as a man/woman's death.

With regard to the *users*, namely physicians, they can commit crimes within the aforementioned patterns too. As for offences within the pattern of *manifest criminality*, in the field of SaMD, these forms of responsibility can consist in violating the ban of merely using devices with certain characteristics, i.e., involving a risk-level that is considered excessive and unacceptable by the law. As for crimes within the pattern of *harmful consequences*, physicians can be found liable for having committed manslaughter due to *medical malpractice*. These offences could fulfil forms of:

1. *direct* liability, when the physician commits *active* deeds, i.e., he/she uses a SaMD, and, as a *direct* consequence, the patient dies; and

2. *derivative* liability, when he/she is responsible for a so-called *commission by omission*.

Indeed, the paradigm of *commission by omission* is based on the criteria of the so-called *duty to avert the harmful event*. The duty to act – a *legal* duty, not merely a *moral* duty – can arise according to:

1. a statute ("other than the criminal statute whose violation is in question");[26]

2. a personal relationship, e.g., between the physician and the patient;

3. a contract;

4. the voluntary assumption of care; and

5. the creation of a risk by the defendant.

In other words, the physician can be held accountable for the event that is to be averted only if – in the light of a special *protection-link* –, in a particular situation, he/she can be considered as the *defender* (the *guarantor*) of a certain legal interest, namely a patient's life.[27] For instance: a physician used a SaMD; he/she had the *duty to control* it, so as to prevent (*as much as possible*) the occurrence of harmful outcomes; nevertheless, he/she failed in doing it (either because he/she *absolutely* did not control the *causal progress* activated by the device's use, or because he/she *erroneously* controlled it, namely not complying with the

24 George P. Fletcher, *Ivi*, 420.

25 Franco Bricola, "Responsabilità penale per il tipo e per il modo di produzione," in *Scritti di diritto penale*, 1, II, ed. Stefano Canestrari and Alessandro Melchionda (Milano: Giuffrè, 1997); Carlo Piergallini, *Danno da prodotto e responsabilità penale: profili dommatici e politico-criminali* (Milano: Giuffrè, 2004), 40-46.

26 Wayne R. LaFave and Austin W. Scott Jr., *Ivi*, 286.

27 This element is common to several legal systems. In this regard, one could mention the Italian notion of "*posizione di garanzia*", or the German concept of "*Garantenstellung*".

applicable guidelines). According to the prevailing view, the perpetrator must know the facts indicating a duty to act, and the jurisprudence retains that the imposition of strict liability in omission cases is inappropriate.[28] Certainly, "sometimes there may be a duty to take care to know the facts, as well as a duty to go into action when the facts are known."[29] Also, the actor must be physically capable of performing the actions necessary to avert the harmful event.[30] In the pattern of *harmful consequences*, when an event occurs due to the use of AI systems, a problem might arise with reference to *causation*. As things stand, we are not able to identify the etiological links within the *neural nets* in terms of certainty. In other words, one can affirm that an AI system causes a certain event *y* (e.g., a diagnosis), from a certain data-set *x* (e.g., some tests on the patient). However, the reconstruction of every single *intermediate* step through which the system reaches that result (especially in cases of DL) is currently impossible.

Nonetheless, by means of the counterfactual analysis, judges only need to establish whether a certain conduct is a contingently essential condition of the event. They never achieve a result that is *deductively sure*. Indeed, their reasonings are almost always *probabilistic* (i.e., rationally plausible), due to the lack of *explicative preconditions* and to the use of *statistical laws* in the explanation of naturalistic events.[31] In this sense, as a matter of common knowledge, the actor's conduct, namely the use of a SaMD, must be the *condicio sine qua non* of the harmful event,[32] or that *but for* the conduct, the event would not have occurred.[33]

In *commission-by-omission* cases, the judge must:

1. reconstruct the omitted action,[34] i.e., the (correct) supervision of the device that would have avoided the deterioration of a patient's health condition;

2. reconstruct the *causal process*, as it actually occurred, i.e., how and why the harmful event occurred; and

3. clarify – in the light of an established scientific law – whether the actor's (*hypothetical*) conduct would have affected the causal process averting the harmful event.

A point must be stressed: the standard of proof expressed by the formula *beyond any reasonable doubt* (*alias* "BARD rule") arises from the consideration that "it is far worse to convict an innocent man than to let a guilty man go free."[35] Therefore, in this field, an established scientific *covering-law* is essential to prove the etiological process and then the actor's liability. Otherwise, the accused physician must be declared *not guilty*.

In order for the crime to be fulfilled, the perpetrator must meet the so-called *mens rea* requirements. In this regard, the offences can be basically committed in the light of the following varieties of fault:[36]

---

28 *Harding v. Price*, [1948] 1 K.B. 695.

29 Wayne R. LaFave and Austin W. Scott Jr., *Ivi*, 290-291.

30 Wayne R. LaFave and Austin W. Scott Jr., *Ivi*, 291.

31 Federico Stella, "La nozione penalmente rilevante di causa: la condizione necessaria," *Rivista Italiana di Diritto e Procedura Penale*, 4, (1988): 1217.

32 For the purpose of criminal law, one can affirm that there is an etiological link between an antecedent (e.g., a – wrong – diagnosis obtained by means of an AI system) and an event (e.g., the patient's death) when two elements subsist, namely:
an established (universal or statistical) covering-law, in the light of which the causal process in question can be explained; and
the agent's conduct (i.e., the use of an AI system) is an *essential* condition *in all the feasible (or probable) explanations*.

33 The American Law Institute, *Model Penal Code*, § 2.03(1) (May 24, 1962).

34 Giovanni Grasso, *Il reato omissivo improprio: La struttura obiettiva della fattispecie* (Milano: Giuffrè, 1983), 370-371.

35 *In re Winship*, 397 U.S. 358 (1970); *Barnes v. United States*, 412 U.S. 873 (1973); *Carella v. California*, 491 U.S. 263; *Albright v. Oliver et al.*, 510 U.S. 266 (1994); *Victor v. Nebraska*, 511 U.S. 1 (1994); *United States v. Gaudin*, 515 U.S. 506; *Spencer v. Kemna*, 523 U.S. 1 (1998). As for the jurisprudence of the Italian Supreme Court see above all: Cassazione Penale, Sezioni Unite, no. 30328 *Franzese* (2002).

36 One might observe that the various legal systems adopt different solutions concerning the varieties of fault. As a matter of fact, this consideration is only partially true. Instead, the comparative analysis demonstrates that, regardless of terminological differences, substantial similarities are often more than dissimilarities. Sure, solutions are rarely the same. Nevertheless, modern jurists' task – living in a globalised world – is to *build bridges not walls*, to find points of contact between different legal traditions, especially in areas – such as AI – where criminality (and its prevention and punishment) no longer concerns one State solely, but requires stronger mechanisms of judicial cooperation. From this point of view, one might refer to the rules and principles set in the Rome Statute of the International Criminal Court (where applicable). Sure, the ICC exercises its jurisdiction over crimes that are other than the ones we are dealing with, i.e., artificial intelligence crimes ("AIC") in the field of healthcare. However, some definitions – especially the ones concerning the so-called *general part* of criminal law – may represent valid references. As a matter of fact, the Statute's drafters (most of which were comparative law experts) made the effort to formulate definitions that could have been a sort of synthesis between the various legal traditions.
In this regard, for example, one might mention the definition of *mens rea*, based on the elements of *intent* and *knowledge*, pursuant to article 30: "[...] a person has intent where: (a) In relation to conduct, that person means to engage in the conduct; In relation to a consequence, that person means to cause that consequence or is aware that it will occur in the ordinary course of events. [...] 'knowledge' means awareness that a circumstance exists or a consequence will occur in the ordinary course of events [...]."

1. intention;
2. knowledge;
3. recklessness;
4. negligence.

Among these varieties of fault, *recklessness* and *negligence* are the most questionable, relating to the notion of *risk* (that must be *substantial* and *unjustifiable*).[37] The actor – respectively – consciously has disregarded it, or should have been aware of it.[38] The threshold of substantiality and unjustifiability of the risk is due to a *gross* deviation from the standard of conduct/care[39] that a *law-abiding/reasonable person*[40] would respect in the actor's situation.[41]

With regards to the aforementioned standard, some rules of care are crucial. As for *producers*, norms on the rigorous clinical validation of AI medical devices can be cited.[42] Instead, the standard of care for *users*, i.e., physicians, guidelines and best practices on healthcare are the main reference. However, in both cases, the judge must investigate whether, in the specific situation, the rule of care that the actor violated was aimed at preventing harmful events of the type of the one that occurred.[43]

A New Paradigm

The abovementioned liability paradigm (still in effect) does not allow to realise the simultaneous *protection of the innocent and the victims.*[44] Another paradigm can be theorised in a *de lege ferenda* perspective.

37 The American Law Institute, *Model Penal Code*, § 2.02(2)(d): "A person acts negligently with respect to a material element of an offense when he should be aware of a substantial and unjustifiable risk that the material element exists or will result from his conduct. The risk must be of such a nature and degree that the actor's failure to perceive it, considering the nature and purpose of his conduct and the circumstances known to him, involves a gross deviation from the standard of care that a reasonable person would observe in the actor's situation."

38 In this respect, we should distinguish three elements: *a)* the *recognisability* ("*Erkennbarkeit*" in Germany, "*riconoscibilità*" in Italy) of the unreasonable/unjustifiable risk; *b)* its *foreseeability* ("*Vorhersehbarkeit*" in Germany, "prevedibilità" in Italy); and *c)* its *preventability* ("*Vermeidbarkeit*" in Germany, "*evitabilità*" in Italy).

39 The notion of "gross deviation from the standard of care" expresses what is commonly understood as the *something extra* that distinguishes the *criminal* (or *gross*) negligence and the *ordinary* negligence, which is relevant in the area of tort law solely.
Analogous (albeit not identical) concepts can be found in the German and Italian legal systems: i.e., the notions of – respectively – "*Leichtfertigkeit*" and "*colpa grave*".

40 Wayne R. LaFave and Austin W. Scott Jr., *Ivi*, 328: "Thus negligence is framed in terms of an objective (sometimes called 'external') standard, rather than in terms of a subjective standard."

41 In this regard, see also Wayne R. LaFave and Austin W. Scott Jr., *Ivi*, 327-328: "'Unreasonable risk" is an expression which takes into account the fact that we all create some risk to others in our everyday affairs without subjecting ourselves to liability for negligence. [...] the test for reasonableness in creating risk is thus said to be determined by weighing the magnitude of the risk of harm against the utility of the actor's conduct. Under such a test, even a slight risk may be unreasonable. [...] Aside from the utility of the actor's conduct, another variable factor involved in the question of whether a particular risk is unreasonable is the extent of the actor's knowledge of the facts bearing upon the risk. [...] Still another variable concerns the nature and the extent of the harm which may be caused by the defendant's conduct, and the number of persons who may be harmed."

42 The International Medical Device Regulators Forum (IMDRF) – of which the WHO is an official observer within the management committee –, in particular the SaMD Working Group, has developed a series of documents in order to provide harmonized principles for individual jurisdictions to adopt based on their own regulatory framework.
The US Food & Drugs Administration (FDA) has adopted these principles. In this respect, see IMDRF SaMD Working Group, *Software as a Medical Device (SaMD): Key Definitions* (December 9, 2013); Id., *"Software as a Medical Device": Possible Framework for Risk Categorization and Corresponding Considerations* (September 18, 2014); Id., *Software as a Medical Device (SaMD): Application of Quality Management System* (October 2, 2015); Id., *Software as a Medical Device (SaMD): Clinical Evaluation* (June 22, 2017).
In Europe, on the 5th April 2017 the European Parliament and the Council of the European Union has adopted the Regulation (EU) 2017/745 *on medical devices*. This "lays down rules concerning the placing on the market, making available on the market or putting into service of medical devices, for human use and accessories for such devices in the Union. This Regulation also applies to clinical investigations concerning such medical devices and accessories conducted in the Union" (article 1.1). The abovementioned guidance developed for medical devices at the international level have been taken into consideration by the EU too, so as "to promote the global convergence of regulations which contributes to a high level of safety protection worldwide, and to facilitate trade" (recital 5). In this way, the results of clinical investigations conducted in the EU would be accepted as documentation outside the EU, and those conducted outside the EU in accordance with international guidelines would be accepted within the EU. In this regard, the rules should be in line with the most recent version of the World Medical Association (WMA) Declaration of Helsinki *on ethical principles for medical research involving human subjects* (recital 64).

43 In Germany and Italy this concept is understood as, respectively, "*Risikozusammenhang*" and "*nesso di rischio*". These expressions' translation might be the *risk-link* between the violation of the duty of care (by the actor) and the harmful event, whose nature is normative, not material.

44 Federico Stella, *Giustizia e modernità: La protezione dell'innocente e la tutela delle vittime* (Milano: Giuffrè, 2003).

In the comparative law-panorama, the U.S. model of assessment of technological risks can be a good reference. It is based on the central role of public agencies (e.g., the Food and Drugs Administration) and on five elements:[45]

1. The agency provides for specific *precautionary rules* that, for example, the companies involved in the production/use of medical devices must respect. These rules set risk-assessment *standards*, i.e. the adequacy threshold of a risk;

2. The adoption of these precautionary rules is *democratically legitimated* (i.e., the Congress approves them);

3. The courts review the rationality of the agencies' decisions, i.e., courts of appeals may invalidate them in case they are *arbitrary* or *capricious*, or not supported by *substantial evidence;*[46]

4. A rigid and preventive *enforcement* system, i.e., agencies may make *inspections* of the companies and issue *injunctions* to them;

5. An *education and compliance assistance* system, i.e., the companies can either ask for the agency's *consultation assistance* in arranging an adequate set of cares able to prevent harmful events or accept a *Voluntary Protection Program* in the light of which the agency itself implements a permanent *on-site analysis* on the internal safety-system.

From a broader perspective, international harmonisation/cooperation is desirable in this field. The World Health Organization (WHO) might set general standards, that national – or hopefully supranational agencies, such as the European Medicines Agency (EMA) – should (not might) subsequently implement into more specific guidelines for producers and users.

As for sanctions, a multi-level system can be theorised. Above all, we should distinguish sanctions for the liability of organisations and sanctions for the liability of individuals.

As for the former, two levels can be hypothesised:

1. Administrative sanctions (imposed by an agency) for minor violations of the precautionary rules concerning the production/use of a SaMD;

2. Criminal sanctions (imposed by the courts) for serious violations of the aforementioned precautionary rules and, eventually, for causing harmful events to the patients because of such violation.

As for the individuals, the view is necessarily different: in a field characterised by *scientific uncertainty*, they might be unconscious that, through their deeds, they create a substantial and unjustifiable risk. Thus, their liability should be limited to conducts covered by the psychological coefficients of *intent* and, at most, of *recklessness*. *Negligent* deeds should remain, in this field, immune from penal sanctions.[47]

Finally, two considerations need to be pointed out. First, regardless of whether the accused is an organisation or an individual, the evidentiary and judgment rule in criminal trials is the one expressed by the formula *beyond any reasonable doubt*. Then, in cases where the prosecutors cannot present sufficient evidence so as to satisfy the aforementioned standard (either because they do not find it, or because – especially with reference to individuals' liability – it does not exist), victims can follow the *path* of tort law, in order to obtain fair compensation for the damages they suffered. In this regard, the view of those who theorise definitions of criminal liability without fault requirement cannot be accepted, and *mens rea* is so crucial to the State's entitlement to sanction individuals depriving them of their freedom that we cannot merely abandon it only because of difficulty in proving it.

45 Sheila Jasanoff, *Science at the Bar: Law, Science, and Technology in America* (Cambridge – Massachusetts/London – England: Harvard University Press, 1995), 69-92; Francesco Centonze, *La normalità dei disastri tecnologici: Il problema del congedo dal diritto penale* (Milano: Giuffrè, 2004), 400-410.

46 *US Administrative Procedure Act* (1946), Section 10.

47 Thomas C. King *et al.*, *Ivi*, 95: "Concerning the knowledge threshold, in some cases the *mens rea* could actually be missing entirely. The potential absence of a knowledge-based *mens rea* is due to the fact that, even if it is understood that an AA [artificial agent] can perform the *actus reus* autonomously, the complexity of the AA's programming makes it possible that the designer, developer, or deployer (i.e., a human agent) will neither know nor predict the AA's criminal act or omission. The implication is that the complexity of AI provides a great incentive for human agents to avoid finding out what precisely the ML [machine learning] system is doing, since the less the human agents know, the more they will be able to deny liability for both these reasons."

Conclusions

To sum up, we have seen that AI applications in healthcare entail both benefits and risks, which need to be balanced. The search for an equilibrium between them involves considerations concerning not only the field of medicine, but also ethics and law.

In particular, the protection of human rights deserves attention. Among them, in the field of criminal law, one should stress the interconnection that exists between, on the one hand, the principles of *legality* (*nullum crimen sine lege*) and of *culpability* (*nullum crimen sine culpa*), and, on the other hand, their processual *pendant*, namely the presumption of innocence and the BARD rule (a rule relating both to the admission and evaluation of evidence, and to the verdict on the defendant's guilty). From the point of view of the accused, the observance of these principles is crucial so as to guarantee the fundamental right to a fair trial.[48] Nevertheless, patients-victims deserve *justice* in case of damages caused by mistakes in applying AI techniques to their clinical situation. We have seen how this simultaneous *protection of the innocent* and *of the victims* can be reached. In the first place, one should not confuse the liability of *producers*, and the one of *physicians*. The former can be found responsible for having committed offences either within the pattern of *manifest criminality* (i.e., they violate the *absolute* ban of producing certain types of devices, or they develop devices without complying with the applicable precautionary rules) or within the pattern of *harmful consequences* (in so far as the violation of the precautionary is etiologically linked to a man/woman's death). Instead, the case of physicians' liability is essentially a matter of medical malpractice. Then, from a different point of view, one should distinguish the cases of liability of *individuals* and *legal persons*, given that several legislations can differently apply to each of them. Finally, in a *de lege ferenda* perspective, we have seen that a new paradigm can be theorised. It is inspired to the US model of risk assessment, and involves the integration of criminal law, civil law, and administrative law measures. I think that this model could guarantee an equilibrium between the fundamental rights of the accused (a fair trial) and of the victims (a compensation for damages they have suffered).

References

Boden, Margaret A. *Artificial Intelligence. A Very Short Introduction*. Oxford: Oxford University Press, 2018 (Italian translation, *L'intelligenza artificiale*. Bologna: Il Mulino, 2019).

Bricola, Franco. "Responsabilità penale per il tipo e per il modo di produzione." In *Scritti di diritto penale*, 1, II, edited by Stefano Canestrari and Alessandro Melchionda. Milano: Giuffrè, 1997.

Centonze, Francesco. *La normalità dei disastri tecnologici: Il problema del congedo dal diritto penale*. Milano: Giuffrè, 2004.

Fletcher, George P. *Rethinking Criminal Law*. Oxford/New York: Oxford University Press, 2000.

Grasso, Giovanni. *Il reato omissivo improprio: La struttura obiettiva della fattispecie*. Milano: Giuffrè, 1983.

Grasso, Giovanni. "La protezione dei diritti fondamentali nella Costituzione per l'Europa e il diritto penale: spunti di riflessione critica." in *Lezioni di diritto penale europeo*, edited by Giovanni Grasso and Rosaria Sicurella. Milano: Giuffrè, 2007.

Jasanoff, Sheila. *Science at the Bar: Law, Science, and Technology in America*. Cambridge – Massachusetts/London – England: Harvard University Press, 1995.

LaFave, Wayne R., and Austin W. Scott Jr. *Substantive Criminal Law*. Volume 1. St. Paul – Minnesota: West Publishing Co., 1986.

Piergallini, Carlo. *Danno da prodotto e responsabilità penale: profili dommatici e politico-criminali*. Milano: Giuffrè, 2004.

Stella, Federico. *Giustizia e modernità: La protezione dell'innocente e la tutela delle vittime.* Milano: Giuffrè, 2003.

Casonato, Carlo. "Potenzialità e sfide dell'intelligenza artificiale." *BioLaw Journal*, 1 (2019): 177-182.

Floridi, Luciano, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Cristoph Luetge, *et al.* "AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations." *Minds and machines* 28, 4 (2018): 689-707. https://doi.org/10.1007/s11023-018-9482-5.

48 Giovanni Grasso, "La protezione dei diritti fondamentali nella Costituzione per l'Europa e il diritto penale: spunti di riflessione critica," in *Lezioni di diritto penale europeo*, ed. Giovanni Grasso and Rosaria Sicurella (Milano: Giuffrè, 2007), 656-659.

Floridi, Luciano. "Digital's Cleaving Power and Its Consequences." *Philos. Technol.* 30, (2017): 123-129. https://doi.org/10.1007/s13347-017-0259-1.

Hallevy, Gabriel. "The Criminal Liability of Artificial Intelligence Entities - from Science Fiction to Legal Social Control." *Akron Intellectual Property Journal* 4, 2, Article 1 (2010): 186. http://ideaexchange.uakron.edu/akronintellectualproperty/vol4/iss2/1.

Jiang, Fei, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen and Yongjun Wang. "Artificial intelligence in healthcare: past, present and future." *Stroke and Vascular Neurology* 2, 4 (2017): 230. http://dx.doi.org/10.1136/svn-2017-000101.

King, Thomas C., Nikita Aggarwal, Mariarosaria Taddeo and Luciano Floridi. "Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions." *Sci. Eng. Ethics* 26, (2020): 90-91. https://doi.org/10.1007/s11948-018-00081-0.

Ledley, Robert S., and Lee B. Lusted. "Reasoning foundations of medical diagnosis; symbolic logic, probability, and value theory aid our understanding of how physicians reason." *Science* 130, 3366 (July 3, 1959);

McCarthy, John, Marvin L. Minsky, Nathaniel Rochester and Claude E. Shannon. "A proposal for the Dartmouth Summer Research Project on Artificial Intelligence; August 31, 1955." *AI Magazine* 27, 4 (2006): 12. https://doi.org/10.1609/aimag.v27i4.1904.

Park, Young-Seuk and Sovan Lek. "Artificial Neural Networks: Multilayer Perceptron for Ecological Modelling." *Developments in Environmental Modelling*, 28 (2016): 124. https://doi.org/10.1016/B978-0-444-63623-2.00007-4.

Rigby, Michael J. "Ethical Dimensions of Using Artificial Intelligence in Health Care." *AMA J Ethics* 21, 2 (2019): 121-124. https://journalofethics.ama-assn.org/article/ethical-dimensions-using-artificial-intelligence-health-care/2019-02.

Saria, Suchi, Atul Butte and Aziz Sheikh. "Better medicine through machine learning: What's real, and what's artificial?." *PloS Med* 15, 12 (2018): 1. https://doi.org/10.1371/journal.pmed.1002721.

Semigran, Hannah L., David M. Levine, Shantanu Nundy and Ateev Mehrotra. "Comparison of Physician and Computer Diagnostic Accuracy." *JAMA Internal Medicine* 176, 12 (2016): 1860-1861. https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/2565684.

Shortliffe, Edward H., and Martin J. Sepúlveda. "Clinical Decision Support in the Era of Artificial Intelligence." *JAMA* 320, 21 (December 4, 2018): 2199.

Stella, Federico. "La nozione penalmente rilevante di causa: la condizione necessaria." *Rivista Italiana di Diritto e Procedura Penale*, 4, (1988): 1217-1268.

Albright v. Oliver et al., 510 U.S. 266 (1994).

Barnes v. United States, 412 U.S. 873 (1973).

Carella v. California, 491 U.S. 263.

Cassazione Penale, Sezioni Unite. no. 30328 *Franzese* (2002).

E.U. Commission. *Artificial intelligence for Europe*, 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, 2 (April 25, 2018).

E.U. High-Level Expert Group on AI. *Ethic Guidelines for Trustworthy AI*, 4 (April 8, 2019). https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Harding v. Price, [1948] 1 K.B. 695.

IMDRF SaMD Working Group. *"Software as a Medical Device": Possible Framework for Risk Categorization and Corresponding Considerations* (September 18, 2014).

IMDRF SaMD Working Group. *Software as a Medical Device (SaMD): Application of Quality Management System* (October 2, 2015).

IMDRF SaMD Working Group. *Software as a Medical Device (SaMD): Clinical Evaluation* (June 22, 2017).

IMDRF SaMD Working Group. *Software as a Medical Device (SaMD): Key Definitions* (December 9, 2013).

In re Winship, 397 U.S. 358 (1970).

Spencer v. Kemna, 523 U.S. 1 (1998).

The American Law Institute. *Model Penal Code*, § 2.03(1) (May 24, 1962).

The European Parliament, and the Council of the European Union. *Regulation (EU) 2017/745 on medical devices* (April 5, 2017).

U.S. Administrative Procedure Act (1946).

United States v. Gaudin, 515 U.S. 506.

Victor v. Nebraska, 511 U.S. 1 (1994).

6. ARTIFICIAL INTELLIGENCE AFFORDANCES: DEEP-FAKES AS EXEMPLARS OF AI CHALLENGES TO CRIMINAL JUSTICE SYSTEMS

Hin-Yan Liu* and Andrew Mazibrada**

Abstract

This paper attempts to provide a unifying conceptual framework to interpret the perils and promises of artificial intelligence (AI) applications in criminal justice systems. As platform technologies that support a myriad of potential applications, the impact of AI systems upon criminal justice will be amplificatory, divergent, and simultaneous. A unifying framework will facilitate a holistic appreciation of why AI systems might destabilise criminal justice systems and suggest appropriate responses depending on the type of criminal legal disruption at its root. We elaborate this theoretical framework by analysing problems posed by generative "deep-fake" technology, in the context of the criminal justice system in England and Wales. The value of this framework is that it forces us to ask questions about pre-existing normative and procedural responses in a way that reveals future problems, instead of bounded discussions of the application to and adaptability of existing systems, which does not.

Keywords: Deep-fake, invented affordance, unperceived affordance, unexploited affordance, artificial intelligence and the criminal justice system.

Introduction

The core question we consider here is not what AI will do to us, or what we can do to each other with AI,[1] but rather the effects that AI will exert in terms of epistemic doubt and uncertainty within criminal justice systems. This has broad ramifications for: the reception of evidence in criminal trials; enabling means of committing new variations of categories of crime such as fraud, blackmail, or electoral interference; revealing fundamental shortcomings in legal processes and principles; and shifting the social-political landscape in ways that privilege the organised, the technological proficient, or the wealthy who engage in criminal activity. In short, these effects demonstrate that AI interferes with processes taking place within both infrastructures *and* actors, who have become habituated to their existence and the way in which they operate, such that technology now occupies an invisible layer that is barely acknowledged until it fails to function effectively.[2] This leaves us vulnerable to manipulation through that technology and, for the specific purposes of our analysis, manifests in, among other things, an assumed veracity of audiovisual recordings.[3]

Affordances: A Conceptual Framework

The concept of 'affordance' was introduced by James Gibson to make simultaneous reference to *all* action possibilities available to an agent in its environment: 'the *affordances* of the environment are what it *offers* the animal, what it *provides* or *furnishes*, either for good or ill'.[4] The concept of affordances was subsequently adapted and popularised by Don Norman to refer to *perceivable* actions; those that agents *consider* as possibilities.[5] Building from these roots, the sociocultural model for affordance theory[6] represents a further development by structuring the interactions between: affordances 'what could be done'; intentionality 'what would be done'; and normativity 'what should be done'. As such, it situates the concept of affordances within broader contexts that influence or structure the possibilities and constraints placed upon an agent by the environment (in the expanded and contextual senses that also take into account normative and social dimensions of the environment, rather than just in the physical geographical senses).[7]

We adopt and adapt this framework to define specific types of affordances at the intersection of these categories, which we suggest will aid research and policy-making in respect of how AI systems will impact, distort, or disrupt criminal justice systems. These intersections reveal '*invented affordances*' where a lack of (technological) capability is no longer a limiting factor for behaviour; '*unperceived affordances*' where the absence of recognising possibilities for action undergirds the restriction in behaviour; and '*unexploited affordances*' where the lack of action or deployment curtails possible behaviours.[8] Our proposed model thus offers additional nuance in identifying and examining the quality and nature of the type of affordance at play. This in turn enables us to refine our analysis of the shifts and perturbations of the sociotechnical landscape[9] at the root of the criminal legal disruption posed by AI to the criminal justice system.

---

1 * Hin-Yan Liu is Associate Professor and Coordinator of the Artificial Intelligence and Legal Disruption Research Group, Faculty of Law, University of Copenhagen. Email: hin-yan.liu@jur.ku.dk
** Andrew Mazibrada is a PhD Fellow at the Centre for Interdisciplinary Studies of Law and the Artificial Intelligence and Legal Disruption Research Group, Faculty of Law, University of Copenhagen. Formerly of the Criminal Bar of England and Wales and the Crown Prosecution Service, London. Email: andrew.mazibrada@jur.ku.dk
Jack M. Balkin, "The Path of Robotics Law," *California Law Review Circuit* 6, (June 2015): 45–60.

2 Daniel Susser, "Invisible Influence: Artificial Intelligence and the Ethics of Choice Architectures," in *AAAI/ACM Conference on AI, Ethics, and Society (AIES '19)*, (New York, NY: ACM, 2019). https://doi.org/10.1145/3306618.3314286. See also the emerging field of postphenomenology, and in particular Daniel Susser, "Transparent Media and the Development of Digital Habits," in *Postphenomenology and Media: Essays on Human-Technology-World* Relations, ed. Yoni Van Den Eede et al. (Lanham: Lexington Books, 2017), 27-44 and Diane Michelfelder, "Postphenomenology with an Eye to the Future," in *Postphenomenological Investigations: Essays on Human-Technology Relations,* eds. Robert Rosenberger and Peter-Paul Verbeek (Lanham: Lexington Books, 2015), 237–46.

3 In this paper, we use 'veracity' to mean the truth of a recording's contents, namely did what it represents actually take place in the manner represented (whether auditorily or visually).

4 James J Gibson, *The Ecological Approach To Visual Perception* (Hillsdale, NJ: Erlbaum, 1986) 127.

5 Donald A Norman, *The Design of Everyday Things* (Cambridge, MA: MIT Press, 2016).

6 Vlad P. Glăveanu, "What Can Be Done with an Egg? Creativity, Material Objects, and the Theory of Affordances", *The Journal of Creative Behavior* 46, no.3, (2012): 192–208.

7 In James Gibson's original formulation.

8 Hin-Yan Liu, Matthijs Maas, John Danaher, Luisa Scarcella, Michaela Lexer, and Leonard Van Rompaey, "Artificial Intelligence and Legal Disruption: A New Model for Analysis," *Law, Innovation and Technology* 12, vol. 2 (2020).

9 Lyria Bennett Moses, "Regulating in the Face of Sociotechnical Change", in *The Oxford Handbook of Law, Regulation and Technology* eds. Roger Brownsword, Eloise Scotford and Karen Yeung (Oxford: Oxford University Press, 2016).

The bulk of contemporary research situate AI systems as either a regulatory target (AI systems need to be subjected to regulation), or in terms of regulatory opportunities (AI systems can displace or distort the contemporary configuration of regulatory modalities). This places much of the current literature in the 'invented affordances' category, where there is arguably nothing new happening. For example, AI has been (controversially) deployed in making risk-assessments for judicial sentencing decisions, and to pivot towards predictive policing strategies.[10] In such settings, however, AI systems are merely tools deployed into processes that already exist, resulting in impacts that can be deemed to be quantitative (exacerbating current challenges) rather than necessarily being qualitative (introducing new forms of challenges).

The framework we set out in this paper suggests and explores two further arenas in which the criminal justice system can be destabilised. We consider situations where AI systems make new types of behaviour possible, but these possibilities have not yet been recognised ('unperceived' affordances); and where AI systems create new types of behaviour that are recognised possibilities, but which have not yet been acted upon ('unexploited' affordances). Because these capabilities are latent, and because recognition and realisation hinder consequential forms of behaviour, we argue that the precise nature of these second order effects are harder to discern and predict and are considerably more disruptive as a result. Recognition and realisation occur non-incrementally and abruptly because they hinge upon insight and creative modes of thinking, and therefore progress is more difficult to monitor, analyse, and evaluate. These characteristics have far-reaching consequences for criminal justice systems because the disruptive impacts do not primarily flow from what the new technology is, nor what it can do: rather such impacts are entirely contingent upon creative recognition of the uses that such technology could be put to, and how that technology is actually deployed. To demonstrate the affordances framework for the purposes of this paper, we will use AI applied as Generative Adversarial Networks (GANs) to facilitate 'deep-fakes'.[11] Furthermore, this contingency extends to how the affordances opened up by a new technology *combine* with affordances already present in the world, as well as the potential affordances presented by other new and emerging technologies. To demonstrate this, we use the ubiquitous adoption of smartphones replete with audiovisual recording, storage, and dissemination capacities.

Taken in isolation, and for reasons we explore below, we suggest the technological capacity to manipulate or fabricate[12] audio and visual material in a convincing manner is something criminal trial processes will, in relatively short order, be able to accommodate. It is when deep-fakes are *combined* with a ubiquitous capacity to record, store, and disseminate audio and visual material that more profound and fundamental challenges are raised for criminal justice systems as a whole.

10 On which see generally, for example: Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, (London: Penguin Allen Lane, 2016); Timo Rademacher, "Artificial Intelligence and Law Enforcement", in Thomas Wischmeyer and Timo Rademacher, *Regulating Artificial Intelligence,* (Cham, Switzerland: Springer, 2020), 226-255; Anupam Chander, "The Racist Algorithm?," *Mich. L. Rev.* 115, (2017): 1023; Ryan Calo, "Artificial Intelligence Policy: A Primer and Roadmap," 51 *U.C.D. L. Rev.* 51 (2017): 399; Ashley Deeks, "The Judicial Demand for Explainable Artificial Intelligence," *Columbia Law Review* 119, no. 7 (2019): 1829-50; and Christopher Markou, "Why using AI to sentence criminals is a dangerous idea," *The Conversation*, May 16, 2017, https://theconversation.com/why-using-ai-to-sentence-criminals-is-a-dangerous-idea-77734.

11 It is outside the scope of this paper to discuss the way in which deep-fakes are created, their historical development, or the myriad threats they pose outside the criminal justice context. For a general discussion of deep-fakes and an introduction to the many issues surrounding them, see, for example, Robert Chesney and Danielle Citron, "Deep-fakes: A Looming Challenge for Privacy, Democracy, and National Security." *California Law Review* 107, no. 6 (December 1, 2019): 1753; Robert Chesney and Danielle Keats Citron, "21st Century-Style Truth Decay: Deep-fakes and the Challenge for Privacy, Free Expression, and National Security," *Maryland Law Review* 78, no. 4 (2019): 882-891; Mary Anne Franks and Ari Ezra Waldman, "Sex, Lies, and Videotape: Deep-fakes and Free Speech Delusions," *Maryland Law Review* 78, no. 4 (2019): 892-898; and Holly Kathleen Hall, "Deep-fake Videos: When Seeing Isn't Believing," *Catholic University Journal of Law and Technology* 27, no. 1 (Fall 2018): 51-76. For a recent analysis of the current state of commodification of deep-fakes, see Robert Volkert and Henry Adjer, "Analyzing the Commoditization of Deep-fakes." *NYU Journal of Legislation and Public Policy*, February 27, 2020, https://nyujlpp.org/quorum/volkert-ajder-analyzing-the-commoditization-of-deep-fakes/ (accessed February 28, 2020) and Henry Ajder, Giorgio Patrini, Francesco Cavalli, and Laurence Cullen, "The State of Deep-fakes: Landscape, Threats, and Impact." Deeptrace Labs, September 2019.

12 In this paper, we use 'manipulation' to refer to the deliberate alteration of a pre-existing recording and 'fabrication' to refer to the creation of a recording from scratch using unrelated pre-existing material that has little or no previous relevance to the facts in issue. Manipulation includes, for example, what are known as 'shallow fakes' or 'cheap-fakes' such as the now infamous video of US Congresswoman Nancy Pelosi, which was slowed down to make the House Speaker appear intoxicated, on which see, for example, Robert Chesney, Danielle Citron, and Quinta Jurecic, "That Pelosi Video: What To Do About 'Cheapfakes' in 2020", *Lawfare* (May 29, 2019), https://www.lawfareblog.com/about-pelosi-video-what-do-about-cheapfakes-2020.

Generative Adversarial Networks (GANs) and Criminal Evidence

We begin by examining the potential impact of deep-fake manipulation and fabrication of audiovisual material offered as evidence in the criminal justice system of England and Wales (hereafter: the English system). We do this because the most obvious effect of the deep-fake phenomenon in this context is on the receipt of evidence into the criminal process.[13] Questions of admissibility in relation to audiovisual evidence exemplify the direct problems posed by deep-fake manipulation and fabrication, and the way in which courts might initially struggle to adequately categorise and respond to the phenomenon. Although problematic, however, deep-fake evidence is not in and of itself the principal problem. We suggest the situation is more complex and it is our proposed affordances framework which reveals that complexity.

We need to make two observations at the outset that frame our analysis. First, locating manipulation or fabrication achieved by GANs with any degree of genuine certainty, let alone proving it to the criminal standard and attributing it responsibility, is not currently possible[14] through technological means. Such proof might of course be possible through other corroborative means, for example through contextual evidence or witness testimony, and this we address in detail below. Second, audio and video manipulation and fabrication are nothing new as such – filmmaking has engaged in iterations of it for as long as there has been cinema, leading to complex CGI and effects-laden films like James Cameron's Avatar.

The paradigm shift we suggest that is now taking place involves diffuse societal and technological contexts which are beginning not only to lower the threshold for engaging in such manipulation and fabrication, but which now make audiovisual material both ubiquitous and more easily deployed.[15] The need for specialist expertise and equipment, a great deal of time, considerable expense, or any combination of these, is quickly declining in relevance. These are 'invented affordances' where the lack of technological capability that was once the limiting factor for behaviour no longer exists. If deep-fakes existed in a vacuum, however, even the lowering of such thresholds to entry would present little cause for alarm. Courts could adapt to the small number of cases where manipulation or fabrication by means of GAN technology was alleged to have taken place. Only when placed in the ubiquity context is the gravity of the effect revealed.

The evidential question affords two key potentialities for a criminal tribunal. The first lies in the ability of any party, particularly the accused, to call into question, regardless of the truth of the matter, the veracity of evidence that previously would have been accepted[16] as factual on its face. The second is the ability of any party to the proceedings to manipulate or fabricate evidence, and potentially with impunity. Both are truly enabled by the ubiquity context. Even now, were the accused to object to the veracity of audiovisual evidence without supporting evidence, it would usually be summarily rejected. In the coming years, the situation will very likely be different as such an allegation is becoming more tenable. Again, these are 'invented' affordances: both of these adversarial forensic tactics existed before deep-fakes – defendants in criminal trials have long alleged evidence planting or that a witness was lying on oath. To some extent, the same can be said of fabricated evidence, such as tendering a forged document into evidence. Criminal justice systems are structured around addressing this behaviour through the weighing and admissibility of evidence, or inter alia through offences

13 Although this may seem obvious, there has been surprisingly little discussion in this area, particularly doctrinally by legal scholars. See, however, Centre for Data Ethics and Innovation (CDEI). "Deepfakes and Audio-visual Disinformation." (September 2019), 10-14 (The Centre for Data Ethics and Innovation is an Expert Committee of the United Kingdom Department for Digital, Culture, Media and Sport). See also Jeff Ward, "10 Things Judges Should Know About AI." *Judicature* 130, no. 1 (2019) 12-18; Marie-Helen Maras and Alex Alexandrou. "Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos." *The International Journal of Evidence & Proof* 23, no.3 (2019): 255–262; Rebecca J. Hamilton. "New Technologies in International Criminal Investigations." *Proceedings of the ASIL Annual Meeting* 112 (2018): 131–33; and Rebecca J. Hamilton. "User-Generated Evidence." *Columbia Journal of Transnational Law* 57, no.1 (2018) American University, WCL Research Paper No. 2018-11.

14 Ibid, fn 11 and post, fn 18. Although work is being done right now to develop and test AI systems capable of detecting deep-fakes (see for example Jay Peters, "Alphabet's Jigsaw unveils a tool to help journalists spot deep-fakes and manipulated images," *The Verge,* February 4th, 2020, https://www.theverge.com/2020/2/4/21122778/alphabet-jigsaw-assembler-tool-news-journalists-deep-fakes-images) there is considerable concern that machine learning (ML) will always outstrip whatever detection algorithms are built (see: James Vincent, "Deep-fakes detection algorithms will never be enough," *The Verge*, June 27, 2019, https://www.theverge.com/2019/6/27/18715235/deep-fakes-detection-ai-algorithms-accuracy-will-they-ever-work). On the 'offence/defence' balance asymmetry in AI, see Ben Garfinkel and Allan Dafoe, "How does the offense-defense balance scale?", *Journal of Strategic Studies*, 42, vol.26, (2019): 736-763, DOI: 10.1080/01402390.2019.1631810.

15 In our use of 'deploy', we refer to all means of dealing with a recording once it has been made: saving it, storing it, using it, showing it, and sharing it. All have become cheaper and easier, and sharing can potentially reach wider target audiences previously either unreachable or more difficult to reach.

16 Usually by way of rebuttable presumption, on which much in this paper turns.

like perjury.[17] Thus, there is a two-fold paradigm shift at play here. First, manipulated or fabricated material will become more readily available and easily deployed, and may soon reach a point where detection of the fake might be permanently impossible[18] – visual clues may effectively disappear and technical clues may be forced to resort to probabilistic statements that do not adequately fit the contemporary evidential paradigm. Second, what might have been a discrete, containable issue is intensified or otherwise morphed into different types of challenges by the ubiquity context.

In the English system, the principle rule of evidence is that, subject to any applicable exclusionary rules or judicial discretion, all evidence which is sufficiently relevant to the facts in issue is admissible, and all evidence which is irrelevant or insufficiently relevant should be excluded. Evidence which is relevant may still be excluded if no reasonable jury, properly directed as to its defects, could place any weight on it.[19] When dealing with audio and video evidence, admissibility and relevance rests upon an implicit presumption of veracity. This stance interrogates the relationship between copies and the original recording (problematically termed 'authenticity' by the case law), while overlooking or otherwise under-appreciating the possibility that the original may itself be a fabrication. This reveals the impact of unperceived and unexploited affordances for criminal justice – it fails to accommodate the increased *combined* possibilities and usages of smartphones and GANs to lower the barriers to entry for manipulating and fabricating audiovisual evidence.

This shortcoming is evidenced by the long-established presumption that mechanical and other instruments of a similar kind, that are usually in working order, were in working order at the time of their use.[20] The party against whom the presumption operates bears an evidential burden to adduce some evidence to the contrary. If that relatively low bar is crossed, and the veracity of a recording is raised as an issue, fundamental questions as to admissibility and weight are raised and there will be no easy answers. Cross-examination has long been one of the key traditional means of testing and assessing contested evidence, but if the originator of a recording is unavailable or not known, a plausible situation in today's global social media environment, this might be impossible. Circumstantial evidence also impacts weight and admissibility and courts might be increasingly forced to rely more on this type of evidence.[21] Problems arise where there is little or no other such evidence and where manipulation or fabrication has not been catered for by ill-considered legislation. By way of example, this paper later considers offences with which the Violence Against Women and Girls

17 It is important to note, in a paper this short, that some arguments will be beyond its scope. This is not to dismiss them as unimportant, however. By way of analogy, there is considerable debate about the possibility of criminalising the making of deep-fakes. This presents considerable, potentially insurmountable, problems. Firstly, the software applications which permit deep-fakes are often coded by numerous parties, frequently on fora where anonymity is prized and protected, almost always from multiple jurisdictions. There are also human rights arguments offered in respect of freedom of expression. The point is made here to demonstrate that there are no clear solutions, as has often been the case in the past, when it comes to the criminalisation of new conduct. Such arguments will be the subject of further research by the authors of this paper.

18 Ibid, fn 11. The important work of Dartmouth College image expert Hany Farid in the area of image forensics subsumes the detection of deepfakes within that field: Hany Farid. "Image Forensics." *Annual Review of Vision Science* 5, no.1 (2019), 549-573 and Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. "Protecting World Leaders Against Deep Fakes." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (2019): 38–45. In the New Zealand context, and using a specifically doctrinal legal methodology, see Curtis Barnes and Tom Barraclough. "Perception Inception: Preparing for deepfakes and the synthetic media of tomorrow." New Zealand Law Foundation, 21 May 2019, https://www.brainbox.institute/report.

19 For an excellent summary, see for example David Ormerod QC and David Perry QC, *Blackstone's Criminal Practice 2019*, (Oxford: OUP 2019), F1.11, where one example cited is *Robinson* [2006] 1 Cr App R 13 (at 221), a case concerning voice recognition evidence. Evidence is relevant 'if it is logically probative or disapprobative of some matter which requires proof.'*DPP v Kilbourne* [1973] AC 729, at p. 756 *per* Lord Simon of Glaisdale.

20 In *Tingle Jacobs & Co. v Kennedy* [1964] 1 All ER 888, Lord Denning MR said (at p. 639) 'when you have a device of this kind set up for public use in active operation … the presumption should be that it is in proper working order unless there is evidence to the contrary'. Additionally, see generally Perry and Ormerod, *Blackstone's Criminal Practice 2019*, F3.59.

21 This is not necessarily problematic in itself, and it does not necessarily follow that the weight attached to circumstantial evidence will be less than accorded to direct evidence. However, circumstantial evidence "works by cumulatively, in geometrical progression, eliminating other possibilities" (*DPP v Kilbourne* [1973] AC 729 *per* Lord Simon, 758). Some see it as cords of a rope, weaker on its own than with other evidence to bind to. Even though circumstantial evidence may sometimes be probative, however, it must always be narrowly scrutinised as fabrication is always a live issue: "It is also necessary before drawing the inference of the accused's guilt from circumstantial evidence to be sure that there are no other co-existing circumstances which would weaken or destroy the inference." (*Teper v The Queen* [1952] AC 480, *per* Lord Normand, 489).

(VAWG) Strategy[22] is concerned, focusing on challenges in evidencing sexual assaults and potential lacunae in offences related to the sharing of fabricated sexual images.

Questions of admissibility also reveal a wider systemic tension in the case of recordings alleged to have been manipulated or fabricated: where such evidence might conceivably have been tainted or its veracity becomes an issue which cannot be resolved (recall that currently the means to detect deep-fakes does not exist), there may be no option but to exclude it on the grounds of fairness to the accused.[23] To use the analogy of expert[24] forensic evidence gathered at a crime scene which is later subject to doubt or tainted in some way,[25] it will often be excluded or its weight diminished. Should the accused discharge an evidential burden, raising some level of doubt in the court's mind that audiovisual material has been manipulated or fabricated, a key source of evidence, perhaps the only corroborative evidence in some instances like those covered in VAWG cases, could be viewed as sufficiently compromised to be unreliable and therefore excluded, which could set problematic future case precedents.

Thus, courts may turn to examining the technical systems that are the initial provenance of the material itself, if they can, and question whether that system could realistically, in some way, have been tampered with. In a CCTV system, the potential for manipulation using generative technology may be less likely than in a more exposed system. To take the forensic evidence analogy further, courts rarely ask whether the forensic evidence could itself have been tampered with. If a similar approach is taken with deep-fakes, such an approach will ignore the technologically lowered barriers to entry that increase the disruptive potential of unrecognized and unperceived affordances: the ability of ordinary people to produce forged footage will soon be far greater than their ability to produce altered forensic samples because of the technological availabilities we have mentioned. Where a smartphone recording could be used to indict or exculpate, it might be unsafe to rely on that evidence if a realistic prospect exists that it might have been forged.[26]

Deep-fake manipulation and fabrication is possible in the context of both audio and video recordings which are dealt with differently by the English system, a curious disjunction which itself reveals some important problems.[27] The contents of audio recordings, even if copies of an original,[28] if produced and played in court,

22 "Violence against Women and Girls (VAWG) Strategy 2017–2020," (London: CPS, 2017), https://www.cps.gov.uk/sites/default/files/documents/publications/VAWG-Strategy-2017-2020.pdf. See also https://www.unwomen.org/en/what-we-do/ending-violence-against-women for the wider strategy of the United Nations itself. Chesney and Citron highlight an analogous issue which they call the 'Liar's Dividend': "Deep fakes will make it easier for liars to deny the truth in distinct ways. A person accused of having said or done something might create doubt about the accusation by using altered video or audio evidence that appears to contradict the claim. … Deep fakes will prove useful in escaping the truth in another equally pernicious way. Ironically, liars aiming to dodge responsibility for their real words and actions will become more credible as the public becomes more educated about the threats posed by deep fakes." Robert Chesney and Danielle Citron, "Deep-fakes: A Looming Challenge for Privacy, Democracy, and National Security." ibid, fn 7, 1785.

23 In the English system, there has long been, either at common law, or as a result of legislation, a discretion to exclude evidence which might affect the fairness of the proceedings. However, as the main statutory vehicle for this discretion, section 78(1) of the Police and Criminal Evidence Act 1984, demonstrates, this is generally focused on evidence adduced by the prosecution: 'In any proceedings the court may refuse to allow evidence on which the prosecution proposes to rely to be given if it appears to the court that, having regard to all the circumstances, including the circumstances in which the evidence was obtained, the admission of the evidence would have such an adverse effect on the fairness of the proceedings that the court ought not to admit it.' At common law a judge has a discretion in a criminal trial 'to exclude evidence if it is necessary in order to secure a fair trial for the accused' per Lord Griffiths in *Scott v R.* [1989] A.C. 1242 at 1256, PC, referring to numerous dicta to that effect in *Selvey* [1970] A.C. 304, HL, and *Sang* [1980] A.C. 402, HL. In both instances, evidence adduced by a defendant does not face the same exclusionary regime.

24 In matters of science or trade, for example, the opinion of an expert, or person intimately acquainted with that science or trade, is admissible to furnish the court with information which is likely to be outside the experience and knowledge of a judge or jury. Only if, on the proven facts, a judge or jury can form their own conclusions without help, shall the opinion of an expert be unnecessary: *Turner (T.)* [1975] Q.B. 834; (1974) 60 Cr. App. R. 80, CA. See generally Mark Lucraft QC, *Archbold: Criminal Pleading, Evidence & Practice 2020*, (London: Sweet and Maxwell, 2020), 10-47.

25 Usually as a consequence of its mishandling or documenting errors which leave open grounds to allege it has been tampered with or might not be the sample recovered. When considering the effect of evidence which falls outside of the court's experience or knowledge, such as technical or expert evidence the provenance of that evidence, and the chain of custody during its analysis, has traditionally required complete transparency.

26 A proven forgery might still be relevant through the fact of its forgery, for example, rather than the truth of its contents in the same way as lies by a defendant or perjury by another witness, subject to appropriate warnings given by the trial judge.

27 The precise reasons for the separate evolution of audio and visual evidence admissibility questions are beyond the scope of this short paper, but certain hypotheses could be posited: for example, the fact that audio recording as a technology preceded video recording might be relevant, and audio may well, at the time of its absorption into the legal landscape, have seemed enough like statements contained with documents to have come to be considered as such. The visual nature of photography and video recordings are sufficiently different to have engendered an entirely separate legal analysis. Further scholarship on this question could be valuable to constructing a new framework to regulate the admissibility of audio and video evidence in criminal trials in England and Wales.

28 It is immaterial how many removes there are between a copy and the original and, as to 'authentication', it is currently thought that the courts are likely to require the same kind of proof that was necessary in the case of copies at common law, namely a proper explanation as to why the originals are not available, and proof of the complete accuracy of the copies in relation to the original: Ormerod and Perry, *Blackstone's Criminal Practice 2019*, F8.53, referring to *Robson* (June 1973, unreported).

may be admitted as evidence of their truth under an exception to the hearsay rule[29] because an audio recording is considered a 'document' for these purposes.[30] Whether the audio has been manipulated or fabricated is an entirely separate issue. Without more, there is a rebuttable[31] presumption of veracity on its face.[32] Conversely, it is clear[33] that video evidence, from a CCTV system for example, is not hearsay. Photographs and films are admissible[34] at common law as a variety of real evidence.[35] Little weight attaches to real evidence in the absence of accompanying evidence identifying it and connecting it with the facts in issue. Anyone seeking adduce a recording in evidence, whether audio or video, must satisfy a court that a prima facie case of *originality* and *authenticity* has been made out by the calling of other evidence which defines and describes the provenance and history of the recording up to the moment of its production in court. Yet, authenticity relates purely to the relationship between copies and the original recording and emphasis is rarely placed on the possibility that the original may itself be a fabrication.[36] The problem for courts now, as we have said, is *veracity*: that the original may itself be a fabrication and no longer represents a reliable and accurate depiction of what it purports to represent. As we will show, this presents a two-sided dilemma.

The positioning of the contents of an audio or visual recording within different admissibility regimes is a function of something well recognised: as technology develops, law develops after it, usually not quickly enough to deal with it, nor comprehensively enough to capture anything more than the essence of the problem posed. This divergent evolution in respect of audio and video recordings will soon become starkly untenable, particularly where such types of evidence are more often combined as they will inevitably be in the age of the ubiquitous smartphone.

Combinatorial Affordances: Deep-fakes in the Ubiquity Context

Attempts to analyse the impact of AI purely in the sense of invented affordances overlook the diverse challenges that result from how actors recognise and exploit the potentialities that the technologies afford. Invented affordances are the easiest challenges to confront only when their combinations with other technologies are ignored. Further, unrecognised and unexploited affordances undergo a combinatorial explosion where these intersect with complementary affordances supplied by other technologies. The true impact of deep-fakes,

29 See, for example, *Senat* (1968) 52 Cr App R 282, which concerned tape recordings of incriminating conversations. The law of England and Wales has 'always insisted that it is ordinarily essential that evidence of the truth of a matter be given in person by a witness who speaks from his own observation or knowledge. It uses the legal expression "hearsay" to describe evidence which is not so given, but rather is given second hand, whether related by a person to whom the absent witness has spoken, contained in a written statement of the absent witness, given in the form of a document or record created by him, or otherwise.' (*Horncastle* [2010] 2 AC 373, *per* Lord Thomas CJ). Such second-hand accounts are unperceived as inferior to primary evidence which may be tested by the trial process, whether by cross-examination or by other forensic means.

30 Criminal Justice Act 2003, s. 133 and s. 134(1).

31 In the case of computer printouts, which may provide support for analogous arguments, before the judge can decide whether they are admissible as real evidence or as hearsay pursuant to statute, it is necessary for appropriate authoritative evidence to be called to describe the function and operation of the computer (*Cochrane* [1993] Crim LR 48).

32 See for example the public online statement of the online CPS Legal Guidance in respect of Exhibits where, in terms of proving the 'authenticity' of the video recording, the Prosecution must be able to show that 'the video film produced in evidence is the original video recording or an authentic copy of the original' and show that it has 'not been tampered with'. In order to do so statements must be available which produce the video evidence as an exhibit and which cover its continuity and security. No mention of its veracity, or anything similar, is made. "CPS Legal Guidance: Exhibits," revised: 9 April 2018 (https://www.cps.gov.uk/legal-guidance/exhibits).

33 From the definition of 'statement' in section 115(1) of the 2003 Act as a representation of fact or opinion made *by a person* (as is the case with an audio voice recording).

34 A photograph or film, the relevance of which can be established by the evidence of someone with personal knowledge of the circumstances in which it was taken or made, may be admissible. Juries are permitted to see still photographs taken by a security camera during an armed robbery (*Dodson* [1984] 1 WLR 971), or a video recording of an incident (*Fowden* [1982] Crim LR 588; *Grimer* [1982] Crim LR 674).

35 Real evidence is usually some material object, the existence, condition, or value of which is in issue or relevant to an issue, produced in court for inspection by the tribunal of fact. In *The Statue of Liberty* [1968] 2 All ER 195, where a film, in effect, contained a statement as to the paths taken by the two ships, Sir Jocelyn Simon P, rejecting a submission that a cinematograph film of radar echoes, recorded by a shore radar station, was inadmissible because it was produced mechanically without human intervention, said (at p. 740): 'If tape recordings are admissible, it seems that a photograph of radar reception is equally admissible – or indeed, any other type of photograph. It would be an absurd distinction that a photograph should be admissible if the camera were operated manually by a photographer, but not if it were operated by a trip or clock mechanism.'

36 Of some concern in the context of our arguments is the line of precedent permitting video evidence to be proved by the parol evidence of witnesses who have seen the photograph or film. In *Taylor v Chief Constable of Cheshire* [1986] 1 All ER 225, a video cassette recording, made by a security camera and showing a person in a shop picking up an item and putting it into his jacket, was played to police officers who identified the person as Taylor. The recording, after it had been returned to the shop, was accidentally erased. Evidence by the officers of what they had seen on the video was held to have been properly admitted, on the ground that what they had seen on the video was no different in principle from the evidence of a bystander who had actually witnessed the incident, despite the fact a bystander could be cross-examined as to the truth of their evidence.

we suggest, is to be found in their combination with near ubiquitous digital audiovisual recording capability through widespread smartphone adoption, the ease of sharing such audiovisual recordings provided by constant interconnectivity and cloud-based systems, and the normative social and economic context where such sharing is encouraged and expected. That proliferation of digital audiovisual material combines with the lowered barriers to deep-fake fabrication and manipulation to create a more general and pervasive epistemic uncertainty concerning the veracity of *any* audiovisual material.

We suggest, therefore, that while the invented affordances provided by deep-fakes may, by themselves, have created initial challenges for the weighing of criminal evidence, their impact would have remained relatively bounded without complementary invented affordances represented by smartphone and cloud computing capabilities. Notwithstanding this, of far greater concern are the recognition and exploitation affordances: in our example, and there are likely many others, this stems from the realisation that deep-fake manipulation and fabrication can be connected to the proliferation of smartphone video and audio, and the exploitation of this realisation.[37] Put another way, the epistemic challenges that arise from the possibility that all video and audio can be manipulated or fabricated would simply not exist without a broader context where digital audio and video are ubiquitous, and also vulnerable to deep-fakes.

Contrast diffuse smartphone use with closed systems like CCTV, where the initial provenance of the footage concerned can usually be discerned and the likelihood of interference can probably be dismissed. In the age of surveillance capitalism,[38] however, where social media encourages the recording and sharing of daily life and experience, as the internet of things becomes a pervasive reality, and where smartphones in the hands of billions of citizens possess not only high quality video capabilities but the storage capacity to retain such videos, there exists the very real prospect of video and audio evidence being presented that does not come from closed systems. Such material might well appear to be relevant to facts in issue – indeed, in case of serious sexual assault or rape, which are again currently receiving media attention for accusations of a risk-averse approach to prosecutions by the Crown Prosecution Service,[39] and consequentially low rates of conviction,[40] evidence from smartphones may provide compelling circumstantial evidence of conduct leading up to the alleged assault, whether inculpatory or exculpatory. Our analysis demonstrates an unwieldy, piecemeal approach to audio and visual evidence in the English system, attempting to fit it into or exclude it from pre-existing doctrines such as hearsay or real evidence, but audio and video evidence in the ubiquity context, we argue, no longer suits any existing category and requires a completely new approach: a category of its own with its own evidential regime.

The VAWG offences arena also reveals other continuums between affordances. We argue that deep-fakes have the potential to both create new offences that do not currently exist, and new means of committing existing offences. We exemplify the tension between unperceived and unexploited affordances, and the slowness of the law to therefore anticipate them, through a brief analysis of section 33 of the Criminal Justice and Courts Act 2015, which recently made it an offence to disclose private sexual photographs or films without the consent of the individual who appears in them and with intent to cause that individual distress.[41] Manipulated

37 For a thorough analysis, see Robert Chesney and Danielle Citron, "Deep-fakes: A Looming Challenge for Privacy, Democracy, and National Security." ibid, fn 7. Also, Katarina Kertysova. "Artificial Intelligence and Disinformation: How AI Changes the Way Disinformation is Produced, Disseminated, and Can Be Countered." *Security and Human Rights* 29 (2018) 55-81, 67 (discussing the 'growing ease of making and sharing fake video and audio content across computers and mobile devices' creating ample opportunities for 'intimidation, blackmail, and sabotage beyond the realm of politics and international affairs.'). Europol also recently noted the potential for criminal use of deep-fake technology: Europol, "Do Criminals Dream of Electric Sheep?" European Union Agency for Law Enforcement Cooperation 2019, 10.

38 The literature on this area is voluminous and approaches the complex of issues from a wide variety of perspectives, but for comprehensive analyses see Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power.* New York: Public Affairs, 2019 and Julie E. Cohen, *Between Truth and Power: The Legal Constructions of Informational Capitalism*, Oxford: OUP, 2019.

39 Caelainn Barr, Alexandra Topping, and Owen Bowcott, "Rape prosecutions in England and Wales at lowest level in a decade," *The Guardian,* September 12, 2019, https://www.theguardian.com/law/2019/sep/12/prosecutions-in-england-and-wales-at-lowest-level-in-a-decade.

40 Caelainn Barr and Owen Bowcott, "Fall in rape convictions 'due to justice system at breaking point'," *The Guardian*, December 19, 2019, https://www.theguardian.com/law/2019/dec/17/fall-in-convictions-due-to-justice-system-at-breaking-point.

41 One related major lacuna in this offence, which demonstrates the wider lack of understanding about the relationship between social media and reputation, relates to its *mens rea*: a person will only be guilty of the offence if the reason for disclosing the photograph, or one of the reasons, is to cause distress to a person depicted in the photograph or film. Anyone sending the message purely because they thought it might be amusing would not be committing the offence. This seems a startling lacuna given the emotional distress caused, and the potentially permanent reputational damage, is likely to be the same regardless of the sender's intent.

or fabricated pornographic images would appear not to be covered by the offence[42] which is drafted so that it 'only applies to material which … originates from an original photograph or film recording.'[43] This, the Crown Prosecution Service asserts in its Legal Guidance for prosecutors, is because 'the harm intended to be tackled by the offence is the misuse of intimate photographs or films.' Is the harm of a fake video, but which looks real, any less than that of a genuine image or video circulated? We would suggest the answer must be no.[44] Thus, deep-fake technology creates new means of committing old offences, even relatively recently-drafted offences, which do not fit into existing systems of detection, culpability, or proof. Section 33 represents an example of law, and its institutions, barely able to keep pace with technological change.

The discussion in respect of admissibly of smartphone recordings reveals a further tension. In the English system, in order to achieve equality of arms in the adversarial process, exclusionary rules tend to apply more to evidence adduced by the prosecution than by the defendant. Clearly, the state has traditionally been far more likely to have relevant and available technical evidence, like audio or video recordings, than was the individual accused. In the age of the smartphone, that will no longer be the case. Imagine an allegation of rape during a party. Evidence of relevant earlier conduct is captured by the victim or other partygoers, who are prosecution witnesses, and by the defendant or other partygoers who have evidence relevant to the defendant. Different admissibility questions will apply to identical types of evidence, obtained through identical technical means. To what extent will the defendant be required to prove, if at all, that the evidence adduced in their defence is not a fake, given that it precisely what the Crown may be required to do?

This tension intensifies when considering the resources and affluence of entities such as corporations accused of white-collar fraud, terrorist groups, or organised criminal networks. If the ability of an individual to secure a manipulated or fabricated audio or video recording seems initially fantastical, it is far less so in the case of such wealthy and connected organisations, for whom the same admissibility tension arises. One means by which terrorist associations are proved, for example, is by way of evidencing interviews given by alleged terrorists to news media in regions such as Syria. These interviews are inevitably edited, leading to one category of admissibility issue in itself, but more challenging problems arise if a suspect declares they were interviewed at all and the video is a complete fabrication. Given the power of disinformation and cyberwarfare in such low-intensity conflicts, this claim might have some foundation, yet obtaining the original, or evidence from the originator, might prove to be impossible. Conversely, were audiovisual alibi evidence produced to demonstrate the defendant could not possibly have been interviewed, it might well be subject to a different standard.

Situations can be envisaged where deep-fakes might also lead to the criminalisation of those who hitherto, when offences required more physical, financial, or emotional input and risk, might not have committed those offences or whose conduct might have been considered relatively minor when previously more self-contained. Deep-fakes in the ubiquity context represent a paradigm shift in the ease with which smartphone users might succumb to temptation in respect of fraud, blackmail, or harassment, for example. In situations of high emotion or arousal, where rational thinking is impaired,[45] technology now offers immediate affordances

42 The Crown Prosecution Service Legal Guidance on this offence, drafted before deep-fakes became so prevalent in public consciousness, demonstrates how quickly technological advances can make legislation outdated. Given this guidance in part helps determine which offences are prosecuted, we suggest it is valuable in assessing the effectiveness of the offence. "CPS Legal Guidance: Revenge Pornography – Guidelines on prosecuting the offence of disclosing private sexual photographs and films", revised 24 January 2017 to add a section on sexting https://www.cps.gov.uk/legal-guidance/revenge-pornography-guidelines-prosecuting-offence-disclosing-private-sexual.

43 "CPS Legal Guidance: Revenge Pornography", ibid. The offence does not include situations where the alteration or combination of a film or photograph only then makes it sexual or if the intended victim is only depicted in a sexual way as a result of the alteration or combination. So, for example, a person who simply transposes the head of a former partner onto a sexual photograph of another person, a common means of using deep-fakes technology, will not commit the offence. Further, images which are 'completely computer generated but made to look like a photograph or film will not be covered by the offence.'

44 The CPS seems to endorse this position in further guidance relating to much older legislation dating back to 1988: communications sent via social media may involve the commission of a range of existing offences against the person, public justice, sexual or public order offence as well as the commission of communications offences contrary to section 1 of the Malicious Communications Act 1988 and/or section 127 of the Communications Act 2003: "CPS Legal Guidance: Social Media – Guidelines on prosecuting cases involving communications sent via social media," revised: 21 August 2018. https://www.cps.gov.uk/legal-guidance/social-media-guidelines-prosecuting-cases-involving-communications-sent-social-media.

45 The 'hot state' of arousal, and its relationship to temptation, is a key area in the field of behavioural economics and the study of cognitive biases. While we do not explore the effect of cognitive biases in this paper, they are many, complex, and outside its scope, precisely how they affect decision-making cannot be divorced from the kinds of issues raised by the many ways and contexts in which deep-fakes could operate to manipulate us. On behavioural economics generally, see three seminal works: Dan Ariely, *Predictably Irrational: The Hidden Forces That Shape Our Decisions,* (London: Harper, 2009); Daniel Kahneman, *Thinking Fast and Slow* (London; Penguin Allen Lane, 2012); Richard H. Thaler and Cass R. Sunstein, *Nudge: Improving Decisions about Health, Wealth, and Happiness.* (London: Penguin Allen Lane, 2009).

to those who might otherwise have been forced to take time to reconsider. The emotional and reputational damage of showing (but retaining) sensitive sexual images to close friends at a party, even manipulated or fabricated ones,[46] is exacerbated exponentially when those same images are shared widely on social media and no longer under the control of the initial disseminator.[47] New technology has historically led to new and often unforeseen behaviour, in particular criminality. The advent of the motor car led to a regime of offences related to driving and road traffic regulation. The introduction of the passport and the later computerisation of identity documentation has facilitated a far more pervasive theft of identity and fraudulent impersonation than was possible before.

Such offences might be seen as entirely new, a genuine and then unforeseen consequence of new technology on an unwitting society, or as a species of older crime evolved into something different and unanticipated. For all practical purposes, this distinction may not matter and perhaps even demonstrates the fact that the affordances within this framework are not discrete and fixed, but rather continuums along which affordances move, evolve, and combine over time and with advances in science and technology. The clear point is that the effects of such advances cannot always be predicted: they are to be considered 'unperceived affordances', and a historical analysis in this arena reveals new behaviour that has been regulated by way of criminalisation.

Concluding Thoughts

We have suggested that there is a two-stage challenge posed by GAN-facilitated deep-fakes to criminal justice systems. In the first stage, when treated in isolation, deep-fakes pose a limited problem concerning their evidential veracity that can be likely accommodated by a robust judicial system. In the second stage, however, the combinatorial nature of the affordances presented by deep-fakes combines within the ubiquity context of audiovisual recording and replication to undermine implicit, yet foundational, epistemic presumptions. If taken seriously, such erosions threaten the very possibility of relying on broad categories of evidence in criminal procedure. Such an analysis is made possible not by examining the intrinsic nature and characteristics of artificial intelligence, nor by dissecting its particular applications, but rather by situating the effects within the conceptual framework of affordances. Yet, even such an examination is insufficient to truly grasp future challenges and it will only be through the more detailed analysis using the combinatorial approach that we will even begin to engage with the impending fundamental shifts in the sociotechnical landscape.

This exercise hints at the oblique nature of the challenges posed by AI for criminal justice that cannot be identified in isolation beforehand, and is a powerful assertion of the Collingridge dilemma.[48] Our suggestion for attempts to frame criminal justice responses to AI challenges goes beyond the conceptual framework of affordances proposed here (which is but one possible approach) and suggests we look to how the deployment of the technology combines both with the existing sociotechnical landscape, and with the recognised and potential uses of other new and emerging technologies. It is likely that, although AI will pose deep challenges in isolation, the more fundamental questions will lie at the Lagrangian points between new technologies.

References

Agarwal, Shruti, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. "Protecting World Leaders Against Deep Fakes." In *Proceedings of the*

46 It is possible to argue that where the fact of the manipulation or fabrication cannot be proved and is therefore in doubt the sense of frustration and injustice might even exacerbate this further.

47 In Denmark, in a recent high-profile case, 1,152 young people have been or are being investigated for sharing child pornography on social media. Some have been sentenced to terms of imprisonment. They were charged with sharing sexual footage of two 15-year-olds after two boys shared the film on the internet: this became known locally as the Umbrella case. The footage came from a private address in North Zealand in 2015, where two boys shared humiliating pictures of a 15-year-old boy and girl. The girl reported the conduct, and both were awarded compensation. Even before this case has been completely resolved, another similar case is currently being investigated, again involving widespread sharing. "Fakta: Flest har fået bøder i Umbrella-sagen", *Jyllandsposten*, January 25, 2020, https://jyllands-posten.dk/indland/politiretsvaesen/ECE11899455/fakta-flest-har-faaet-boeder-i-umbrellasagen/; and Mads Klitgaard, "I en af Danmarks største sager om deling af seksuelle optagelser slipper mange unge med bøder", *Berlingske*, August 29, 2019, https://www.berlingske.dk/samfund/i-en-af-danmarks-stoerste-sager-om-deling-af-seksuelle-optagelser-slipper.

48 David Collingridge, *The Social Control of Technology,* (Frances Pinter, 1980). It can be succinctly summarized as: 'When change is easy, the need for it cannot be foreseen; when change is apparent, change has become expensive, difficult, and time consuming.', 11.

IEEE Conference on Computer Vision and Pattern Recognition Workshops (2019): 38–45.

Ajder, Henry, Giorgio Patrini, Francesco Cavalli, and Laurence Cullen. "The State of Deep-fakes: Landscape, Threats, and Impact." Deeptrace Labs, September 2019.

Ariely, Dan. *Predictably Irrational: The Hidden Forces That Shape Our Decisions.* London: Harper, 2009.

Balkin, Jack M. 'The Path of Robotics Law'. *California Law Review Circuit* 6 (2015): 45–60.

Barnes, Curtis and Barraclough, Tom. 'Perception Inception: Preparing for deepfakes and the synthetic media of tomorrow.' New Zealand Law Foundation, 21 May 2019, https://www.brainbox.institute/report.

Barr, Caelainn, Alexandra Topping, and Owen Bowcott. "Rape prosecutions in England and Wales at lowest level in a decade." *The Guardian,* September 12, 2019, https://www.theguardian.com/law/2019/sep/12/prosecutions-in-england-and-wales-at-lowest-level-in-a-decade.

Barr, Caelainn, and Owen Bowcott. "Fall in rape convictions 'due to justice system at breaking point'." *The Guardian*, December 19, 2019, https://www.theguardian.com/law/2019/dec/17/fall-in-convictions-due-to-justice-system-at-breaking-point.

Calo, Ryan. "Artificial Intelligence Policy: A Primer and Roadmap." *U.C.D. L. Rev.* 51 (2017): 399-435.

Chander, Anupam. "The Racist Algorithm?" *Mich. L. Rev.* 115 (2017): 1023

Chesney, Robert, and Danielle Keats Citron. "Deep-fakes: A Looming Challenge for Privacy, Democracy, and National Security." *California Law Review* 107, no. 6 (December 1, 2019): 1753

Chesney, Robert and Danielle Keats Citron. "21st Century-Style Truth Decay: Deep-fakes and the Challenge for Privacy, Free Expression, and National Security." *Maryland Law Review* 78, no. 4 (2019): 882-891.

Chesney, Robert, Danielle Citron, and Quinta Jurecic. "That Pelosi Video: What To Do About 'Cheapfakes' in 2020." *Lawfare* (May 29, 2019), https://www.lawfareblog.com/about-pelosi-video-what-do-about-cheapfakes-2020.

Cohen, Julie E., *Between Truth and Power: The Legal Constructions of Informational Capitalism.* Oxford: OUP, 2019.

Crown Prosecution Service. "CPS Legal Guidance: Revenge Pornography – Guidelines on prosecuting the offence of disclosing private sexual photographs and films", revised: 24 January 2017 to add a section on sexting.
https://www.cps.gov.uk/legal-guidance/revenge-pornography-guidelines-prosecuting-offence-disclosing-private-sexual.

Crown Prosecution Service. "Violence against Women and Girls (VAWG) Strategy 2017–2020." (London: CPS, 2017),
https://www.cps.gov.uk/sites/default/files/documents/publications/VAWG-Strategy-2017-2020.pdf.

Crown Prosecution Service. "CPS Legal Guidance: Exhibits." revised: 9 April 2018 (https://www.cps.gov.uk/legal-guidance/exhibits).

Crown Prosecution Service. "CPS Legal Guidance: Social Media – Guidelines on prosecuting cases involving communications sent via social media." revised: 21 August 2018.
https://www.cps.gov.uk/legal-guidance/social-media-guidelines-prosecuting-cases-involving-communications-sent-social-media.

Collingridge, David. *The Social Control of Technology*. Frances Pinter, 1980

Deeks, Ashley. "The Judicial Demand for Explainable Artificial Intelligence," *Columbia Law Review* 119, no. 7 (2019): 1829-50.

Europol, "Do Criminals Dream of Electric Sheep?" European Union Agency for Law Enforcement Cooperation 2019.

"Fakta: Flest har fået bøder i Umbrella-sagen." *Jyllandsposten*, January 25, 2020.
https://jyllands-posten.dk/indland/politiretsvaesen/ECE11899455/fakta-flest-har-faaet-boeder-i-umbrellasagen/.

Farid, Hany. "Image Forensics." *Annual Review of Vision Science* 5, no.1 (2019): 549-573.

Franks, Mary Anne, and Ari Ezra Waldman. "Sex, Lies, and Videotape: Deep-fakes and Free Speech Delusions." *Maryland Law Review* 78, no. 4 (2019): 892-898.

Garfinkel, Ben, and Allan Dafoe. "How does the offense-defense balance scale?" *Journal of Strategic Studies.* 42, vol.26 (2019): 736-763.

Glăveanu, Vlad P.. "What Can Be Done with an Egg? Creativity, Material Objects, and the Theory of Affordances." *The Journal of Creative Behavior* 46, no.3 (2012): 192–208.

Gibson, James J. *The Ecological Approach To Visual Perception.* Hillsdale, NJ: Erlbaum, 1986.

Hall, Holly Kathleen. "Deep-fake Videos: When Seeing Isn't Believing." *Catholic University Journal of Law and Technology* 27, no. 1 (Fall 2018): 51-76.

Hamilton, Rebecca J. "New Technologies in International Criminal Investigations." *Proceedings of the ASIL Annual Meeting* 112 (2018): 131–33. doi:10.1017/amp.2019.18.

Hamilton. "User-Generated Evidence." *Columbia Journal of Transnational Law* 57, no.1 (2018) American University, WCL Research Paper No. 2018-11.

Kahneman, Daniel. *Thinking Fast and Slow.* London: Penguin Allen Lane, 2012.

Klitgaard, Mads. "I en af Danmarks største sager om deling af seksuelle optagelser slipper mange unge med bøder." *Berlingske*, August 29, 2019. https://www.berlingske.dk/samfund/i-en-af-danmarks-stoerste-sager-om-deling-af-seksuelle-optagelser-slipper

Liu, Hin-Yan, Matthijs Maas, John Danaher, Luisa Scarcella, Michaela Lexer, and Leonard Van Rompaey. "Artificial Intelligence and Legal Disruption: A New Model for Analysis." *Law, Innovation and Technology* 12, no.2 (2020).

Lucraft QC, Mark. *Archbold: Criminal Pleading, Evidence & Practice 2020.* London: Sweet and Maxwell, 2020.

Kertysova, Katarina. "Artificial Intelligence and Disinformation: How AI Changes the Way Disinformation is Produced, Disseminated, and Can Be Countered." *Security and Human Rights* 29 (2018) 55-81.

Maras, Marie-Helen, and Alex Alexandrou. "Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos." *The International Journal of Evidence & Proof* 2019, Vol. 23(3) 255–262.

Markou, Christopher. "Why using AI to sentence criminals is a dangerous idea." *The Conversation*, May 16, 2017. https://theconversation.com/why-using-ai-to-sentence-criminals-is-a-dangerous-idea-77734.

Michelfelder, Diane. "Postphenomenology with an Eye to the Future." In *Postphenomenological Investigations: Essays on Human-Technology Relations,* edited by Robert Rosenberger and Peter-Paul Verbeek, 237–46. Lanham: Lexington Books, 2015.

Norman, Donald A. *The Design of Everyday Things.* Cambridge, MA: MIT Press, 2016.

O'Neil, Cathy. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* London: Penguin Allen Lane, 2016.

Ormerod QC, David, and David Perry QC. *Blackstone's Criminal Practice 2019.* Oxford: OUP, 2019.

Peters, Jay. "Alphabet's Jigsaw unveils a tool to help journalists spot deep-fakes and manipulated images." *The Verge,* February 4th, 2020. https://www.theverge.com/2020/2/4/21122778/alphabet-jigsaw-assembler-tool-news-journalists-deep-fakes-images)

Rademacher, Timo. "Artificial Intelligence and Law Enforcement." In *Regulating Artificial Intelligence,* edited by Thomas Wischmeyer and Timo Rademacher, 226-255. Cham, Switzerland: Springer, 2020.

Susser, Daniel. "Invisible Influence: Artificial Intelligence and the Ethics of Choice Architectures," in *AAAI/ACM Conference on AI, Ethics, and Society (AIES '19).* New York, NY: ACM, 2019. https://doi.org/10.1145/3306618.3314286.

Susser, Daniel. "Transparent Media and the Development of Digital Habits." In *Postphenomenology and Media: Essays on Human-Technology-World Relations*, edited by Yoni Van Den Eede et al., 27-44. Lanham: Lexington Books, 2017.

Thaler, Richard H., and Cass R. Sunstein. *Nudge: Improving Decisions about Health, Wealth, and Happiness.* London: Penguin Allen Lane, 2009.

Vincent, James. "Deep-fakes detection algorithms will never be enough." *The Verge*, June 27, 2019. https://www.theverge.com/2019/6/27/18715235/deep-fakes-detection-ai-algorithms-accuracy-will-they-ever-work.

Volkert, Robert, and Henry Adjer, "Analyzing the Commoditization of Deep-fakes." *NYU Journal of Legislation and Public Policy.* February 27, 2020. https://nyujlpp.org/quorum/volkert-ajder-analyzing-the-commoditization-of-deep-fakes/ (retrieved February 28, 2020).

Zuboff, Shoshana. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power.* New York: Public Affairs, 2019.

7. ARTIFICIAL INTELLIGENCE AND LAW ENFORCEMENT: THE USE OF AI-DRIVEN ANALYTICS TO COMBAT SEX TRAFFICKING

Clotilde Sebag*

Abstract

This paper reviews the use of AI-driven analytics by law enforcement to combat sex trafficking and evaluates the impacts it is having on law enforcement. Police agencies around the world are increasingly using AI-driven analytics software to help them in the fight against sex trafficking, as this technology now allows for the huge amounts of online data left by traffickers to be collected, processed, and analysed. The ability to do so has prompted the spread of the policing approach known as data-driven policing – by using actionable intelligence in real time as provided by the AI-driven analytics software, law enforcement can develop much more proactive and in-depth responses while utilising resources more efficiently. However, this reliance on data is also a concern, as the quality of intelligence produced by AI-driven analytics is only so good as the data it processes. Data gaps and misleading data can weaken police responses. Moreover, the risk of developing incomprehensible and unaccountable AI systems is also high. It is recommended for law enforcement to establish a framework that guides the implementation of AI technologies in this domain. Overall, the use of AI-driven analytics to fight sex trafficking represents how crime prevention now lies at the intersection of technology and data.

Keywords: Artificial Intelligence, AI-driven Analytics, Sex Trafficking, Data, Law Enforcement, Analysis, Proactive, Data-driven Policing, Black Box, Data Gaps, AI framework.

Introduction

As part of the increasing use of Artificial Intelligence (AI) within law enforcement agencies,[1] the anti-trafficking community has been focusing on how technology can help in combatting sex trafficking. At the centre of this are "data-based efforts",[2] with public and private initiatives seeking to restructure how sex trafficking investigations are carried out by leveraging the power of AI, data, and advanced analytics.[3] In this paper it

1 * Clotilde Sebag is research assistant at the Institute of Security and Global Affairs of Leiden University. Based in the Intelligence and Security Research group, she is interested in the nexus between international security, intelligence, and organised crime.

Artificial Intelligence and Robotics for Law Enforcement (Interpol and UNICRI, 2019), v, https://issuu.com/unicri/docs/artificial_intelligence_robotics_la/1?ff.

2 Maria Mayorga et al., "Countering Human Trafficking Using ISE/OR Techniques," in *Emerging Frontiers in Industrial and Systems Engineering: Success through Collaboration*, ed. Harriet B. Nembhard, Elizabeth A. Cudney, and Katherine M. Coperich (CRC Press, 2019), 245, PDF.

3 Katie Kackenmeister, "AI and the fight against human trafficking," Lehigh University, last modified October 30, 2019, accessed March 3, 2020, https://engineering.lehigh.edu/news/article/ai-and-fight-against-human-trafficking.

has been chosen to use the term *AI-driven analytics* to describe the application of AI technologies to advanced analytics capabilities. While initially developed for enterprises, it is now being deployed by law enforcement agencies, prosecutors' offices, and non-governmental organisations (NGOs) across the world to combat and investigate sex trafficking.[4] Indeed, AI-driven analytics is being credited with helping rescue hundreds of victims and providing the necessary leads to bring down traffickers.[5] Some of the most popular software today include Memex, Traffic Jam, XIX, and Spotlight.

AI-driven analytics has had profound impacts on law enforcements' crime prevention capabilities and is one of the most adopted AI technologies by police agencies around the world. As such, its practical application to sex trafficking investigations merits to be carefully studied in order to gain a better understanding of its impacts on law enforcements' approaches, outlooks, and capacities. Examining this particular application also helps shed light on the wider benefits and challenges of integrating more AI technologies within the police. This article begins by contextualising what is meant by sex trafficking; then, it reviews AI-driven analytics – how the technology works and how it is being practically applied to disrupt sex trafficking. Finally, it evaluates the main impacts that it is having on law enforcement.

It finds that AI-driven analytics is extremely useful for law enforcement agencies to garner greater insights into sex trafficking patterns and trends while providing invaluable leads for investigations and operations. More than that though, the use of AI-driven analytics reflects an ongoing shift in law enforcement toward a greater reliance on big data, giving rise to an increasingly data-driven policing approach.[6] With the exponential growth in the volume of data, law enforcement will continue to require powerful analytical capabilities to process and analyse it. Thus, as represented by anti-trafficking efforts, combatting crime increasingly lies at the intersection of technology and data.

Contextualisation of Sex Trafficking

The United Nations Office on Drugs and Crime (UNODC) divides human trafficking into three categories: trafficking for sexual exploitation; trafficking for forced labour; and trafficking for other purposes. While the patterns of trafficking and the most prevalent forms of exploitation vary around the world, the most detected form is that of trafficking for sexual exploitation, with 59 percent of the detected victims in 2016 being trafficked for this reason. The second most common form of trafficking is for forced labour, which makes up 33 percent of detected victims. Nonetheless, there remain considerable gaps in our knowledge of trafficking patterns, and impunity largely prevails.[7]

The Protocol to Prevent, Suppress and Punish Trafficking in Persons Especially Women and Children, also known as the Palermo Protocol, was adopted by the United Nations in 2000 in order to establish an international approach against human trafficking. It lays down the internationally accepted definition of "Trafficking in persons" in its Article 3:

The recruitment, transportation, transfer, harbouring or receipt of persons, by means of the threat or use of force or other forms of coercion, of abduction, of fraud, of deception, of the abuse of power or of a position of vulnerability or of the giving or receiving of payments or benefits to achieve the consent of a person having control over another person, for the purpose of exploitation. Exploitation shall include, at a minimum, the exploitation of the prostitution of others or other forms of sexual exploitation, forced labour or services, slavery or practices similar to slavery, servitude or the removal of organs.[8]

The European Union Agency for Law Enforcement Training (CEPOL) identifies five processes that make up a human trafficking crime. These are: advertisement by the traffickers; finding accommodation and workplaces

4 Ranjit Bose, "Advanced analytics: opportunities and challenges," *Industrial Management & Data Systems* 109, no. 2 (2009): 155, https://doi.org/10.1108/02635570910930073.

5 Mayorga et al., "Countering Human," 244.

6 Andrew Guthrie Ferguson, *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement* (New York: New York University Press, 2017), 16.

7 UNODC, *Global Report on Trafficking in Persons 2018*, 13, 14, 29, 31, December 2018, accessed February 11, 2020, https://www.unodc.org/documents/data-and-analysis/glotip/2018/GLOTiP_2018_BOOK_web_small.pdf.

8 *Protocol to Prevent, Suppress and Punish Trafficking in Persons Especially Women and Children, supplementing the United Nations Convention against Transnational Organized Crime*, General Assembly resolution 55/25. (Nov. 2000). https://www.ohchr.org/Documents/ProfessionalInterest/ProtocolonTrafficking.pdf.

for the victims; transportation of victims; finances; and communication between the criminals, victims, and customers.[9] Each of these activities has been facilitated by developments in technology with sex traffickers constantly exploiting them to facilitate their criminal enterprise and exploitation.[10] Digital technology has increased their anonymity; facilitated the recruitment and exploitation of victims; made it easier to access expanded ranges of criminal activity; and increased the means through which to control and exploit victims.[11]

Recruitment via the Internet is increasingly part of how traffickers operate, as suitable targets can easily be identified and contacted.[12] Travel and accommodation logistics are also facilitated – bookings and financial payments can be carried out quickly and anonymously online. Furthermore, it is easier to communicate and coordinate between the different components and elements that make up a trafficking network. The exploitation of victims has also been eased by the Internet: the marketing of victims has radically changed, with most of them now being advertised online to a greater audience and potential buyers.[13] It is estimated that 63 percent of child sex trafficking survivors were advertised online at some point.[14]

While at first glance this may seem discouraging, cyberspace is one of the rare places where "the notoriously clandestine activity of human trafficking actually surfaces" with some of its processes becoming more visible, particularly the online advertisement of the victims.[15] All the operations carried out by traffickers online create and leave digital footprints – this should be seen as data that can be used to fight against them.

AI-driven Analytics

Because of the development of sex trafficking in the online domain, the approach to fighting it is increasingly data-based; law enforcement agencies need to process a colossal amount of data from disparate sources, and turn it into useful and actionable intelligence which they can act on.[16] This "ever-expanding volume, velocity, and variety of data", also known as big data, can only be processed and analysed by AI-driven analytics.[17] This section will review what is meant by this term, as well as how exactly it is being used in the fight against sex trafficking.

AI-driven analytics is a subfield of advanced analytics.[18] Advanced analytics was first developed for use in commercial businesses; it is defined as "the autonomous or semi-autonomous examination of data or content using sophisticated techniques and tools [...] to discover deeper insights, make predictions, or generate recommendations."[19] Analytics is about using analytical methods to discover and interpret meaningful patterns, correlations, and previously unseen connections in data that humans could not have detected.[20]

Within advanced analytics there are processes that are driven by Artificial Intelligence and that can 'unlock'

9 European Union Agency for Law Enforcement Training, *Trafficking in Human Beings*, accessed February 15, 2020, https://enet.cepol.europa.eu/index.php?id=online-courses&no_cache=1.

10 Sigrid Raets and Jelle Janssens, "Trafficking and Technology: Exploring the Role of Digital Communication Technologies in the Belgian Human Trafficking Business," *European Journal on Criminal Policy and Research*, 2019, 2, accessed February 20, 2020, https://doi.org/10.1007/s10610-019-09429-z ; and John Richmond, "Taking a Lesson From Traffickers: Harnessing Technology To Further the Anti-Trafficking Movement's Principal Goals," speech, April 8, 2019, U.S. Mission to the OSCE, last modified April 8, 2019, accessed February 26, 2020, https://osce.usmission.gov/taking-a-lesson-from-traffickers-harnessing-technology-to-further-the-anti-trafficking-movement/.

11 Inter-agency Coordination Group Against Trafficking in Persons, *Human Trafficking and Technology: Trends, Challenges and Opportunities*, issue brief no. 7, 1, 2, 2019, accessed March 4, 2020, https://icat.network/sites/default/files/publications/documents/Human%20trafficking%20and%20technology%20trends%20challenges%20and%20opportunities%20WEB....pdf.

12 Jorn van Rij and Ruth McAlister, "Using Criminal Routines and Techniques to Predict and Prevent the Sexual Exploitation of Eastern-European Women in Eastern Europe," in *The Palgrave International Handbook of Human Trafficking*, ed. John Winterdyk and Jackie Jones (Palgrave Macmillan, 2020), 1693, PDF; and Raets and Janssens, "Trafficking and Technology," 6.

13 Raets and Janssens, "Trafficking and Technology," 8-9.

14 "Thorn, Digital Defenders of Children," McCain Institute, accessed March 8, 2020, https://www.mccaininstitute.org/thorn-digital-defenders-children/.

15 Raets and Janssens, "Trafficking and Technology," 11.

16 M.I Pramanik et al., "Big data analytics for security and criminal investigations," *WIREs Data Mining and Knowledge Discovery* 7, no. 4 (July/August 2017): 1, accessed February 13, 2020, https://doi.org/10.1002/widm.1208.

17 Ferguson, *The Rise,* 16.

18 Janet Wagner, "Advanced analytics vs. artificial intelligence: How are they different?," ZyloTech, https://www.zylotech.com/blog/advanced-analytics-vs.-artificial-intelligence-how-are-they-different.

19 Gartner, quoted in: Jeremy Rose et al., "The Advanced Analytics Jumpstart: Definition, Process Model, Best Practices," *Journal of Information Systems and Technology Management* 14, no. 3 (September/December 2017), accessed February 10, 2020, https://doi.org/10.4301/s1807-17752017000300003.

20 Tom Sabo, "An Artificial Intelligence Framework to Combat Human Trafficking," lecture, Bright Talk, last modified August 5, 2019, accessed February 14, 2020, https://www.brighttalk.com/webcast/17718/368329/an-artificial-intelligence-framework-to-combat-human-trafficking.

more from data; this is referred to as AI-driven analytics. AI technologies that are applied include "data/text mining, machine learning, pattern matching, forecasting, visualization, semantic analysis, sentiment analysis, network and cluster analysis, multivariate statistics, graph analysis, simulation, complex event processing, neural networks."[21] AI is one of the most significant sub-fields within analytics as it has transformed data mining methods, and as such allows for "new efficiencies and insights" to be gleaned from data, thereby generating much more value from it than before.[22]

AI-driven Analytics to Combat Sex Trafficking

With law enforcement agencies often lacking the infrastructure, resources, and skills to develop an AI-driven analytics capability, we are seeing that in this field, the actors involved in developing new technological initiatives are varied, "multi-sector, multidisciplinary and international."[23] Commercial enterprises work alongside actors from the public sector including law enforcement agencies, academia, and non-governmental organisations. Law enforcement agencies, prosecutors' offices, District Attorney offices, and NGOs across the world are using these tools: more than 30 law enforcement agencies across the world are using Memex software to help in human trafficking investigations; Spotlight is currently used by more than 4,000 law enforcement officials across 780 agencies in the United States, having helped to identify more than 6,300 victims of sex trafficking in the US so far;[24] Traffic Jam has long-standing partnerships with law enforcement agencies, prosecutors' offices and non-profit organisations.[25]

These solutions have been developed for use specifically in the anti-sex trafficking field – they leverage the online connections that buyers and sellers have and the data this creates. First, cutting-edge "search and analytics engine[s]" are designed for law enforcement.[26] Then, online data is collected, after which a host of AI technologies mine it to garner deeper insights, derive patterns, conduct analyses, and provide predictions autonomously[27] at a "scale, speed and depth" that cannot be matched by human analysts.[28] It then delivers the results of the analyses in intuitive visual tools that aid in decision-making. In this field, most of the data that is used comes from online ads for sexual services.

The search systems are very "user-centric" – they are developed around user requirements and specifications.[29] Based on these requirements, automated web crawlers are written to extract relevant data from websites across the Internet.[30] Different companies build their own crawling architecture to reflect their data needs.[31] These crawlers browse the millions of online ads which contain "extremely rich data" such as images, phone numbers, locations, working names, physical attributes, titles of ads, rates, etc.;[32] and there is plenty of data to collect – it is estimated that every day 400,000 to 500,000 "new adult services, ads, reviews, and discussion threads" are posted.[33]

Despite the vast amount of publicly available information, there exists a challenge in exploiting all the available sources as the advertisements are usually a mix of structured and unstructured data, the latter being extremely

21 Gartner, "Advanced Analytics," Gartner Glossary.
22 Paul B. Symon and Arzan Tarapore, *Defense Intelligence Analysis in the Age of Big Data*, 4, March 30, 2016, accessed March 2, 2020, https://css.ethz.ch/en/services/digital-library/articles/article.html/195263/pdf.; and Pramanik et al., "Big data," 2.
23 Nadya Bliss, "Towards a Pipeline – Technology, Techniques and Training," in *Code 8.7: Conference Proceedings* (United Nations University, 2019), 13, accessed February 28, 2020, https://collections.unu.edu/eserv/UNU:7313/UNU_Code8.7_Final.pdf.
24 Andrea Fox, "3 Tools Helping Law Enforcement Agencies Stop Sex Trafficking," Efficient Gov, last modified February 22, 2017, accessed March 2, 2020, https://www.efficientgov.com/technology/articles/3-tools-helping-law-enforcement-agencies-stop-sex-trafficking-nq91QzbMdg3MasTS/ .
25 Brody, "How Artificial,".
26 Mayank Kejriwal and Pedro Szekely, "Technology-assisted Investigative Search: A Case Study from an Illicit Domain," *CHI'18 Extended Abstracts*, 2018, 2, accessed February 23, 2020, https://doi.org/10.1145/3170427.3174364.
27 Rose et al., "The Advanced".
28 Bracket Foundation, Bracket Capital, and Value for Good, *Artificial Intelligence: Combating Online Sexual Abuse of Children*, 14, 2019, accessed February 2020, https://www.flipbookpdf.net/web/files/uploads/765c57681ad3259906107226b5934880ca9dbebfFBP17764427.pdf.
29 Kejriwal and Szekely, "Technology-assisted Investigative," 2.
30 "Memex Human Trafficking Summary," DeepDive, accessed February 24, 2020, http://deepdive.stanford.edu/showcase/memex.
31 Kejriwal and Szekely, "Technology-assisted Investigative," 2.
32 "Memex Human," DeepDive.
33 "The TellFinder Alliance," TellFinder, accessed February 27, 2020, https://www.tellfinder.com/.; and Andreas Olligschlaeger, (Senior Research Scientist Marinus Analytics LLC), interview by the author, Online, February 4, 2020.

labour intensive to collect and mine.[34] Furthermore, much of the interesting data is unstructured and thus 'hidden' within ads.[35] Traffickers may use Unicode to hide important information such as telephone numbers or addresses.[36] Moreover, to avoid detection, they often post temporary ads, or frequently change small details such as the phone numbers or the names in the ads.[37] Thus, reliance on human manual collection and analysis or basic analytical tools is not enough; AI can help identify and gather critical information, including that which is in freeform or unstructured data.[38]

Once extracted, the data needs to be indexed and catalogued in order to analyse it. Traffic Jam, for example, has built a database of "images, phone numbers, and location data" that goes back to 2011.[39] By having crawlers scrape websites very regularly, ads can be indexed even if posted only temporarily. From 2014 to 2016, the MEMEX team indexed "80 million web pages and 40 million images."[40] With these huge databases, AI technologies such as machine learning, natural language processing, and text and image mining are applied to analyse the collected information and gain deeper insights on the phenomenon. These databases with terabytes of data can also be searched for very broad criteria. For example, one could run a phone number found on a suspect ad, and instantly find other ads on other sites that are linked to the same phone number.

Most of the current applications involve applying AI technologies, such as machine learning, to identify cases of sex trafficking in escort ads, and discover trafficking networks.[41] AI can help detect which ads are displaying trafficked people by discovering "online behavioural signals in the ads" that indicate that the person might be being trafficked. Ads can be analysed with mining techniques that rely on visual cues, for instance, if tattoos are present in multiple images; or textual cues, including "specific styles of ad-writing; keywords, etc." to help differentiate suspicious from non-suspicious ads.[42] Another key task of advanced AI-driven analytics is to "group ads by *true* owner", and not by the claimed authors.[43] Entity resolution involves AI discovering "a set of explicit links between entities extracted from different sources."[44] As a result, entities such as phone numbers and email addresses can be used to find patterns and connect disparate ads that are most likely written by the same traffickers across different geographical areas.[45] This is known as record linkage and enables law enforcement to identify a same entity that is referenced across different sources and different times.[46] Some of the software, including Traffic Jam and XIX, also rely on image analysis capacities – these tools can scrape images or identify similar photos that appear in disconnected ads.[47] Traffic Jam also relies on a facial recognition software to identify trafficking victims – law enforcement can run a photo of a missing person against other photos in the database to help identify if the person is being advertised online.[48] Finally, AI-driven analytics can also help identify critical links and nodes within a network and its structure. Subsequently, it is possible to expose the most valuable patterns and trends, such as meaningful relationships, key people at the core of the network, and flows between nodes (e.g. of goods, finance, or information).

Further value can be garnered from data through innovative visualisation of the data and the insights garnered through analysis. Building a cohesive and contextualised understanding of human trafficking based on a

34 Pedro Szekely et al., "Building and Using a Knowledge Graph to Combat Human Trafficking," in *The Semantic Web–ISWC 2015 14th International Semantic Web Conference Bethlehem, PA, USA, October 11–15, 2015 Proceedings, Part II*, ed. Marcelo Arenas, et al. (Cham: Springer, 2015), 9367:206; and Rose et al., "The Advanced Analytics".

35 Renata A. Konrad et al., "Overcoming human trafficking via operations research and analytics: Opportunities for methods, models, and applications," *European Journal of Operational Research* 259, no. 2 (June 1, 2017): 738, accessed March 3, 2020, https://doi.org/10.1016/j.ejor.2016.10.049.

36 Szekely et al., "Building and Using," 9367:206.

37 Konrad et al., "Overcoming human," 735.

38 Kackenmeister, "AI and the fight," Lehigh University.

39 Rachel Kaser, "This company is using facial recognition to fight human trafficking," The Next Web, last modified June 1, 2018, accessed March 9, 2020, https://thenextweb.com/world/2018/06/01/this-company-is-using-facial-recognition-to-fight-human-trafficking/; and Olligschlaeger, interview by the author.

40 Anastasija Mensikova and Chris A. Mattmann, "Ensemble Sentiment Analysis to Identify Human Trafficking in Web Data," *GTA3 2018*, February 2018, 1, accessed March 5, 2020, https://doi.org/10.475/123_4.

41 Rebecca S. Portnoff et al., "Backpage and Bitcoin: Uncovering Human Traffickers," *KDD 2017*, 2017, 1596, accessed March 5, 2020, https://doi.org/10.1145/3097983.3098082.

42 Raets and Janssens, "Trafficking and Technology," 13; and Mensikova and Mattmann, "Ensemble Sentiment," 1.

43 Portnoff et al., "Backpage and Bitcoin," 1596, emphasis added by the author.

44 Szekely et al., "Building and Using," 9367:215.

45 Szekely et al., "Building and Using," 9367:2047, 215; and Raets and Janssens, "Trafficking and Technology,"

46 Kyle Rossetti and Rebecca Bilbro, "Basics of Entity Resolution with Python and Dedupe," District Data Labs, accessed March 3, 2020, https://www.districtdatalabs.com/basics-of-entity-resolution.

47 Sean Captain, "This AI can spot signs of human trafficking in online sex ads," *Fast Company*, November 4, 2019, accessed March 6, 2020, https://www.fastcompany.com/90424645/how-image-recognition-ai-is-busting-sex-traffickers.

48 Kaser, "This company," The Next Web.

flood of data relies on AI building visualisations such as interactive user interfaces, dashboards, graphs, heat map matrices, "spark lines to show temporal fluctuation", or maps with arrows indicating the different cities where the same phone number is being advertised.[49] Analytic visualisation enables officers to respond more effectively to the data as pertinent findings become more intuitive to read and understand, and thus, more well-informed decisions can be taken.[50] This visualisation tool is also highly relevant when communicating findings to policy- and decision-makers, who may lack subject-specific knowledge.[51]

Evaluation

Overall, AI-driven analytics assists investigators in: finding ads that are most likely to be advertising trafficked people; clarifying the habits and structure of trafficking organisations; prioritising scarce law enforcement resources; and detecting the possible and most effective interventions.[52] It is important to emphasise that AI-driven analytics does not remove humans from the fight against human trafficking. Instead, this technology is mostly used to create investigative *leads*, with "human intelligence and decision-making" still required for action to be taken.[53] Indeed, when ads are highlighted as potentially being cases of trafficking it is still up to human agents to review the classification, decide how and where to act, and how to allocate resources. What we are seeing is a "tandem" between traditional techniques and new technologies that assist law enforcement.[54] Based on this review of the technology as applied to sex trafficking investigations, this section will evaluate three significant impacts that AI-driven analytics has had on law enforcement.

More Proactive Responses

By unlocking insights and connections from data that would have otherwise not been found, AI-driven analytics is transforming the very approach that officers are taking towards sex investigations. This new approach to crime fighting has been coined "data driven policing", and it involves the automatic collection and analysis of large datasets, in this case through AI-driven analytics, allowing police officers to shape their measures and responses around real-time information gathered.[55]

With more information, law enforcement gains a better understanding of which leads are more urgent and how limited resources should be prioritised.[56] Automation is also key in reducing the time needed to carry out investigations and operations.[57] This not only frees up an incredible amount of time for officers to focus on other key tasks, but it also increases the number of investigations that can be carried out, likely increasing the number of rescued victims. For instance, XIX is able to compile an intelligence report for law enforcement to act on in six hours instead of the previous 22 days,[58] and Manhattan's District Attorney has been able to increase the number of human trafficking investigations they carry out yearly from 30 to 300 since working with an AI-driven analytics software.[59]

Moreover, real time information allows officers to design more "purposeful interventions", and have a much more proactive approach to policing.[60] The visualisation of the various links that make up a trafficking network allows law enforcement officials to develop operations that block essential flows between traffickers, or that

49 Mayorga et al., "Countering Human," 248.
50 Marco Annunziata, "AI And Data Visualization: How AI Helps Companies See Through The Fog Of Data," *Forbes*, February 9, 2019, accessed March 1, 2020, https://www.forbes.com/sites/marcoannunziata/2019/02/09/ai-and-data-visualization-how-ai-helps-companies-see-through-the-fog-of-data/#73268a7f7cf3.
51 Davina Durgana, "Mining Government Data to Reach Target 8.7," in *Code 8.7: Conference Proceedings* (United Nations University, 2019), 6, accessed March 2, 2020, https://collections.unu.edu/eserv/UNU:7313/UNU_Code8.7_Final.pdf.
52 "Memex Human," DeepDive.
53 Bracket Foundation, Bracket Capital, and Value for Good, *Artificial Intelligence*, 20.
54 Raets and Janssens, "Trafficking and Technology," 18.
55 https://cops.usdoj.gov/RIC/Publications/cops-w0558-pub.pdf; and Ferguson, *The Rise.*
56 Babuta, Oswald, and Janjeva, *Artificial Intelligence*, 7.
57 Ferguson, *The Rise*, 103.
58 Captain, "This AI can spot,".
59 Brody, "How Artificial,".
60 *Code 8.7: Conference Proceedings* (United Nations University, 2019), 11, accessed March 7, 2020, https://collections.unu.edu/eserv/UNU:7313/UNU_Code8.7_Final.pdf.

target critical actors which, if taken out, might destabilise or make the network collapse.[61] Furthermore, they can set up operations that target other actors who "set up, contribute and perpetuate" sex trafficking such as financial backers, customers, recruiters, and so on;[62] and by detecting trafficking routes, transportation methods, safe-houses, and recruitment hotspots, it is easier to plan rapid operations that can rescue victims at these points.[63] Finally, understanding the flows, and the temporal and geographical patterns of trafficking, is crucial in developing adapted and decisive responses. For instance, data from Traffic Jam helped disprove the common belief that there was an increase in human trafficking in the times before and during events like the Super Bowl.[64] This knowledge prevents law enforcement from misplacing limited resources during events like this.

Law Enforcement's Increased Reliance on Data

With the ongoing production of digital data and the benefits that a data-driven approach brings, it seems likely that law enforcement's use of AI-driven analytics will only become greater. However, this presents certain limitations. This technology depends on large and accurate datasets; yet, there exist significant "data gaps".[65] Data on "the distribution of victims, traffickers, buyers, and exploiters" is often unreliable or partial.[66] Furthermore, there are knowledge gaps of trafficking flows in large parts of the world, such as areas in Africa and the Middle East, and estimates mostly stem from non-methodical or incomplete data based on reported cases only.[67] This paints an inaccurate picture of the "nature and scale" of human trafficking globally, thus hindering the global framework and responses envisaged by the anti-trafficking community.[68]

Moreover, since "Big data collection will not count those whom it cannot see",[69] reliance on data constrains law enforcement to only act on crimes that create online data at some point. Sex trafficking is perhaps one of the crimes that can best be combatted through data-driven efforts due to the online advertisements placed by traffickers.[70] Still, these represent only a fraction of the entire sex trafficking chain; most of it remains invisible, creating little data and thus little information to act on. Moreover, reliance on data also presents a security risk as criminals may discover how to better hide their data and make it unusable; or they may alter data to feed disinformation to law enforcement.

AI as a Black Box

Law enforcement agencies often do not have the in-house infrastructure, resources, nor skills to develop an advanced AI-driven analytics capability.[71] It seems therefore likely that the current commercial model whereby private companies develop tools which they contract out to law enforcement agencies will not change. However, this poses several concerns as the software is often proprietary and there is a lack of transparency regarding how algorithms were created and how they work.[72]

One of the risks involved is that private companies develop what is known as 'black box' AI systems. This refers to cases where, because of the complexity of the technology, AI systems' data processing methods are opaque and may not be understood by its operators.[73] The lack of interpretability may render law

61 Farrell and de Vries, "Measuring the Nature," 157.
62 Rosalva Resendiz and Lucas E. Espinoza, "The International Law Enforcement Community: Cooperative Efforts in Combatting Human Trafficking," in *The SAGE Handbook of Human Trafficking and Modern Day Slavery*, ed. Jennifer Bryson Clark and Sasha Poucki (Sage, 2019), 480, accessed February 12, 2020.
63 Konrad et al., "Overcoming human," 737.
64 Alexandra Ossola, "AI Tool Helps Law Enforcement Find Victims of Human Trafficking," *Futurism*, April 16, 2018, accessed March 5, 2020, https://futurism.com/ai-tool-law-enforcement-stop-human-trafficking.
65 Ferguson, *The Rise*, 184.
66 Konrad et al., "Overcoming human," 736.
67 UNODC, *Global Report*, 15.
68 Farrell and de Vries, «Measuring the Nature,» 148.
69 Ferguson, *The Rise*, 185.
70 Raets and Janssens, "Trafficking and Technology," 11.
71 *Artificial Intelligence*, 24.
72 Randy Rieland, "Artificial Intelligence Is Now Used to Predict Crime. But Is It Biased?," *Smithsonian Magazine*, March 5, 2018, accessed March 7, 2020, https://www.smithsonianmag.com/innovation/artificial-intelligence-is-now-used-predict-crime-is-it-biased-180968337/.
73 Alexander Babuta, Marion Oswald, and Christine Rinik, *Machine Learning Algorithms and Police Decision-Making: Legal, Ethical and Regulatory Challenges*, report no. 3-18, 17, September 2018, accessed March 9, 2020, https://rusi.org/sites/default/files/201809_whr_3-18_machine_learning_algorithms.pdf.pdf.

enforcement officials unable to explain the evidence obtained against suspected traffickers during court cases, which can undermine "the transparency of the overall justice process"[74] and hinder sex trafficking convictions.[75] Furthermore, algorithms and the data sets they are based on are rarely subject to review nor audited, potentially allowing biases to be embedded into the programmes.[76] In order for police officers to retain accountability over the final decision-making, it would be necessary for the systems to be designed so that "non-technically skilled users can interpret and critically assess key technical information".[77]

Despite this risk, Babuta, Oswald and Rinik stress that the multidisciplinary nature of the partners involved in developing these kinds of software can actually reduce the risk of developing and implementing biased and untransparent algorithms. There are "technical, organisational and legal complexities" involved in such a task and it is suggested that cross-sectorial partnerships is a way for all these complexities to be considered.[78] In addition, many companies involved with the anti-trafficking community have been developing open source software, a trend which can mitigate the black box effects.

Nonetheless, a wider and overarching framework should be established to guide these efforts and the wider spread of AI technologies within law enforcement. Establishing ethical and legal guidelines is primordial, but it must also be considered how law enforcement training should evolve to reflect these changes. Officers should be able to comprehend and work with AI technologies in a transparent and explainable manner; but they must also stay abreast of the latest technological innovations to understand how these may practically impact law enforcement. The full range of consequences that incorporating AI into the criminal justice system would have has not yet been fully forecasted,[79] and while the benefits are numerous, this review has also demonstrated that important limitations remain, and should continue to be monitored as AI technologies are more widely used by law enforcement.

Conclusion

Technology has transformed and reconfigured sex trafficking, but also anti-trafficking efforts: with the prevalence of available data, a more proactive and data-driven policing approach has been adopted. AI-driven analytics is a particularly important technology that can analyse large datasets and generate valuable insights and actionable intelligence for law enforcement.

It is an instance of fruitful partnership between the tech and anti-trafficking fields. Its application to fight sex trafficking is having concrete and positive impacts on how law enforcement can combat this crime. More than just building software though, the use of this technology is impacting the very conception of how crime is to be fought – at the intersection of technology and data. Advanced AI-driven analytic innovations should be expected to continue spreading, becoming an indispensable tool for law enforcement. It is thus necessary to understand not just the technology, but also take into account the wider contexts it operates in, e.g. legal and societal. An overarching framework surrounding the use of AI-advanced analytics in law enforcement should be established; and while the application of AI-driven analytics is now solely to sex trafficking investigations, new opportunities to harness previously unusable data may appear and extend these technologies to fight other types of trafficking or crime in general.

74 Babuta, Oswald, and Rinik, *Machine Learning*, 17.
75 Olligschlaeger, interview by the author.
76 Ronald Yu and Gabriele Spina Ali, "What's Inside the Black Box? AI Challenges for Lawyers and Researchers," *Legal Information Management* 19, no. 1 (March 2019): accessed March 2, 2020, https://doi.org/10.1017/S1472669619000021.
77 Babuta, Oswald, and Janjeva, *Artificial Intelligence*, viii.
78 Babuta, Oswald, and Rinik, *Machine Learning*, 27.
79 Babuta, Oswald, and Rinik, *Machine Learning*, 10.

References

Annunziata, Marco. "AI And Data Visualization: How AI Helps Companies See Through The Fog Of Data." *Forbes*, February 9, 2019. Accessed March 1, 2020. https://www.forbes.com/sites/marcoannunziata/2019/02/09/ai-and-data-visualization-how-ai-helps-companies-see-through-the-fog-of-data/#73268a7f7cf3.

Artificial Intelligence and Robotics for Law Enforcement. Interpol | UNICRI, 2019. https://issuu.com/unicri/docs/artificial_intelligence_robotics_la/1?ff.

Babuta, Alexander, Marion Oswald, and Ardi Janjeva. *Artificial Intelligence and UK National Security: Policy Considerations*. Royal United Services Institute, 2020. https://rusi.org/sites/default/files/ai_national_security_final_web_version.pdf.

Babuta, Alexander, Marion Oswald, and Christine Rinik. *Machine Learning Algorithms and Police Decision-Making: Legal, Ethical and Regulatory Challenges*. Report no. 3-18. September 2018. Accessed March 9, 2020. https://rusi.org/sites/default/files/201809_whr_3-18_machine_learning_algorithms.pdf.pdf.

Bliss, Nadya. "Towards a Pipeline – Technology, Techniques and Training." In *Code 8.7: Conference Proceedings*, 12-13. United Nations University, 2019. Accessed February 28, 2020. https://collections.unu.edu/eserv/UNU:7313/UNU_Code8.7_Final.pdf.

Bose, Ranjit. "Advanced analytics: opportunities and challenges." *Industrial Management & Data Systems* 109, no. 2 (2009): 155-72. https://doi.org/10.1108/02635570910930073.

Bracket Foundation, Bracket Capital, and Value for Good. *Artificial Intelligence: Combating Online Sexual Abuse of Children*. 2019. Accessed February 2020. https://www.flipbookpdf.net/web/files/uploads/765c57681ad3259906107226b5934880ca9dbebfFBP17764427.pdf.

Brody, Liz. "How Artificial Intelligence Is Tracking Sex Traffickers." *OneZero*, May 8, 2019. Accessed March 11, 2020. https://onezero.medium.com/how-artificial-intelligence-is-tracking-sex-traffickers-276dcc025ecd.

Captain, Sean. "This AI can spot signs of human trafficking in online sex ads." *Fast Company*, November 4, 2019. Accessed March 6, 2020. https://www.fastcompany.com/90424645/how-image-recognition-ai-is-busting-sex-traffickers.

Cockayne, James. "Creating Incentives for Action – Research, Regulation and Rewards." In *Code 8.7: Conference Proceedings*, 14-15. United Nations University, 2019. Accessed March 11, 2020. https://collections.unu.edu/eserv/UNU:7313/UNU_Code8.7_Final.pdf.

Dubrawski, Artur, Kyle Miller, Matthew Barnes, Benedikt Boecking, and Emily Kennedy. "Leveraging Publicly Available Data to DiscernPatterns of Human-Trafficking Activity." *Journal of Human Trafficking* 1, no. 1 (2015): 65-85. Accessed March 12, 2020. https://doi.org/10.1080/23322705.2015.1015342.

Duong, Kim Anh. "A Comprehensive Gender Framework to Evaluate Anti-trafficking Policies and Programs." In *The Palgrave International Handbook of Human Trafficking*, edited by John Winterdyk and Jackie Jones, 247-69. Palgrave Macmillan, 2020. PDF.

Durgana, Davina. "Mining Government Data to Reach Target 8.7." In *Code 8.7: Conference Proceedings*, 5-6. United Nations University, 2019. Accessed March 2, 2020. https://collections.unu.edu/eserv/UNU:7313/UNU_Code8.7_Final.pdf.

European Union Agency for Law Enforcement Training. *Trafficking in Human Beings*. Accessed February 15, 2020. https://enet.cepol.europa.eu/index.php?id=online-courses&no_cache=1.

Farrell, Amy, and Ieke de Vries. "Measuring the Nature and Prevalence of Human Trafficking." In *The Palgrave International Handbook of Human Trafficking*, edited by John Winterdyk and Jackie Jones, 147-62. Palgrave Macmillan, 2020. PDF.

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

Fox, Andrea. "3 Tools Helping Law Enforcement Agencies Stop Sex Trafficking." Efficient Gov. Last modified February 22, 2017. Accessed March 2, 2020. https://www.efficientgov.com/technology/articles/3-tools-helping-law-enforcement-agencies-stop-sex-trafficking-nq91QzbMdg3MasTS/.

Gartner. "Advanced Analytics." Gartner Glossary. Accessed January 24, 2020. https://www.gartner.com/en/information-technology/glossary/advanced-analytics.

Hoang, Thi. *Addressing Tomorrow's Slavery Today*. Compiled by OHCHR. Accessed March 9, 2020. https://www.ohchr.org/Documents/Issues/Slavery/SR/AddressingTomorrowSlaveryToday/TechnologyAgainstTrafficking.pdf.

Inter-agency Coordination Group Against Trafficking in Persons. *Human Trafficking and Technology: Trends, Challenges and Opportunities*. Issue brief no. 7. 2019. Accessed March 4, 2020. https://icat.network/sites/default/files/publications/documents/Human%20trafficking%20and%20technology%20trends%20challenges%20and%20opportunities%20WEB....pdf.

Jones, Paul, and Chloe Setter. "Finding Hidden Populations – Orphanage Trafficking." In *Code 8.7: Conference Proceedings*, 10-11. United Nations University, 2019. Accessed March 7, 2020. https://collections.unu.edu/eserv/UNU:7313/UNU_Code8.7_Final.pdf.

Kackenmeister, Katie. "AI and the fight against human trafficking." Lehigh University. Last modified October 30, 2019. Accessed March 3, 2020. https://engineering.lehigh.edu/news/article/ai-and-fight-against-human-trafficking.

Kaser, Rachel. "This company is using facial recognition to fight human trafficking." The Next Web. Last modified June 1, 2018. Accessed March 9, 2020. https://thenextweb.com/world/2018/06/01/this-company-is-using-facial-recognition-to-fight-human-trafficking/.

Kejriwal, Mayank, and Pedro Szekely. "Technology-assisted Investigative Search: A Case Study from an Illicit Domain." *CHI'18 Extended Abstracts*, 2018, 1-9. Accessed February 23, 2020. https://doi.org/10.1145/3170427.3174364.

Konrad, Renata A., Andrew C. Trapp, Timothy M. Palmbach, and Jeffrey S. Blom. "Overcoming human trafficking via operations research and analytics: Opportunities for methods, models, and applications." *European Journal of Operational Research* 259, no. 2 (June 1, 2017): 733-45. Accessed March 3, 2020. https://doi.org/10.1016/j.ejor.2016.10.049.

Landman, Todd. "Vulnerability Mapping and Modelling." In *Code 8.7: Conference Report*, 1-2. United Nations University, 2019. Accessed March 3, 2020. https://collections.unu.edu/eserv/UNU:7313/UNU_Code8.7_Final.pdf.

Leventhal, Barry. "An introduction to data mining and other techniques for advanced analytics." *Journal of Direct, Data and Digital Marketing Practice* 12, no. 2 (2010): 137-53.

Mayorga, Maria, Laura Tateosian, German Velasquez, Reza Amindarbari, and Sherrie Caltagirone. "Countering Human Trafficking Using ISE/OR Techniques." In *Emerging Frontiers in Industrial and Systems Engineering: Success through Collaboration*, edited by Harriet B. Nembhard, Elizabeth A. Cudney, and Katherine M. Coperich, 237-57. CRC Press, 2019. PDF.

"Memex Human Trafficking Summary." DeepDive. Accessed February 24, 2020. http://deepdive.stanford.edu/showcase/memex.

Mensikova, Anastasija, and Chris A. Mattmann. "Ensemble Sentiment Analysis to Identify Human Trafficking in Web Data." *GTA3 2018*, February 2018, 1-6. Accessed March 5, 2020. https://doi.org/10.475/123_4.

Olligschlaeger, Andreas, (Senior Research Scientist Marinus Analytics LLC). Interview by the author. Online. February 4, 2020.

Ossola, Alexandra. "AI Tool Helps Law Enforcement Find Victims of Human Trafficking." *Futurism*, April 16, 2018. Accessed March 5, 2020. https://futurism.com/ai-tool-law-enforcement-stop-human-trafficking.

Pellerin, Cheryl. "DARPA Program Helps to Fight Human Trafficking." Defense. Last modified January 4, 2017. Accessed March 4, 2020. https://www.defense.gov/Explore/News/Article/Article/1041509/darpa-program-helps-to-fight-human-trafficking/.

Portnoff, Rebecca S., Danny Yuxing Huang, Periwinkle Doerfler, Sadia Afroz, and Damon McCoy. "Backpage and Bitcoin: Uncovering Human Traffickers." *KDD 2017*, 2017, 1595-604. Accessed March 5, 2020. https://doi.org/10.1145/3097983.3098082.

Pramanik, M.I, Raymond Y.K. Lau, Wei T. Yue, Yunming Ye, and Chunping Li. "Big data analytics for security and criminal investigations." *WIREs Data Mining and Knowledge Discovery* 7, no. 4 (July/August 2017): 1-19. Accessed February 13, 2020. https://doi.org/10.1002/widm.1208.

Protocol to Prevent, Suppress and Punish Trafficking in Persons Especially Women and Children, supplementing the United Nations Convention against Transnational Organized Crime, General Assembly resolution 55/25. (Nov. 2000). https://www.ohchr.org/Documents/ProfessionalInterest/ProtocolonTrafficking.pdf.

Raets, Sigrid, and Jelle Janssens. "Trafficking and Technology: Exploring the Role of Digital Communication Technologies in the Belgian Human Trafficking Business." *European Journal on Criminal Policy and Research*, 2019, 1-24. Accessed February 20, 2020. https://doi.org/10.1007/s10610-019-09429-z.

Resendiz, Rosalva, and Lucas E. Espinoza. "The International Law Enforcement Community: Cooperative Efforts in Combatting Human Trafficking." In *The SAGE Handbook of Human Trafficking and Modern Day Slavery*, edited by Jennifer Bryson Clark and Sasha Poucki, 469-85. Sage, 2019. Accessed February 12, 2020.

Richmond, John. "Taking a Lesson From Traffickers: Harnessing Technology To Further the Anti-Trafficking Movement's Principal Goals." Speech, April 8, 2019. U.S. Mission to the OSCE. Last modified April 8, 2019. Accessed February 26, 2020. https://osce.usmission.gov/taking-a-lesson-from-traffickers-harnessing-technology-to-further-the-anti-trafficking-movement/.

Rieland, Randy. "Artificial Intelligence Is Now Used to Predict Crime. But Is It Biased?" *Smithsonian Magazine*, March 5, 2018. Accessed March 7, 2020. https://www.smithsonianmag.com/innovation/artificial-intelligence-is-now-used-predict-crime-is-it-biased-180968337/.

Rose, Jeremy, Mikael Berndtsson, Gunnar Mathiason, and Peter Larsson. "The Advanced Analytics Jumpstart: Definition, Process Model, Best Practices." *Journal of Information Systems and Technology Management* 14, no. 3 (September/December 2017). Accessed February 10, 2020. https://doi.org/10.4301/s1807-17752017000300003.

Rossetti, Kyle, and Rebecca Bilbro. "Basics of Entity Resolution with Python and Dedupe." District Data Labs. Accessed March 3, 2020. https://www.districtdatalabs.com/basics-of-entity-resolution.

Sabo, Tom. "An Artificial Intelligence Framework to Combat Human Trafficking." Lecture. Bright Talk. Last modified August 5, 2019. Accessed February 14, 2020. https://www.brighttalk.com/webcast/17718/368329/an-artificial-intelligence-framework-to-combat-human-trafficking.

Symon, Paul B., and Arzan Tarapore. *Defense Intelligence Analysis in the Age of Big Data*. March 30, 2016. Accessed March 2, 2020. https://css.ethz.ch/en/services/digital-library/articles/article.html/195263/pdf.

Szekely, Pedro, Craig A. Knoblock, Jason Slepicka, Andrew Philbot, Amandeep Singh, Chengye Yin, Dipsy Kapoor, Prem Natarajan, Daniel Marcu, Kevin Knight, David Stallard, Subessware S. Karunamoorthy, Rajagopal Bojanapalli, Steven Minton, Brian Amanatullah, Todd Hughes, Mike Tamayo,

David Flynt, Rachel Artiss, Shih-Fu Chang, Tao Chen, Gerald Hiebel, and Lidia Ferreira. "Building and Using a Knowledge Graph to Combat Human Trafficking." In *The Semantic Web–ISWC 2015 14th International Semantic Web Conference Bethlehem, PA, USA, October 11–15, 2015 Proceedings, Part II*, edited by Marcelo Arenas, Oscar Corcho, Elena Simperl, Markus Strohmaier, Mathieu d' Aquin, Kavitha Srinivas, Paul Groth, Michel Dumontier, Jeff Heflin, Krishnaprasad Thirunarayan, and Steffen Staab, 205-21. Vol. 9367. Cham: Springer, 2015.

Tarinelli, Ryan. "Online sex ads rebound, months after shutdown of Backpage." *National Post* (Dallas), November 28, 2018. Accessed March 12, 2020. https://nationalpost.com/pmn/news-pmn/online-sex-ads-rebound-months-after-shutdown-of-backpage.

"The TellFinder Alliance." TellFinder. Accessed February 27, 2020. https://www.tellfinder.com/.

"Thorn, Digital Defenders of Children." McCain Institute. Accessed March 8, 2020. https://www.mccaininstitute.org/thorn-digital-defenders-children/.

UNODC. *Global Report on Trafficking in Persons 2018*. December 2018. Accessed February 11, 2020. https://www.unodc.org/documents/data-and-analysis/glotip/2018/GLOTiP_2018_BOOK_web_small.pdf.

van Rij, Jorn, and Ruth McAlister. "Using Criminal Routines and Techniques to Predict and Prevent the Sexual Exploitation of Eastern-European Women in Eastern Europe." In *The Palgrave International Handbook of Human Trafficking*, edited by John Winterdyk and Jackie Jones, 1689-708. Palgrave Macmillan, 2020. PDF.

Wagner, Janet. "Advanced analytics vs. artificial intelligence: How are they different?" ZyloTech. https://www.zylotech.com/blog/advanced-analytics-vs.-artificial-intelligence-how-are-they-different.

Winterdyk, John, and Jackie Jones, eds. *The Palgrave International Handbook of Human Trafficking*. Palgrave Macmillan, 2020. PDF.

Yu, Ronald, and Gabriele Spina Ali. "What's Inside the Black Box? AI Challenges for Lawyers and Researchers." *Legal Information Management* 19, no. 1 (March 2019): 2-13. Accessed March 2, 2020. https://doi.org/10.1017/S1472669619000021.

8. DATA REGIMES: AN ANALYTICAL GUIDE FOR UNDERSTANDING HOW GOVERNMENTS REGULATE DATA

Hunter Dorwart* and Olena Mykhalchenko**

Abstract

Artificial intelligence (AI) poses new problems for effective data governance on the local, national and international levels. Given the variety of historical, cultural and institutional structures between nation-states, governments around the world are beginning to diverge in how they regulate the big data ecosystem and assert power over the digital sphere. The absence of a consistent framework for analyzing these different stances towards data governance has prevented the harmonization of national interests and hampered the efforts of international organizations to create clear obligations towards emerging technological issues such as AI, digital surveillance, and cross-border data flows. Such harmonization, we argue, is necessary to keep new technologies in compliance with human rights frameworks, prevent the balkanization of the digital sphere and avoid the creation of police states. This paper coins a new term — data regime — to address this shortcoming and to better understand the emerging fragmentation of digital space by offering a new classification framework of how states assert power over digital technologies and the Internet. A data regime not only encompasses the policy tools countries use to pursue their objectives with respect to technology but also how those instruments relate to the larger normative values that guide global interaction between states. Finally, by aiding in the identification of national data interests, the framework codifies a novel way of viewing international data governance.

Keywords: Data, artificial intelligence, data governance, data ecosystem, data privacy.

Introduction

The trend of fragmentation that underscores the process of globalization is likely to intensify as countries continue to struggle with a plethora of governance problems. Such problems now concern many areas including fiscal and monetary instability, shallow economic forecasts, demographic and migration issues, supply chain vulnerabilities, ecological concerns, and public health crises. Indeed, as governments pursue their own national interests amid growing domestic political pressure, cooperative engagement in the

international domain is readily giving way to competitive retraction.

One such area of fragmentation concerns digital space and the global Internet. In response to emerging technological trends such as the "Fourth Industrial Revolution" and Artificial Intelligence (AI), governments are diverging in how they regulate data, assert control over digital processes, and manage interconnected communication systems.[1] Indeed, the situation is creating conflict scenarios over the future rules of digital space and the normative values that guide global interaction.[2] If unaddressed in a meaningful way, these scenarios threaten to cause more instability and heighten the risks associated with contemporary trends such as ecological catastrophe, global decoupling, and economic dislocation.

Technology is key to overcoming the major challenges facing humanity in the 21st century. Yet recent phenomena indicate that technology can harm cooperation and even accelerate global conflict: fears over autonomous weaponry, biotechnology, state surveillance, and disinformation now occupy a core part of the discourse around digital governance.[3] In order to minimize the harm of technological abuse by state and non-state actors, policymakers must understand how the global internet is fragmenting to anticipate and address the key sources of potential conflict. One dimension of this understanding involves answering how and why countries are diverging in their data governance approaches. The absence of a consistent framework for analyzing these different approaches has prevented the harmonization of national interests and hampered the efforts of international organizations to create clear obligations towards emerging technological issues such as AI, digital surveillance, and cross-border data flows.

In order to address the conceptual problem of digital fragmentation, this paper coins a new term — data regime. A data regime refers to the unique combination of governance capacity, economic policies, and behavioral practices that concretize a cognizable posture towards data governance. We determine the typology of data regimes through a set of guiding criteria that help classify the different variations in how governments regulate data. This paper does not make any normative judgments about data regimes but rather aims to describe how a government's interest of asserting more control in a fragmenting world crystalize into concrete data governance strategies. In addition, the framework outlined below does not specifically answer how states can harmonize their data regimes. Rather, it serves as an overarching conceptual lens through which states may not only rationalize norm creation in the digital sphere but also understand the possible divergence of that sphere in the near future.

Finally, this concept emerges in response to a perceived gap in the recent literature on topics concerning data privacy, artificial intelligence, and digital policy. While there are a plethora of frameworks guiding the conversation of the role of government in technology (and the role of *technology in government*), many of these frameworks deal with specific and narrow topics without addressing how these issues relate to much larger processes and phenomena currently facing the world today.[4] To be sure, these frameworks serve as useful tools for highlighting important problems regarding the intersection of technology and government and therefore form a crucial foundation of further research on the subject. The concept of data regimes builds upon these models in order to further shade some emerging trends in the next decade. The authors hope such a lens will be valuable to policymakers, the private sector, and other stakeholders to bridge diverging interests, prevent widespread harm, and engage international actors in a meaningful way.

---

1 * Hunter Dorwart is currently an independent researcher living in Washington D.C., who has engaged with numerous organizations in technology policy including Telecommunications Industry Association (TIA) and the International Bar Association (IBA).
** Olena Mykhalchenko is an experienced human rights attorney focusing on the nexus of law and technology. After her studies in the U.S. as a Fulbright and Edmund S. Muskie scholar, she became a fellow at the Portulans Institute, where she examines emerging issues with internet governance and technology.

The Fourth Industrial Revolution refers to the predicted widespread economic and social change brought on by new technologies that fuse the physical, the digital and the biological realms into a single process of production. Scholars associate the term with previous revolutions in human organization that saw profound transformations in the fundamental ways societies operate and harness technology. See. Tessaleno Devezas et al. *Industry 4.0.: Entrepreneurship and Structural Change in the New Digital Landscape*, (2016); Klaus Schwab. *The Fourth Industrial Revolution*, (2016).

2 "Confronting the Crisis of Global Governance." *Report of the Commission on Global Security*, (2015).

3 Anupam Chander & Uyên P. Lê. "Data Nationalism." *Emory Law Journal*, Vol. 64 (2015), pp. 677-739.

4 See e.g., "Government Artificial Intelligence Readiness Index 2019," *Oxford Insights* (2019), https://www.oxfordinsights.com/ai-readiness2019; Human Rights in the Age of Artificial Intelligence, *Access Now* (2018); Mark Latonero, "Governing Artificial Intelligence: Upholding Human Rights," *Data & Society* (2018); Eileen Donahoe & Megan MacDuffee Metzger, "Artificial Intelligence and Human Rights" *Journal of Democracy*, Vol. 30, No. 2 (2019); "How to Prevent Discriminatory Outcomes in Machine Learning" White Paper, *World Economic Forum* (2018); Samm Sacks & Justin Sherman, "Global Data Governance: Concepts, Obstacles, and Prospects" *New America* (2019).

Data Regime: Definition and Criteria

Generally speaking, data regimes form the basis through which countries regulate the Internet and use digital technology to achieve stated and unstated goals.[5] States may regulate digital space through a variety of tools including economic policies, privacy legislation, market and infrastructural control, and military power. Data regimes also reflect the normative values that states ascribe to their interests in the digital sphere. Countries may use these tools in a number of ways and for a number of reasons given the complexity of the overlapping interests involved.

Data regimes encompass more than data governance models or data governance approaches. The latter concepts often refer to specific policy choices that governments make with respect to cross-border data flows and information privacy.[6] As such, these terms tend to emphasize how governments regulate the private sector's collection and use of data or the transference of data beyond national borders.[7] Data regimes go further than this by encompassing the larger strategic goals that governments adopt to *actively define and create* the parameters of digital space and the underlying values or norms that govern the Internet on the international level. Indeed, data regimes are at the crux of how states assert their strategic interests over the Internet and refashion the scope of their sovereignty in the digital age. Finally, while data regimes incorporate data governance models, they also contain a geopolitical component that goes beyond influencing private sector behavior. As discussed below, data regimes further reveal how states are actively shaping the market and possibility of multilateral engagement and coalition-building in digital space.

This paper offers the following criteria to make sense of the divergence in data regimes and better situate how the different tools and reasons behind technology policy overlap in consistent and predictable ways. Each criterion represents a central axis of a state's data regime and includes specific policies and their rationales. The authors selected these criteria based on the wide policy discussions they receive, the centrality to economic growth they possess, and the importance for extending digital influence they offer. In addition, this paper assembled these criteria to make up for perceived gaps in other data governance typologies while reaffirming the major insights already produced in the emerging literature on data and policy.

Criterion 1 – Cross-border Data Flows and Localization

Digitalization now affects all aspects of economic activity and will increasingly become more central to everyday life.[8] The process has created a whole new economy centered around information: industries such as cloud computing, big data analytics, and e-commerce now make significant contributions to GDP.[9] It has also transformed the manufacturing process and been key to the reconfiguration of global supply chains following the structural transformation of the 1970s.[10]

The digital revolution has resulted in an unparalleled increase in the cross-border flow of data. Estimates indicate that over the past 10 years, global data flows have increased world GDP by roughly 10% with the amount of global bandwidth growing 148 times larger in the same period.[11] Cross-border data flows include data accompanying international trade, computer and information services, and data sharing within corporations.[12] This staggering volume of data transference has given rise to a series of attempts by governments to regulate what is now a critical component of economic activity and trade.

5 Stated goals are official rationales such as economic growth, job creation, or national security. Unstated goals are those that underly larger historical trends such as trade leverage, retaliation, and control.

6 Barbara L. Cohn, "Data Governance: A Quality Imperative in the Era of Big Data, Open Data, and Beyond." *Journal of Law and Policy for the Information Society* (2014).

7 See Kenneth A. Bamberger & Deirdre K. Mulligan, *Privacy on the Ground* (2015).

8 By digitalization we mean the process of automating and coordinating tasks through information communication technologies (ICT).

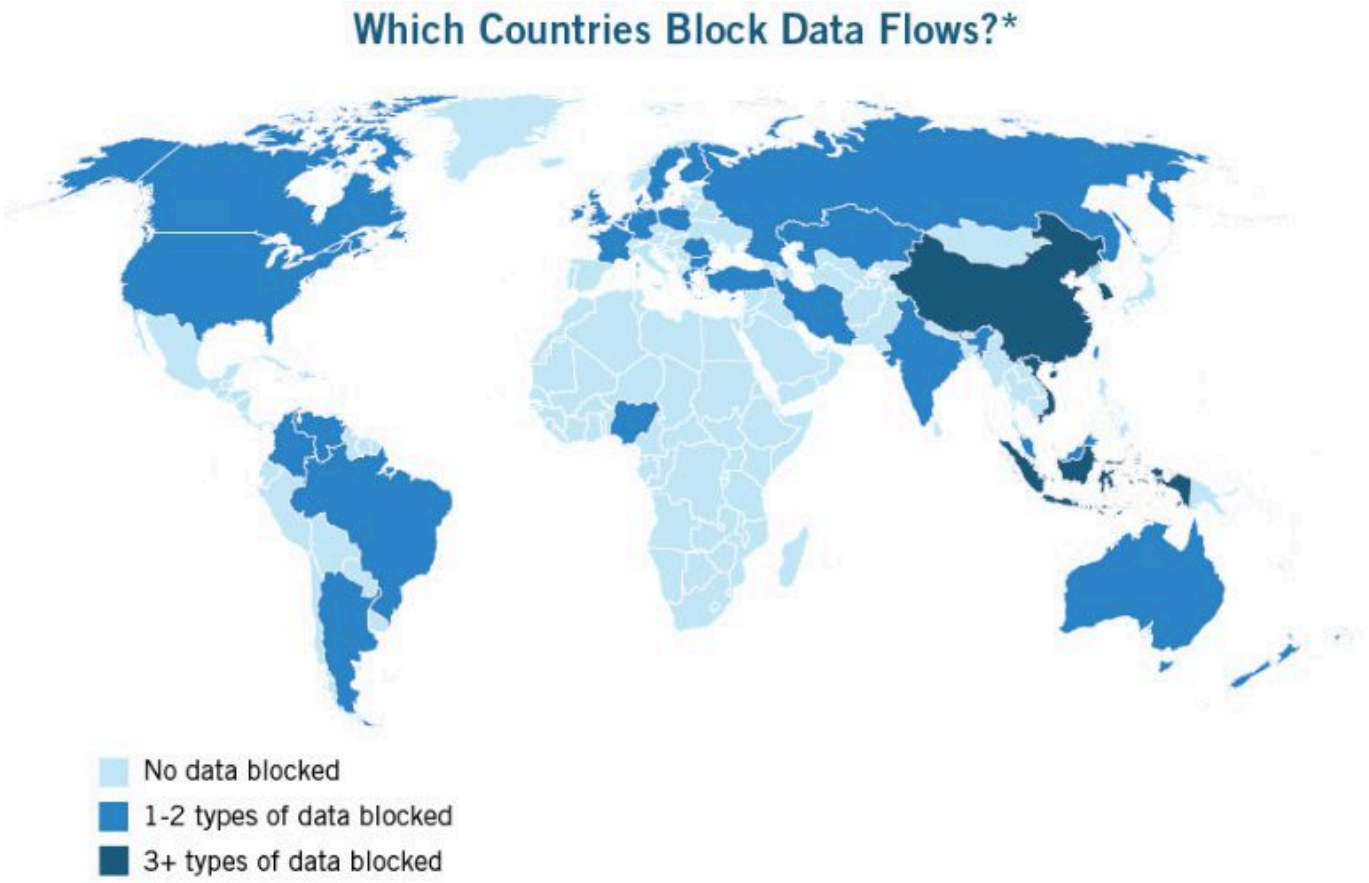
9 "Measuring the Digital Transformation: A Roadmap for the Future." *OECD* (2019), p. 20.

10 Scott Lash & John Urry. *The End of Organized Capitalism*. (1987), pp. 196-209. Jeffery Henderson. *The Globalization of High Technology Production*. (1989).

11 "Globalization in Transition: The Future of Trade and Value Chains." *McKinsey Global Institute*. (2019), p. 14.

12 "Digital Globalization: The New Era of Global Flows." *McKinsey Institute*. (2016).

To reflect this, the first criterion revolves around the regulation of data transference *between countries* and the rationales behind the policies. Restrictions on cross-border data flows take several forms. Generally speaking, these involve (i) categorical prohibitions on transferring data outside borders ("Hard Localization"); (ii) requirements for companies to maintain copies of data within local servers ("Soft Localization"); (iii) requirements for companies to obtain consent or meeting certain conditions before transfer ("Conditional Restrictions").[13]



Source: Information Technology and Innovation Foundation (ITIF)

Some accounts have highlighted a bifurcated approach to data regulation based on models from international trade.[14] On one end of the spectrum are countries that tend to support a liberalized trade regime; on the other are those that find these principles harmful for a handful of reasons, not the least including the asymmetrical effects of liberalized trade on domestic industries within disadvantaged countries.[15]The liberal/protectionist typology helps explain the underlying rationales behind data flow restrictions. We offer five main reasons:

- Protection of citizens' personal privacy
- State control over data for law enforcement purposes
- National security such as preventing foreign surveillance
- Promoting economic competitiveness through "digital import substitution"

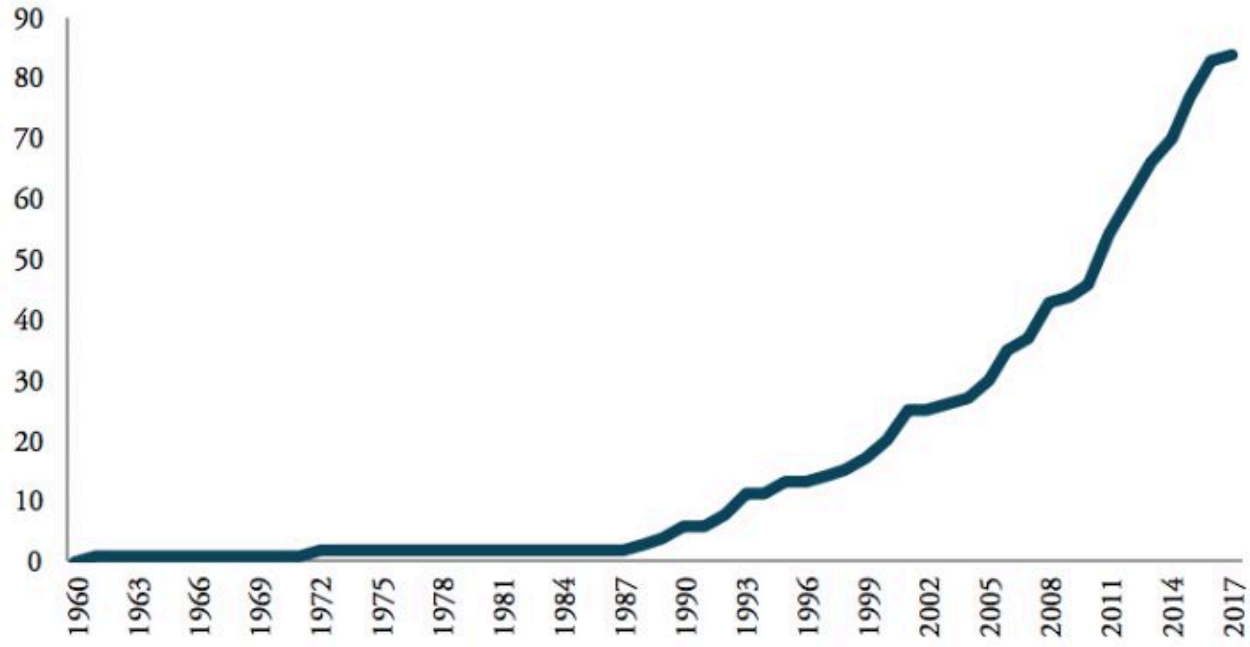
13 The line between these policies is often blurry as countries adopt combinations of localization measures. See. Meltzer, J. Lovelock, P. "Regulating for a Digital Economy: Understanding the Importance of Cross-Border Data Flows in Asia. *Brookings Institution*. (2018), p. 5.

14 See e.g., Nir Khsetri, "Success of Crowd-based Online Technology in Fundraising: An Institutional Perspective." *Journal of International Management*, Vol. 21, No. 2. (2015), pp. 100-116. Dani Rodrick. "Political Economy of Trade Policy." *Handbook of International Economics*. (1995), pp. 1457-1494; Shamel Azmeh & Christopher Foster. "The TPP and the Digital Trade Agenda." *International Development*, No. 16-175. (2016).

15 Liberalized here means the trade principles implemented through GATT and the WTO trade regime such as non-discrimination, coordinated predictability, and a general reduction of trade-barriers such as tariffs and quotas. Jagdish Bhagwati, Pravin Krishna, & Arvind Panagariya. "The World Trade System: Trends and Challenges." Presented at the Conference on *Trade and Flag: The Changing Balance of Power in the Multilateral Trade System*. (2014).

- Negotiating leverage with more advanced digital economies[16]

In practice, these rationales often overlap in myriad ways.[17] Each of these rationales occupy a particular point on the protectionist spectrum and reflect certain policy objectives that inform a data regime.[18]



Restrictions on cross-border data have steadily increased from 1960-2017.

Image: ECIPE

Criterion 2 – Personal Data Protection and Privacy Frameworks

Our second criterion involves how governments interact with the growing awareness of *social issues* related to the digital economy.[19] Concomitant with the digital revolution, the proliferation of personal data through databases, communication networks, and data-generating devices has resulted in a profound transformation in the relationships between citizens and governments.[20] The sheer amount of data generated today — estimated at 10.6 zettabytes — far surpasses levels measured even a decade ago, which has bolstered the emergence of a complicated data ecosystem and market projected at $49 billion.[21]

In response to these changes, policymakers have begun to implement legal requirements on the collection,

16 Meltzer, J. Lovelock, P. "Regulating for a Digital Economy: Understanding the Importance of Cross-Border Data Flows in Asia. *Brookings Institution*. (2018), p. 5.

17 "Data Localization: A Challenge to Global Commerce and the Free Flow of Information." *Albright Stonebridge Group* (2015), pp. 12-15.

18 This paper does not take a normative position on the value of these laws or the underlying political regimes of states that pass them.

19 These social issues are multifaced, multivariate, and occupy a core position within traditional theories of state-engagement and democratic participation. For an introductory summary see. Carole Pateman. *Participation and Democratic Theory*. (1970), pp. 40-53. For a more comprehensive treatment see. Joseph Schumpeter. *Capitalism, Socialism, and Democracy* (1943); Jurgen Habermas. *Legitimation Crisis* (1973); Harry Eckstein. *Division and Cohesion in Democracy* (1966); Robert A. Dahl. *Who Governs?* (1961); G. William Domhoff. *Who Rules America?* (1967); T.H. Marshall. "Citizenship and the Social Class" *Cambridge University Lecture Series* (1949).

20 Daniel Trottier & Christain Fuchs. "Theorizing Social Media, Politics and the State." *Social Media, Politics and the State: Protests, Revolutions, Riots, Crime and Policing in the Age of Facebook, Twitter and Youtube.* (2014).

21 Shanhong Liu, "Big Data-Statistics & Facts" *Statista* (2019). https://www.statista.com/topics/1464/big-data/; Thomas A Singlehurst et al, "ePrivacy and Data Protection" *CitiGroup*.

processing, and use of personal information.[22] This plethora of policy instruments —including laws, regulations, industry standards, governmental initiatives, and international agreements — forms an essential backbone to public sector data governance.[23]While a comprehensive analysis of these laws is beyond the scope of this paper, the proliferation of privacy frameworks in the public and private sectors has become a key nexus of policymaking and organizational management and has varied throughout national contexts.[24]

This variance has created an emerging fragmentation of data regimes and the legal architecture that underly them.[25] Rationales for imposing limitations on how private and public actors use personal data are crucial for understanding this fragmentation. On one side of the spectrum are countries that fashion privacy laws for economic purposes (i.e., consumer protection) or political choice reasons (i.e., public sector control over the digital economy).[26] These laws and regulations often prioritize cybersecurity above individual privacy and reflect goals of strengthening security interests from either cyberespionage or foreign interference in communications infrastructure.[27] On the other side are governance models that tend to protect privacy as a *fundamental human right*.[28]These models tend to emphasize the harmful social consequences of emerging technologies such as algorithmic bias and discrimination, pervasive surveillance, and job loss.[29] We acknowledge that in practice, countries often vacillate between these two conceptual rationales.[30]

22 Data protection requirements are embedded within cross-border data provisions such as the current E.U.-U.S. Privacy Shield. But while most countries have some type of data protection measure, the degree and scope of these restrictions vary. In addition, the stated rationales behind these provisions are also inconsistent, creating more fragmentation for countries' efforts to harmonize each other's digital trade agendas. For instance, one setback in the ongoing negotiation process for the Regional Comprehensive Economic Plan (RCEP) concerns disagreement over the proper data protection framework for cross-border data transfer. While Japan has been robustly committed to liberalizing data flows in coordination with its Act on Protection of Personal Information (APPI), some countries like Malaysia and Thailand have raised their own concerns with this type of legal imposition. In addition, similar problems exist for harmonizing the Asian Pacific Economic Cooperation (APEC) Privacy Framework with the ASEAN Framework on Digital Data Governance and establishing a consistent regime for privacy compliance. See generally, Thio Tse Gan, "Data and privacy protection in ASEAN – what does it mean for business in the region?" *Deloitte*. (2018); Akemi Suzuki & Tomohiro Sekiguchi, "Data Protection & Privacy" *Nagashima Ohno & Tsunematsu*. (2019). https://gettingthedealthrough.com/area/52/jurisdiction/36/data-protection-privacy-japan/; Kensaku Takase, "GDPR matchup: Japan's Act on the Protection of Personal Information." *IAPP* (2017). https://iapp.org/news/a/gdpr-matchup-japans-act-on-the-protection-of-personal-information/;

23 OECD "Data Governance in the Public Sector," *The Path to Becoming a Data-Driven Public Sector*. (2019); Daniel J. Solove & Paul M. Schwartz. *Privacy Law Fundamentals*. (2015).

24 Many of the key privacy frameworks have evolved from the Organization of Economic Cooperation and Development's (OECD) Fair Information Practices. The EU's 1995 Directive and subsequent General Protection Data Regulation (GDPR) both share substantial overlap with the principles as well as the 2018 California Consumer Privacy Act (CCPA). Robert Gellman. "Fair Information Practices: A Basic History." *SSRN* (2019). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2415020

25 "The Keys to Data Protection." *Privacy International*. (2018); Sean McDonald & Xiao Mina, "The War-Torn Web" *Foreign Policy*. (Dec. 19. 2018). https://foreignpolicy.com/2018/12/19/the-war-torn-web-internet-warring-states-cyber-espionage/#map

26 For example, the Clarifying Lawful Overseas Use of Data (CLOUD) Act in the United States was passed in 2018 as a direct result of the *Microsoft* decision which limited the power of law enforcement agencies to compel digital companies to produce communications of individuals. See. Stephen P. Mulligan. "Cross-Border Data Sharing Under the CLOUD Act." R45173 *Congressional Research Service*. (2018). In China, the CyberSecurity Law (CSL) aims at ensuring that digital and telecommunication companies operating within Chinese jurisdiction and processing personal data of Chinese citizens do not interfere with national economic policies. See. "The Data Protection Regime in China." *European Parliament's Directorate General for International Policies*. (2015).

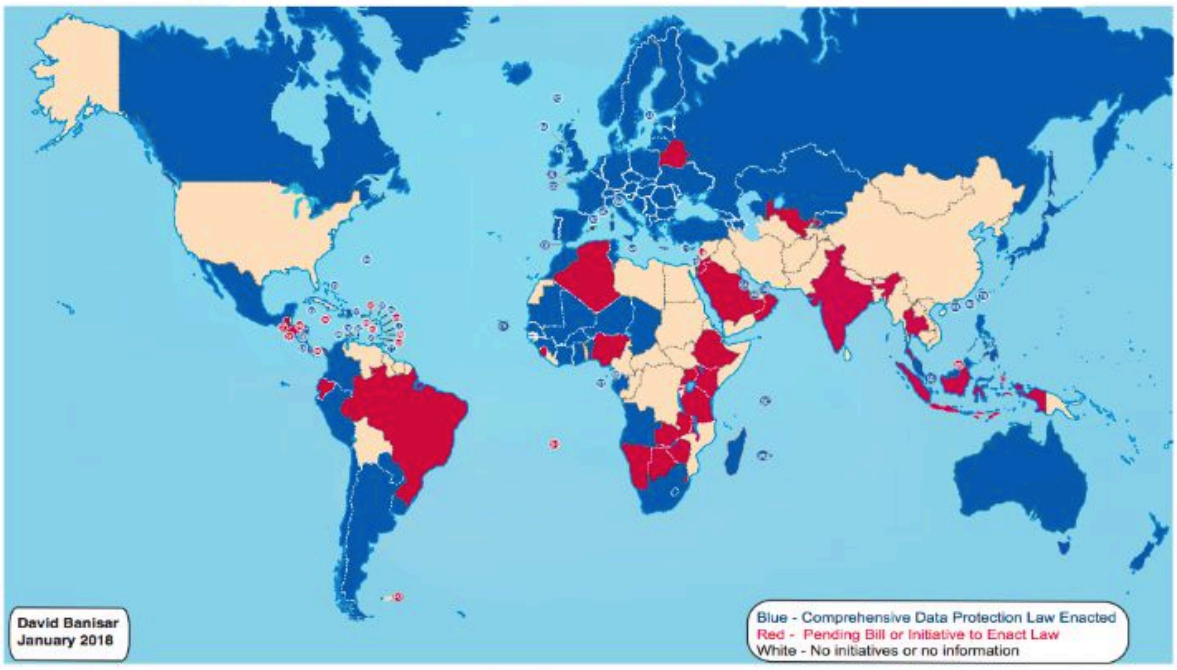
27 See e.g., The Cyber Security Basic Act in Japan (2014), China's Cybersecurity Law (2016), Cybersecurity Act in Singapore (2018), the Law on Cybersecurity in Vietnam (2018).

28 This tradition in part evolves from key international human rights agreements. For example, Article 12 of the Universal Declaration of Human Rights (UDHR) contains an oft-cited reference to privacy protection that has influenced other major treaties like the International Covenant on Civil and Political Rights (ICCPR) and the Convention on the Rights of the Child (CRC). Other examples of this include the 2016 UN General Assembly Resolution Right to Privacy in the Digital Age, GA Res. 71/199, the European Convention for the Protection of Individuals with Regard to Automatic Processing of Personal Data (Convention 108), and the Arab Charter on Human Rights. See. "International Privacy Standards." *The Electronic Frontier Foundation.* (2018). https://www.eff.org/issues/international-privacy-standards

29 "The Keys to Data Protection." *Privacy International*. (2018); "Human Rights in the Age of Artificial Intelligence." *Access Now* (2018).

30 For instance, Mexico has enshrined privacy protection into its Constitution and has passed a handful of federal privacy laws such as the General Law on the Protection of Personal Data by Public Entities, mainly to comply with EU requirements for commercial interaction. "State of Privacy in Mexico." *Privacy International* (Jan. 26, 2019). https://privacyinternational.org/state-privacy/1006/state-privacy-mexico

National Comprehensive Data Protection/Privacy Laws and Bills 2018



Source: Banisar, David, National Comprehensive Data Protection/Privacy Laws and Bills 2018 (January 25, 2018). Available at SSRN:https://ssrn.com/abstract=1951416 or http://dx.doi.org/10.2139/ssrn.1951416

Criterion 3 – Digital Industrial Policy

Digital industrial policy refers to a state's attempt to create favorable environments for technological development through allocating resources to help industries that revolve around data.[31] Industrial policy around AI and the Internet has become one mechanism through which governments are actively constructing their own visions of digital space.[32] These policies come in many forms including: subsidization (both direct and indirect), innovation parks, fiscal and monetary stimulus, long-term strategic planning, digital service taxes, acquisitional structuring, licensing arrangements, and foreign direct investment.[33]

The asymmetrical structure of global value chains (GVC) has produced variance in how countries pursue their national digital strategies.[34] Countries that initially dominated digitalization typically pursue strategies that will favor the expansion of their digital production into other markets while preventing competitors from challenging their dominance.[35] By contrast, countries "catching up" may pursue both defensive policies that *restrict* the penetration of global data firms into their domestic markets and aggressive policies that *spur* the development of their own technological capabilities and structural positions.[36]

31 Rather than viewing AI as a "catch-all" revolutionary term that implies every emerging issue with technology, this paper adopts a narrow definition confined to an extension of data processing technologies and the various national policies that underscore them. For a nice summary of the debate on industrial policy see. Christopher Foster & Shamel Azmeh. "Latercomer Economies and National Digital Policy: An Industrial Policy Perspective." *The Journal of Development Studies* (2019).

32 Some argue, for example, that the Russian "Sovereign Internet Law" operates to assert more control over the Internet by routing traffic through state-controlled infrastructure. See. Alena Epifanova. "Deciphering Russia's 'Sovereign Internet Law'" No. 2. *German Council on Foreign Relations* (2020). https://dgap.org/en/research/publications/deciphering-russias-sovereign-internet-law

33 Christopher Foster & Shamel Azmeh. "Latercomer Economies and National Digital Policy: An Industrial Policy Perspective." *The Journal of Development Studies* (2019).

34 Sean McDonald & Xiao Mina, "The War-Torn Web" *Foreign Policy*. (2018).

35 For a nice overview of this in the industrial context see. Ha-Joon Chang, *Kicking Away the Ladder: Development Strategy in Historical Perspective* (2003); Sanjaya Lall, "Technological Capabilities and Industrialization." *World Development* (1992); Immanual Wallerstein *The Modern World System*. (1974); Giovanni Arrighi. *The Long Twentieth Century* (1994); Nick Srnicek, *Platform Capitalism* (2016).

36 For instance, India's National Program on AI proposes to utilize its large talent pool to create cross-cutting projects while focusing more domestic resources on developing key data-intensive industries at home. In addition, India's proposed data localization law would have arguably raised transaction costs for foreign firms, thereby altering the preferability of certain services in the country. See. "National Strategy for Artificial Intelligence #AIForAll." *Niti Aayog*. (2018); Arindrajit Basu et al. "The Localization Gambit" *The Centre for Internet and Society, India*. (2019); "Unlocking the Potential of India's Data Economy. *Omidyar Network India*. (2019).

National policies can take a conciliatory, a competitive, or an isolationist posture vis-à-vis the perceived threat from other digital firms. With conciliatory engagement, countries negotiate development strategies in conformity with preexisting norms and customs in order to cooperate with digital multinationals for perceived mutual benefit.[37] In contrast is a competitive option whereby countries develop their own digital capacities and then aggressively compete with other multinationals in exporting goods and services to acquire market share abroad.[38] Finally, countries may choose an isolationist route — that is, allocating resources to develop self-sufficiency in knowledge-intensive industries or extending juridical sovereignty over digital territory.[39] The framework acknowledges that in practice, countries adopt combinations of these policy postures.[40]

Criterion 4 – Artificial Intelligence Research and Development (R&D)

This criterion takes into consideration the private sector financing of AI and the academic research and talent pool of the global AI landscape. The sheer volume of literature outpouring on AI is increasing with considerable pace.[41] Indeed, one figure puts the global volume of peer-reviewed AI papers at a 300% increase from 1998-2018.[42] Coupled with this has been the dramatic growth of AI venture capital financing and public-sector participation in spurring the development of new technologies.[43] In 2019, global private AI investment totaled over $70 billion and of that ($37 billion) went to startups while the other ($34 billion) went to mergers and acquisitions ("M&A").[44]

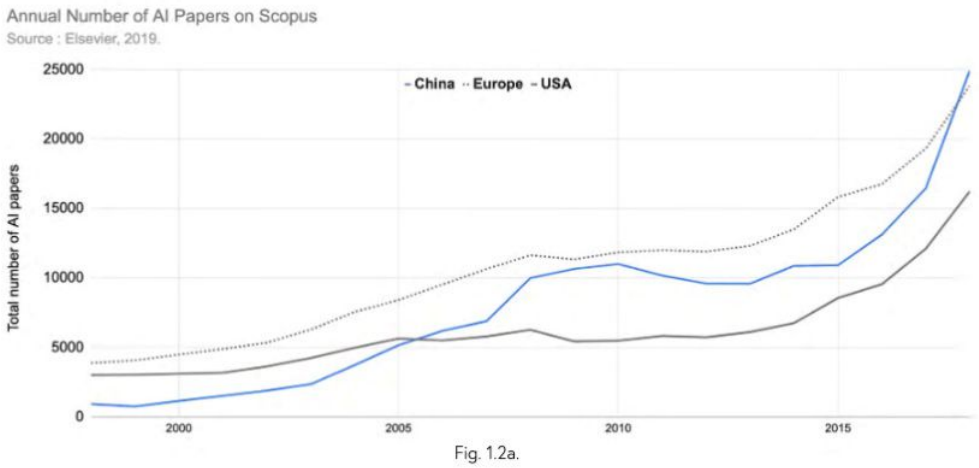


Fig. 1.2a.

As such, the degree of resources committed within and by countries to the development of AI indicates their power projection *capabilities and intentions*. While there is disagreement over which countries have invested

37 The Zimbabwean government signed a data-sharing agreement with the Guangzhou-based startup CloudWalk Technology in 2018 to develop smart city infrastructure, in what has been perceived as a mutually beneficial agreement. Lynsey Chutel. "China Is Exporting Facial Recognition Software to Africa, Expanding its Vast Database." *Quartz Africa*. (May 25, 2018). https://qz.com/africa/1287675/china-is-exporting-facial-recognition-to-africa-ensuring-ai-dominance-through-diversity/

38 China occupies this position in relation to the United States and its global digital firms. While this paper declines to get into the debate, it acknowledges the role of China's industrial policy in spurring its own high-tech capabilities and now forward-looking development plans (e.g., Belt and Road Initiative (一带一路), Made in China 2025). See. Jeffery Ding. "Deciphering China's AI Dream." *Future of Humanity Institute*. (2018); Martina F. Ferracane & Hosuk Lee-Makiyama. "China's Technology Protectionism and its Non-Negotiable Rationales. *European Centre for International Political Economy*. (2016).

39 In reality, the former is highly unlikely given the fixed capital cost threshold for high-tech development but could materially manifest in certain contexts. For example, Cuba's recent technological trajectory illustrates this possibility. Anne Nelson. "Cuba's Parallel Worlds: Digital Media Crosses the Divide." *Center for International Media Assistance*. (2016), p. 13.

40 For example, Kenya has at various points adopted all three depending on the context and negotiating partner involved. Karishma Banga & Dirk Willem te Velde. "How to Grow Manufacturing and Create Jobs in a Digital Economy: 10 Policy Priorities for Kenya." *Supporting Economic Transformation* (2018), pp. 7-13.

41 See. Grace Kiser & Yoan Mantha. "Global AI Talent Report 2019." *Jfgagne*. (2019). https://jfgagne.ai/talent-2019/

42 "Artificial Intelligence Index Report 2019." *Stanford Institute for Human-Centered Artificial Intelligence* (2019), p. 5.

43 For a global study in the private sector initiatives see. Michael Chui et al. "Notes from the AI Frontier" *McKinsey Global Institute*. (2018). For a breakdown of public sector initiatives see. Tim Dutton, "An Overview of National AI Strategies." *Medium* (2018). "Mapping Regulatory Proposals for Artificial Intelligence in Europe." *Access Now* (2018).

44 "Artificial Intelligence Index Report 2019." *Stanford Institute for Human-Centered Artificial Intelligence* (2019), pp. 4-5.

more and are "leading" the development of AI, we decline to get into this debate.[45] Instead, this framework focuses generally on the degree to which states have invested and attracted human and physical capital into propping up their respective AI reputations. In comparison with digital industrial policy, this criterion reflects pure economic and investment power rather than any specific set of policy choices. For instance, a country may have little AI capability but nonetheless proposes a robust national strategy for reputational reasons —this criterion serves to account for this aspect and further clarify the *actual* pace of talent acquisition and foreign direct investment ("FDI").

Moreover, educational capacity to work in the global innovation space also implicates the ability of governments to analyze and catalogue public statistical data such as census blocks or health and consumer behavior. States lagging in this *statistical capacity* often have a hard time incorporating data analytics into value chains and face limitations in developing private sector technology solutions. Without public data, businesses lack access to the key resource they use to develop, test, and implement their services, which hinders a government's larger data strategy and reinforces a country's lack of know-how or high-tech investment.[46] These data gaps further harm educational initiatives by creating feedback loops: diminished talent pools reinforce administrative gaps and foster incentives for brain-drains because of diminished opportunities at home.[47] While the scope of this framework primarily focuses on states and their institutional actors, it nonetheless does not ignore the importance of private sector behavior for data governance.

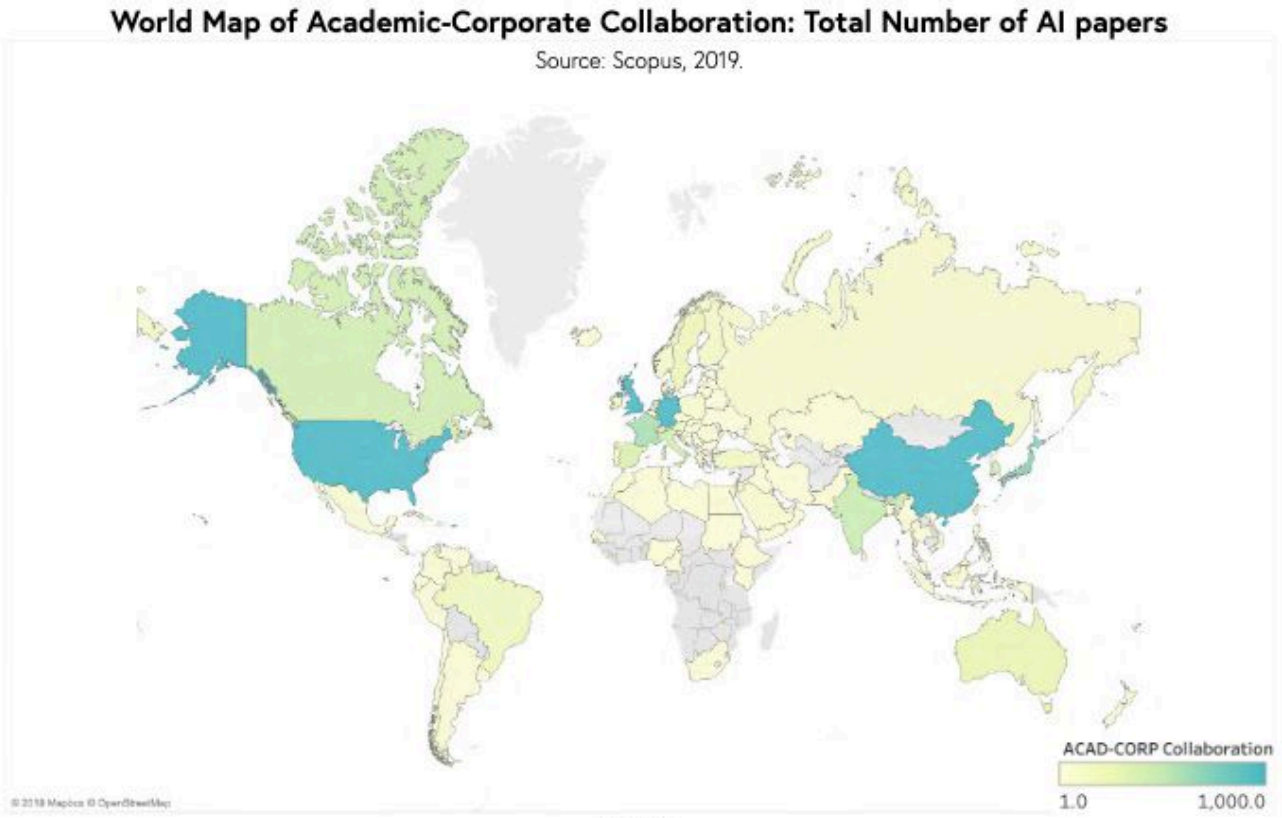


Fig. 1.5a.

45 Rarely do any studies deny that China and the U.S. account for the largest share of AI-related R&D in the world. The debate largely centers around which of these two is truly leading the next revolution of digital technologies. See generally. William A. Carter & William D. Crumpler. "Smart Money on Chinese Advances in AI." *Center for Strategic and International Studies*. (2019), p. 13. Ryan Hass & Zach Balin. "US-China Relations in the Age of Artificial Intelligence." *The Brookings Institution*. (2019). https://www.brookings.edu/research/us-china-relations-in-the-age-of-artificial-intelligence/

46 Steve MacFeely & Nour Barnat. "Statistical Capacity Building for Sustainable Development: Developing the fundamental pillars necessary for modern national statistical systems." *United Nations Economic Commission for Europe* (2016), pp. 2-4.

47 "Global Talent 2021: How the new geography of talent will transform human resource strategies." *Oxford Economics*. (2020); "Global Talent Risk – Seven Responses." *World Economic Forum* (2011), p. 9.

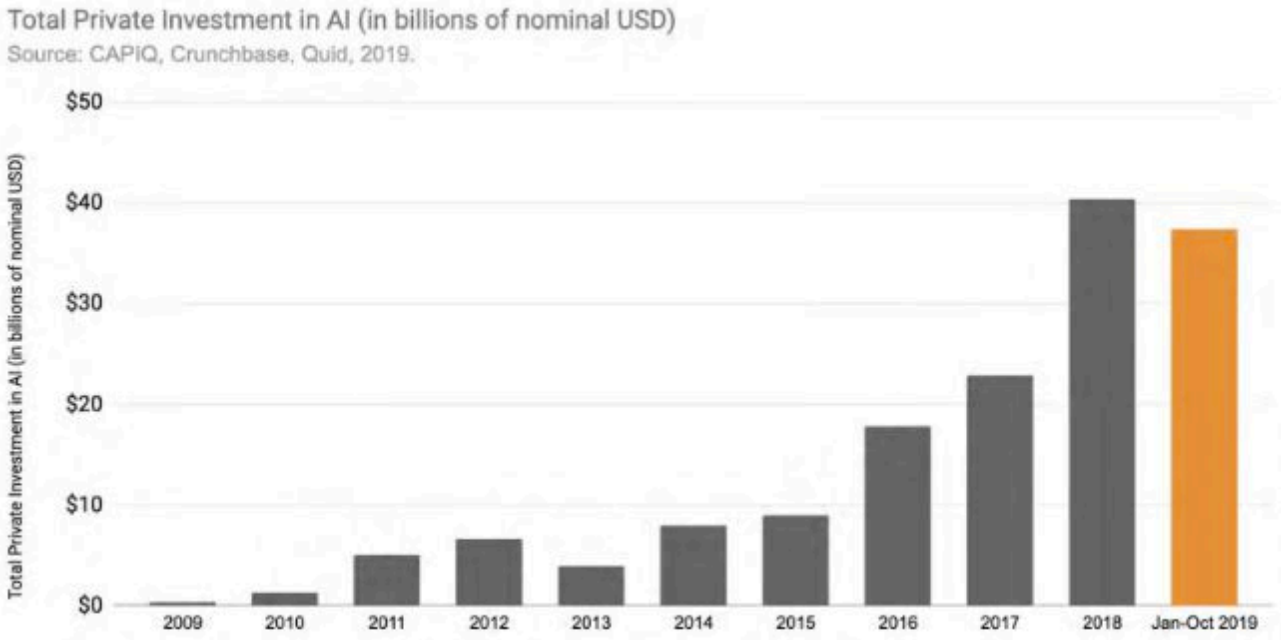


Fig. 4.2.1a.

Criterion 5 – Infrastructure and Development

Digital infrastructure refers to the physical and networked systems that people use to transport information, services, and digital bits. This definition includes both telecommunications infrastructure and the physical infrastructure that contains both informational and digital components ("Hybrid Infrastructure").[48] A thorough digital governance typology must contain an analysis of the levels of digital infrastructure development. On one hand, the capacity for networked communications helps determine a country's historical sequence of engaging with the digital revolution *in the past*.[49] Indeed, activity in the online space is impossible without some form of network capacity. On the other hand, infrastructure helps situate a country's current position with respect to *future* digital development. Indeed a country's ability to project power to consolidate its interests in digital space and influence norm-creation hinges upon its ability to engage in the game in the first place.[50] Moreover, countries can proclaim their infrastructural commitments in tandem with other policymaking goals such as closing the digital divide, developing the smart cities of the future, and preparing the workforce of tomorrow.[51]

Like other data governance indicators, there is a wide discrepancy in the compositional index of infrastructural development between countries that have claimed a dominant position in the global digital economy and countries that have arrived late to the process.[52] But unlike other sources of digital power, telecommunications infrastructure is not entirely dispositive with respect to geopolitical control over digital space.[53]A country or region may have strong broadband deployment but lack other capacities (i.e., talent acquisition, funding) to

48 Aykut Atali, Chandra Gnanasambandam, & Bhargs Srivathsan. "Transforming Infrastructure Operations for a Hybrid-Cloud World. *McKinsey & Company* (2019). https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/transforming-infrastructure-operations-for-a-hybrid-cloud-world#

49 Tom Forester. *The Information Technology Revolution*, (1980); Melvin Kranzberg. "The Information Age: Evolution or Revolution?" *Information Technologies and Social Transformation*, (1985).

50 Robert D. Atkinson. "A Policymaker's Guide to Digital Infrastructure." *Information Technology & Innovation Foundation*. (2016), pp. 3-5.

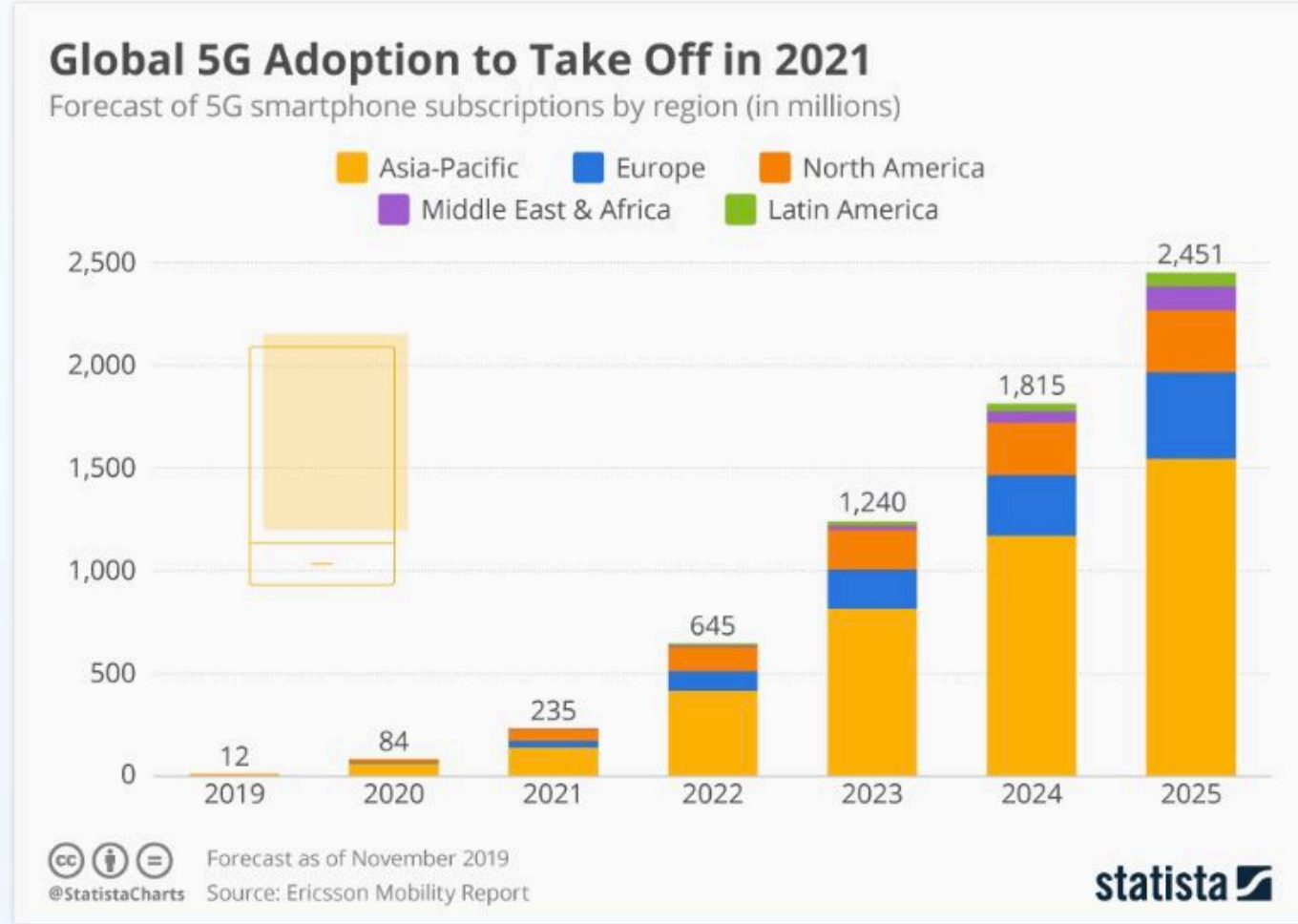
51 Durban, South Africa, is one example where municipal authorities, with the help of the State, can use smart technologies to foster innovative and sustainable growth and attract attention from other countries in the process. A key component of Durban's transformation, which saw the city selected by the Rockefeller Foundation as a pioneer for its Resilient Cities program, was the ability to link infrastructural management with data analytics. "Global Lessons from Durban's Climate Change Challenges." *The New Humanitarian*. (May 24, 2011). http://www.thenewhumanitarian.org/news/2011/05/24/global-lessons-durban-s-climate-change-challenges

52 "Global Telecommunications Study 2019" *Ernst & Young* (2019), p. 4.

53 The U.S.-China 5G competition illustrates that while network capacity is a key indicator of power, it by itself cannot determine leadership in digital space. While many have highlighted the connection between the proliferation of Huawei and the ability of the Chinese to influence standard setting, no one could doubt that major U.S. tech firms have led the digital revolution and will continue to occupy a central position in its ongoing process.

independently *lead* digital innovation.[54] Rather, network infrastructure is a crucial *foundation* through which all countries channel their digital governance interests.

This criterion acknowledges the importance of 5G and the Internet of Things (IoT) for understanding the recent fragmentation of digital governance and the fight over digital territory.[55] Previous advancements in wireless communications saw multiple state actors assert leadership in deploying hardware equipment and creating the specifications and technical standards that support the interoperability of digital technologies.[56] The difference with 5G is precisely the current fragmentation process that has turned data innovation into such a contentious sphere of geopolitical and economic competition. Coupled with the increasing importance of IoT for daily economic transactions, network infrastructure helps measure a country's future digital prospects and position in the larger competitive arena.[57]



Criterion 6 – Administrative Capacity and Regulation

Administrative capacity within a country reflects not only the degree to which states can enforce data governance mechanisms but also the structural relationship between the private and public sectors. While the precise definition of this term is hotly contested, the authors approach this with two major conceptual fulcra in mind. First, administrative capacity reflects a *functional* dimension in the public administration sense:

54 Europe is a prime example of this situation. See. Nick Wallace & Daniel Castro. "The Impact of the EU's New Data Protection Regulation on AI." *Center for Data Innovation.* (2018), pp. 2-4; John H. Chestnut. "U.S. vs. European Broadband Deployment: What Do the Data Say?" *Center for Technology, Innovation and Competition*. (2014).

55 5G refers to the next generation of wireless communications that will see unprecedented increases in the amount of dataflow and network connectivity. IoT refers to the proliferation of smart devices that bridge the digital and physical worlds and the process that will further enable a complex data-driven economy.

56 Leadership in wireless communications innovation is a hotly contested area because "first mover" status usually gives companies an advantage in acquiring early market share through the "network effect." Catherine Tucker. "Network Effects and Market Power: What Have We Learned in the Last Decade?" *Antitrust* (2018). http://sites.bu.edu/tpri/files/2018/07/tucker-network-effects-antitrust2018.pdf

57 Stockholm, Sweden is one notable example of how digital infrastructure is influencing sustainable development strategies. Hydropower and strong electricity grids form the backbone of the country's leadership with respect to sustainable city design. In comparison, Toronto, Canada, also boasts of smart city technological deployment.

it is the ability of complex bureaucracies to carry out actions through legal instruments from the local to the national levels.[58] In addition, this paper also acknowledges that the term contains a component of *power* — that is, the ability of the state to *forcefully* implement political decisions, regulate social relationships, allocate resources in determined ways, and mobilize collaboration in pursuit of goals.[59]

Because this paper uses the term in this descriptive way, it declines to address the role of corruption or the type of government for analyzing data regimes. Studies that focus exclusively on governmental structure tend to assume that the correct standards of data governance are those correlated with certain governmental regimes or existing international norms.[60] Nonetheless, the authors acknowledge that in many circumstances, these indicators do explain complex behavior and serve as useful tools for pinpointing limitations in policy implementation, effective governance, and transparency.

Following this, administrative capacity helps measure the enforceability of laws and therefore operates as another criterion for analyzing data governance. The existence and proliferation of laws, regulations, and policy objectives with respect to digital innovation and cyberspace would amount to little if the state could not actually implement them and influence concrete behavior. Indeed, many countries have promulgated data protection laws but find it difficult to enforce them fail to enforce them due to lack of resources and procedural delays.[61] This variability in enforcement alters the behavior of actors who either (i) must comply with the laws of a given country; (ii) negotiate with a given country internationally; or (iii) pass, implement or review the legislation themselves.

With respect to the first instance, corporations and other business entities must adjust their risk assessment models when engaging in market opportunities to balance the costs of compliance and uncertainty. In addition, reputations for lack of enforceability may harm the ability of a country to project its digital power abroad or influence the creation of norms. Particularly, reputation linked to administrative capacity also involves the ability of states to use technology to respond to crises or severe problems.[62] Such reputation contains both prospective and respective components: states may use technological capacity to address an anticipated problem or respond to an existing one. These considerations influence the posture a state will adopt in order to pursue its own interests.

Moreover, administrative capacity also illustrates the unique balance of interests between public and private stakeholders within a given country. This balance of interests accounts in part for the specific formation of policy as well as the normative approach states take towards the digital sphere.[63] It therefore covers variance in the state's role in the economy and the approach taken in altering private behavior. Some states have taken a heavy-hand approach in asserting influence over the digital economy.[64]

---

58 Peter J. May. "Policy Design and Implementation." *Handbook of Public Administration*, (2003); Beryl. A. Radin. "The Instruments of Intergovernmental Management." *Handbook of Public* Administration, (2003). Peter Morgan. "Capacity Development: An Introduction." *Emerging Issues in Capacity Development: Proceedings of a Workshop*. (1993).

59 Michael Mann breaks this concept into two components: "despotic" and "infrastructural" power. Michael Mann. *The Sources of Social Power,* (1986). For a treatment of Mann's work, see. Joel S. Migdal. *State in Society*, (2012).

60 Susan Aaronson, "The Turn to Trade Agreements to Regulate the Internet," in Jean-Baptise Velut et al., in *Understanding Mega-Free Trade Agreements: The Political and Economic Governance of New Cross-Regionalism*, (2017); Steven Feldstein. "The Global Expansion of AI Surveillance." *Carnegie Endowment for International Peace*, (2019).

61 See Rule of Law Index. *World Justice Project*, (2019).

62 Importantly, the public health measures taken by governments in response to the spread of Covid-19 showcases this capacity. Countries in East Asia like China and South Korea utilized complex data analytics in order to predict the spread of the virus and leveraged their e-commerce infrastructure to deliver goods directly to infected individuals. Justin Fendos. "Lessons from South Korea's COVID-19 Outbreak: The Good, Bad, and Ugly." *The Diplomat*, (Mar. 10, 2020). https://thediplomat.com/2020/03/lessons-from-south-koreas-covid-19-outbreak-the-good-bad-and-ugly/

63 Saudi Arabia's Vision 2030 contains a digital initiative management program to recalibrate the country's economic model through innovative technologies and young talent. The creation of the strategy emerged out of the recognition by Saudi leaders for the need to diversify economic growth away from oil production and refashion the country's overall digital posture in the 21st century. "Unlocking the Digital Economy Potential of the Kingdom of Saudi Arabia." *Ernst & Young,* (2019).

64 In addition to the United States and China, some other notable illustrations are South Korea, Indonesia and Vietnam. See. Lee Junkyu. "Korea's Trade Structure and Its Policy Challenges." *The Future of Korean Trade Policy*, (2012); Teresa Umali. "Developing Indonesia's Digital Ecosystem." *OpenGov Asia* (Aug. 29, 2019). https://www.opengovasia.com/developing-indonesias-digital-ecosystem/ Prema-chandra Athukorala. "Trade Policy Reforms and the Structure of Protection in Vietnam." *The World Economy*, Vol. 29. No. 2. (2006), pp.161-187.

Others have, at various points, been less active in this respect and instead have focused on accommodating the private sector to lead the digitalization process.[65]

Criterion 7 – Leadership and Power

Lastly, any meaningful classification of digital governance must have some measurement of leadership and power. Like administrative capacity, power is a complicated term that often draws more disagreement than consensus.[66] We refer to power as the ability of the state to pressure other states into acquiescing their interests. However, this ability is not always straightforward. As discussed below, there are a range of tools available for a state to exercise its power including brute force (both military and economic), agenda setting, and global norm creation.[67] While each of these tools may allow a state to exert pressure on others, the manner in which they operate reflects notable theoretical differences.

Brute force, perhaps the simplest but most aggressive tactic, only goes so far because it operates on a case-by-case basis and is relatively easy to lose.[68] Retaining brute force capacity requires ensuring that all actors cannot effectively challenge the dominant power. Agenda-setting may emerge from brute force (i.e., negative conditioning) but can also be acquired independently.[69] Because agenda-setting allows an actor to *remove* the possibility of another actor fulfilling its interest, this form of power is more sophisticated than brute force and goes further in its operation.[70] Finally, while agenda-setting may alter the constellation of possible options, it does not alter the original desire of the actor to pursue its interest. Global norm creation, by contrast, exerts even more power than agenda-setting because it not only determines the possible options on the table but also alters how a state defines its own interest.[71] Striking at the core of how other actors define their strategic interests constitutes an incredibly powerful tool and represents one of the highest forms of power a state can aspire to.

These three axes of power are interwoven into how states assert their influence over the internet, digital space, and the future of technology through data regimes. They also create hierarchies of leadership within the global constellation of states. Leadership is an important component of data regimes because it in part represents how states view each other. For this criterion, the paper proposes the following dimensions.

First, leadership contains a reputational component that we define as the ability of a state to use its *perceived position of authority* to pursue its policy objectives and assert its interests in digital space.[72] The source of this reputation stems from a complex set of factors including history, economic capability, innovation readiness, norm-creation, multilateral and plurilateral engagement, and vision-setting. The degree of leadership varies depending on the forum of engagement and the leadership position of the states involved in the interaction. For example, a country that lacks influence in relation to a digital leader may find itself in a better position among a smaller cohort.[73]

---

65 Israel is a somewhat neglected example of this phenomenon. Its economy is ripe with innovative capacity and has since the early 2000s become a hotbed for digital activity, hosting a high concentration of startups, VC funds, and corporate R&D centers, all without any protectionist laws. However, most of these startups face challenges competing with foreign companies and are usually acquired by the larger players. Dan Senor & Saul Singer. *Start-up Nation: The Story of Israel's Economic Miracle* (2009).

66 Stephen McNamee & Michael Glasser. "The Power Concept in Sociology: A Theoretical Assessment." *Humboldt Journal of Social Relations*, Vol. 15, No. 1. (1987), pp. 79-104.

67 For a discussion of the transformation of power projection tools over the past half century see Martin van Creveld. *The Transformation of War* (1991). For a more philosophical account of this schema of power see Steven Lukes. *Power: A Radical View* (1974).

68 Ibid, pp. 110-112.

69 Margaret Groarke, "Power, agency and structure," *New Political Science*, Vol. 14, No. 1 (1993).

70 Adam Przeworski, *Capitalism and Social Democracy* (1985).

71 Steven Lukes. *Power: A Radical View* (1974).

72 The authors acknowledge the debate in international relations theory about the limitations of defining what a state really "wants." See. Alexander Wendt. *Social Theory of International Politics* (1999). Nonetheless, we choose to accept that states do act in their own interest and pursue goals based off this interest.

73 Finland, Latvia, and Estonia have emerged as trend-setters in combating foreign interference of elections and have gained diplomatic and negotiating leverage on this topic over other states. Sebastian Bay & Guna Šnore. "Protecting Elections: A Strategic Communications Approach." *NATO Strategic Communications Centre of Excellence,* (2019), p. 5.

Second, leadership also contains a dimension that extends beyond soft power indicators.[74] It also involves military hard power and the ability of the state to resort to the threat or use of force in exceptional circumstances.[75] The authors acknowledge that while international legal mechanisms play an influential role in shaping a state's cyber-related decisions, many of the current practices in cyberwarfare have emerged through the "grey zones" in conventional and customary arrangements.[76] This twilight area of international law in part accounts for how a state approaches data governance as a matter of leadership and power because it represents unchartered areas where precedent-setting carries a tremendous advantage for those who can do so successfully.

From a geopolitical perspective, digital governance involves two dimensions. On the one hand, the ability of the state to exert military power through cyberspace *externally* on other actors indicates a degree of power. On the other, the ability of the state to use digital policing tactics *domestically* on its own people indicates not only the *intention* of the state actor but also the *pervasiveness* of its digital control. While associating certain political philosophies with either external or domestic methods of digital control may yield interesting correlations, it does not by itself present a thorough understanding of the fragmentation of the global Internet because almost all countries have now instituted a certain level of surveillance infrastructure.[77] For this reason, our typology does not focus on surveillance capacity as an *independent* criterion for evaluating digital governmental policy and instead opts to include it within a general overview of military power.

External power projections involve leveraging cyberspace capabilities to exploit vulnerabilities in a target country's key communication, energy, financial, economic, and political infrastructure in order to pressure leaders through social disruption.[78] Cyber-capabilities in the *offensive* sense include both Computer Network Exploitation ("CNE"), Computer Network Attack ("CNA") and Computer Network Disruption ("CND").[79] Many countries are already beginning to invest in massive defense resources to develop strategic readiness for the purpose of deterring foreign attacks.[80]Moreover, current and planned defense agreements between states illustrates the formation of a loose cyber alliance system and may indicate future trends within international relations.[81]

74 The term soft power, which means the ability of the state to attract and influence other states through diplomacy, is associated with Joseph Nye. *Soft Power: The Means to Success in World Politics*, (2004).

75 Military hard power refers to the either the aggressive or defensive use of economic or military means to influence the behavior of another state. Joseph Nye. *Soft Power: The Means to Success in World Politics*, (2004). For a broader contextualization of this concept in relation to state theory see Max Weber, "Politics as a Vocation." *Max Weber: Essays in Sociology*, translated by H. H. Gerth and C. Wright Mills (1946); Giorgio Agamben, *Homo Sacer* (1995).

76 For example, while the UN Charter may govern the right of states to resort to the threat of use of force, many of the norms that exist with respect to cyberwarfare have emerged outside the black letter of the law. Robert Reiner. *The Politics of the Police*, (2010); Kubo Mačák. "Is the International Law of Cyber Security in Crisis?" *2016 8th International Conference on Cyber Conflict*, (2016), pp. 132-134.

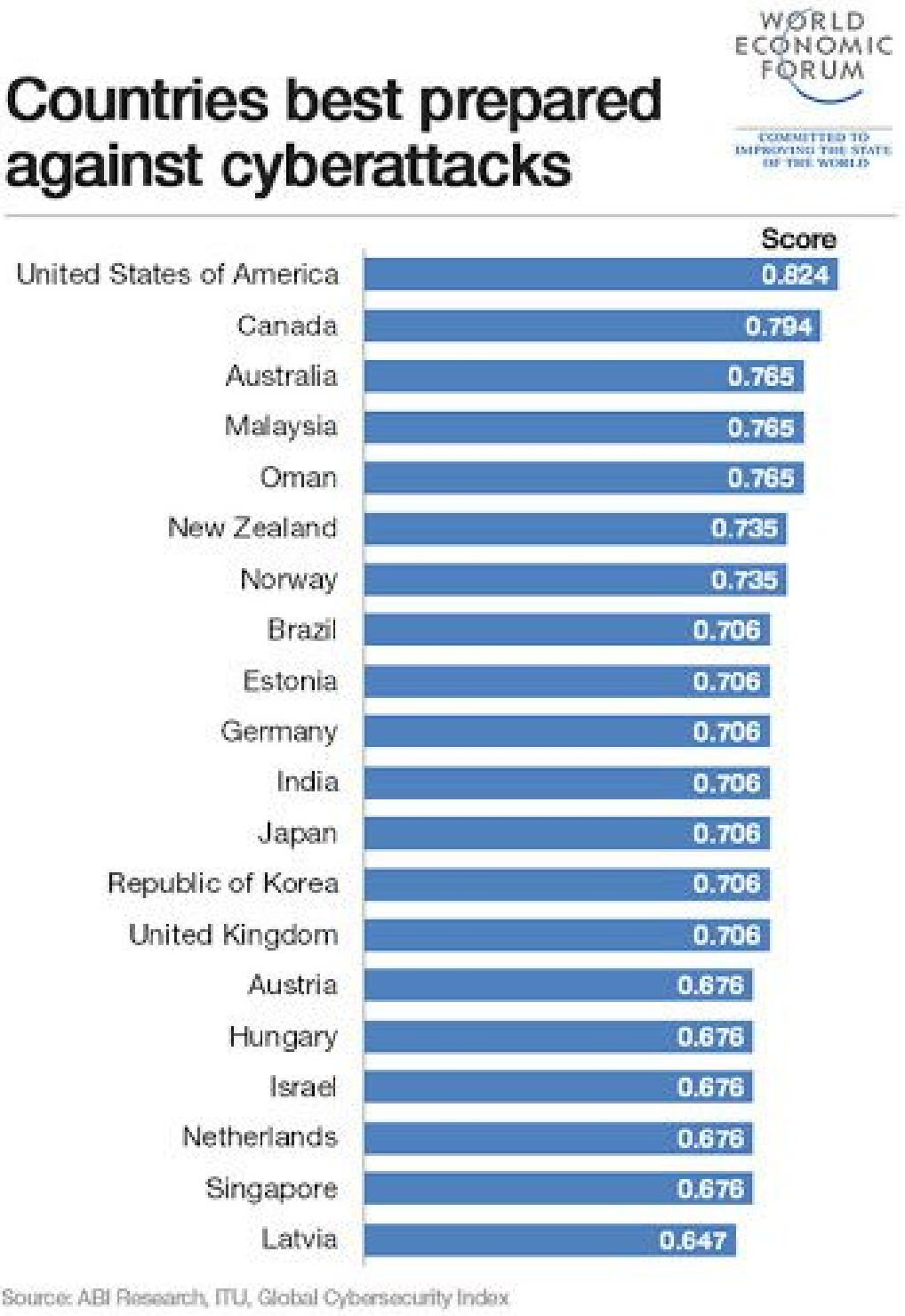
77 Sean McDonald & Xiao Mina, "The War-Torn Web" *Foreign Policy*. (2018).

78 Some notorious cyberattacks include: the 2014 Yahoo! hack which compromised 200 million identities, the 2017 Equifax credit card number data breach affecting 143 million American, Canadian, and British customers, the 2017 Wannacry Virus, and the Stuxnet cyber-worm that targeted an Iranian nuclear facility in Natanz in 2010. See. "Commodification of Cyber Capabilities: A Grand Cyber Arms Bazaar." *2019 Public-Private Analytic Exchange Program.* (2019), pp. 14-18.

79 For a comprehensive discussion of these terms see. Ross M. Rustici. "Cyberweapons: Leveling the International Playing Field." *Parameters*. (2011).

80 For instance, the Australian 2016 Cyber Security Strategy created an AU$230 million fund to improve infrastructure resiliency to cyber-attacks. The strategy initiates the creation of the Australian Cyber Security Centre (ACSC) which coordinates intelligence analysis with other Australian institutions, as well bolsters the resources dedicated to financial and energy stability and election security. Stuart Fowler et al. "Developing Cyber-Security Policies that Penetrate Australian Defense Acquisitions." *Australian Defense Force Journal*. (2017)

81 Benjamin M. Jensen, Brandon Valeriano, & Ryan Maness. *Cyber Strategy: The Evolving Character of Power and Coercion.* (2018).



A Brief Typology

This section contains a brief visualization of the criteria discussed above in order to describe the typological framework behind classifying data regimes. The first two criteria — cross-border data flows and personal data protection — differentiate data regimes based on the *nature* of the policies. For these two criteria, countries either meet a qualification or they do not. The remaining criteria involve differences in *degree* — that is, they measure the position that a country occupies on a spectrum from high to low capacity based on the content of the criterion. The table below is not exhaustive and should serve as a starting guide for further analysis.

Criteria	Data regime variance		
Cross Border Data Flows and Localization	Hard Localization – protectionist	Conditional restrictions – hybrid and soft localization	No localization requirements – liberal
Personal Data Protection and Privacy Frameworks	Data protection laws – consumer protection, human rights, cybersecurity	Data protection laws with lack of robust enforceability – conditional trade orientation	Little to no data privacy laws
Digital Industrial Policy	Incumbency – strengthening capacity to prevent competitive takeover	Influencer – Using industrial policy to further develop digital capacity	Consolidator – taking advantage of fragmentation to consolidate power
Artificial Intelligence Research & Development	High to medium levels – robust VC and startup environments, talent attraction	Medium to low levels – growing market potential, mild talent attraction, talent creators	Low levels to no specified R&D – lagging behind significantly, little talent attraction, lack of educational capacity
Infrastructure and Development	High capacity – robust network connectivity, active digital culture and information economy	Medium capacity – network connectivity but lagging in coverage, online population but lacking independent e-commerce and e-government services.	Low capacity – little or failing infrastructure, low online capacity and digital services
Administrative Capacity and Regulation	Strong – enforcement of laws on local levels, ability to harness data analytics and build databases	Moderate – enforcement of laws but lagging in some governmental services	Low – little to no enforcement, problems with governmental control
Leadership and Power	High – perceived and actual global leadership in digital space and cybersecurity	Medium – regional leadership, cybersecurity capacity and influence	Low – little to no independent cybersecurity capacity, lack of leadership

Table 1: Data Regime Typological Matrix

Conclusion

As the Internet continues to fragment and create space for the formation of new norms in the digital sphere, governments around the world are beginning to adopt a plethora of policies in order to assert more control over their national destinies and pursue their own interests. The divergence of data governance practices among states and the emergence of special data regimes deserves particular attention since the future rules of interaction and digital control will emerge largely from contemporary trends in cyberspace. In order to better prepare policymakers for the task of minimizing harm in a global data conflict, this paper proposed the concept of data regimes, or the unique combination of governance capacity, economic policies, and behavioral practices that concretize a cognizable posture towards data governance. Specifically, the concept of data regimes helps address some of the gaps in other notable typologies of emerging issues in data governance, artificial intelligence and the digital economy by incorporating the geopolitical and economic divergences of current approaches to regulating data. As such, data regimes strike at the core of how states are actively creating and defining the scope of digital space and the Internet. The paper outlined their core criteria and the future trends likely to emerge over the next decade.

The paper outlined the components of data regimes by proposing a set of criteria in order to create a workable framework for understanding the growing fragmentation of digital space. First, policies that restrict the flow of data between countries highlight two important features of data regimes: 1) the ability of the state to directly control digital trade; and 2) the ability of the state to monitor digital industries within its borders. Such policies have multiple rationales including national security and trade leverage. One such rationale, the protection of personal data, occupies a special position for data regimes because it involves the social issue

of privacy and reflects how a government perceives its relationship to its people. Next, emerging trends in digital industrial policy also play a role in global fragmentation: countries may pursue initiatives to better equip their domestic industries for the digital future and in doing so create harmful competition with other states.

Yet not all states may participate equally in this competition. Those with higher levels of concentrated R&D in digital industries and robust educational systems have a greater chance of steering the general future of data and using the private sector to create global norms than those countries lacking in these indicators. Similarly, without the infrastructural network capacity to make digital activity possible, a state stands on unequal footing with larger players, which in turn influences the choice of policy that reflect differences in data governance approaches. Moreover, data regimes also hinge on the capacity of the state to functionally administer its laws in consistent and predictable ways. Like infrastructural development, a country with legislative initiatives to influence digital space may fall short of its objective without the ability to enforce its laws or adequately balance private and public interests. Consequently, this may adversely affect a country's reputation for digital influence and therefore will help account for divergence in data regimes.

Finally, any framework for understanding how governments regulate data must take into account leadership and power. Countries may possess a perceived leadership role in the creation of digital norms and projection of digital power. Leadership comes from both diplomatic finesse and military power. Military power concerns not only the ability of states to aggressively target key assets through cyberwarfare but also the ability to defend and deter foreign influence. Such abilities implicate existential concerns of societies as economies everywhere become more reliant on digital technologies.

As the world continues to fragment, understanding where and how countries diverge in their interests will become key to identifying and preventing widespread abuse of technology, harmful global power competition, and increased international tension. We hope that the conceptual lens of data regimes may help address these concerns.

References

Adam Przeworski, *Capitalism and Social Democracy* (Cambridge: Cambridge University Press,

1986).

Akemi Suzuki & Tomohiro Sekiguchi, "Data Protection & Privacy" *Nagashima Ohno &*

Tsunematsu. (2019). https://gettingthedealthrough.com/area/52/jurisdiction/36/data-protection-privacy-japan/

Alena Epifanova. "Deciphering Russia's 'Sovereign Internet Law'" No. 2. *German Council on*

Foreign Relations (January 2020). https://dgap.org/en/research/publications/deciphering-russias-sovereign-internet-law

Alexander Wendt. *Social Theory of International Politics* (Cambridge: Cambridge University

Press, 1999).

Anne Nelson. "Cuba's Parallel Worlds: Digital Media Crosses the Divide," *Center for*

International Media Assistance (August 2016)

Anupam Chander & Uyên P. Lê. "Data Nationalism." *Emory Law Journal*, Vol. 64 (November

2015), pp. 677-739.

Arindrajit Basu et al. "The Localization Gambit" *The Centre for Internet and Society, India*.

(March 2019), https://cis-india.org/internet-governance/resources/the-localisation-gambit.pdf

Aykut Atali, Chandra Gnanasambandam, & Bhargs Srivathsan. "Transforming Infrastructure

Operations for a Hybrid-Cloud World. *McKinsey & Company* (October 2019) https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/transforming-infrastructure-operations-for-a-hybrid-cloud-world#

Barbara L. Cohn, "Data Governance: A Quality Imperative in the Era of Big Data, Open Data, and Beyond." *Journal of Law and Policy for the Information Society* Vol. 10, No. 3 (2015), pp. 811-826.

Benjamin M. Jensen, Brandon Valeriano, & Ryan Maness. *Cyber Strategy: The Evolving Character of Power and Coercion* (Oxford: Oxford University Press 2018).

Beryl. A. Radin. "The Instruments of Intergovernmental Management." *Handbook of Public* Administration (London: Sage 2003).

Carole Pateman. *Participation and Democratic Theory* (Cambridge: Cambridge University Press 1970).

Catherine Tucker. "Network Effects and Market Power: What Have We Learned in the Last Decade?" *Antitrust* (2018). http://sites.bu.edu/tpri/files/2018/07/tucker-network-effects-antitrust2018.pdf

Christopher Foster & Shamel Azmeh. "Latercomer Economies and National Digital Policy: An Industrial Policy Perspective." *The Journal of Development Studies* (November 2019).

Dan Senor & Saul Singer. *Start-up Nation: The Story of Israel's Economic Miracle* (New York City: Twelve 2009).

Daniel J. Solove & Paul M. Schwartz. *Privacy Law Fundamentals*. (2015).

Daniel Trottier & Christian Fuchs. "Theorizing Social Media, Politics and the State." *Social Media, Politics and the State: Protests, Revolutions, Riots, Crime and Policing in the Age of Facebook, Twitter and Youtube.* (New York: Routledge 2014).

Dani Rodrick. "Political Economy of Trade Policy." *Handbook of International Economics* Vol. 3 (1995), pp. 1457-1494.

Eileen Donahoe & Megan MacDuffee Metzger, "Artificial Intelligence and Human Rights" *Journal of Democracy*, Vol. 30, No. 2 (April 2019).

G. William Domhoff. *Who Rules America?* (Upper Saddle River: Prentice-Hall 1967).

Grace Kiser & Yoan Mantha. "Global AI Talent Report 2019." *Jfgagne*. (April 2019). https://jfgagne.ai/talent-2019/

Ha-Joon Chang, *Kicking Away the Ladder: Development Strategy in Historical Perspective* (New York: Anthem Press 2003);

Harry Eckstein. *Division and Cohesion in Democracy* (Princeton: Princeton University Press 1966).

Giorgio Agamben, *Homo Sacer: Sovereign Power and Bare Life* (1995).

Giovanni Arrighi. *The Long Twentieth Century* (New York: Verso 1994).

Immanuel Wallerstein, *The Modern World System*. (Cambridge: Cambridge University Press, 1974);

J.W. Henderson. *The Globalization of High Technology Production*. (London: Routledge 1989).

Jagdish Bhagwati, Pravin Krishna, & Arvind Panagariya. "The World Trade System: Trends and

Challenges." Presented at the Conference on *Trade and Flag: The Changing Balance of Power in the Multilateral Trade System* (2014).

Jeffery Ding. "Deciphering China's AI Dream." *Future of Humanity Institute* (March 2018).

Joel S. Migdal. *State in Society* (Cambridge: Cambridge University Press 2012).

John H. Chestnut. "U.S. vs. European Broadband Deployment: What Do the Data Say?" *Center for Technology, Innovation and Competition* (June 2014).

Joseph Nye. *Soft Power: The Means to Success in World Politics*, (New York City: PublicAffairs 2004).

Joseph Schumpeter. *Capitalism, Socialism, and Democracy* (New York: Harper & Brothers 1943).

Justin Fendos. "Lessons from South Korea's COVID-19 Outbreak: The Good, Bad, and Ugly."

The Diplomat, (March 2020). https://thediplomat.com/2020/03/lessons-from-south-koreas-covid-19-outbreak-the-good-bad-and-ugly/

Jurgen Habermas. *Legitimation Crisis* (Boston: Beacon Press 1973).

Karishma Banga & Dirk Willem te Velde. "How to Grow Manufacturing and Create Jobs in a Digital Economy: 10 Policy Priorities for Kenya." *Supporting Economic Transformation* (November 2018).

Kenneth A. Bamberger & Deirdre K. Mulligan, *Privacy on the Ground* (Cambridge: MIT Press 2015).

Kensaku Takase, "GDPR matchup: Japan's Act on the Protection of Personal Information."

IAPP (August 2017). https://iapp.org/news/a/gdpr-matchup-japans-act-on-the-protection-of-personal-information/

Klaus Schwab. *The Fourth Industrial Revolution* (New South Wales: Currency, 2016).

Lee Junkyu. "Korea's Trade Structure and Its Policy Challenges." *The Future of Korean Trade Policy* (2012).

Lynsey Chutel. "China Is Exporting Facial Recognition Software to Africa, Expanding its Vast Database." *Quartz Africa*. (May 2018). https://qz.com/africa/1287675/china-is-exporting-facial-recognition-to-africa-ensuring-ai-dominance-through-diversity/

Margaret Groarke, "Power, Agency and Structure," *New Political Science*, Vol. 14, No. 1 (1993).

Mark Latonero, "Governing Artificial Intelligence: Upholding Human Rights," *Data & Society* (October 2018), https://datasociety.net/wp-content/uploads/2018/10/DataSociety_Governing_Artificial_Intelligence_Upholding_Human_Rights.pdf

Martina F. Ferracane & Hosuk Lee-Makiyama. "China's Technology Protectionism and its Non-Negotiable Rationales. *European Centre for International Political Economy* (June 2017).

Martin van Creveld. *The Transformation of War* (New York: The Free Press 1991).

Max Weber, "Politics as a Vocation." *Max Weber: Essays in Sociology*, translated by H. H. Gerth and C. Wright Mills (Oxford: Oxford University Press 1946);

Meltzer, J. Lovelock, P. "Regulating for a Digital Economy: Understanding the Importance of Cross-Border Data Flows in Asia. *Brookings Institution* (March 2018).

Melvin Kranzberg, "The Information Age: Evolution or Revolution?" in *Information Technologies and Social Transformation*, edited by Bruce Guile (Washington, D.C.: National Academy Press, 1985).

Michael Chui et al. "Notes from the AI Frontier" *McKinsey Global Institute* (September 2018).

Michael Mann. *The Sources of Social Power* (Cambridge: Cambridge University Press 1986).

Nick Srnicek, *Platform Capitalism* (Cambridge: Polity 2016).

Nick Wallace & Daniel Castro. "The Impact of the EU's New Data Protection Regulation on AI." *Center for Data Innovation* (March 2018).

Nir Khsetri, "Success of Crowd-based Online Technology in Fundraising: An Institutional Perspective." *Journal of International Management*, Vol. 21, No. 2. (2015), pp. 101-116.

OECD "Data Governance in the Public Sector," *The Path to Becoming a Data-Driven Public Sector* (November 2019).

Peter J. May. "Policy Design and Implementation." *Handbook of Public Administration* (London: Sage 2003)

Peter Morgan. "Capacity Development: An Introduction." *Emerging Issues in Capacity Development: Proceedings of a Workshop* (1993).

Prema-chandra Athukorala. "Trade Policy Reforms and the Structure of Protection in Vietnam." *The World Economy*, Vol. 29. No. 2. (2006), pp.161-187.

Robert D. Atkinson. "A Policymaker's Guide to Digital Infrastructure." *Information Technology & Innovation Foundation* (2016).

Robert A. Dahl. *Who Governs?* (New Haven: Yale University Press 1961).

Robert Gellman. "Fair Information Practices: A Basic History." *SSRN* (June 2019). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2415020

Robert Reiner. *The Politics of the Police*. (Oxford: Oxford University,2010); Kubo Mačák. "Is the International Law of Cyber Security in Crisis?" *2016 8th International Conference on Cyber Conflict*, (2016), pp. 132-134.

Ross M. Rustici. "Cyberweapons: Leveling the International Playing Field." *Parameters*. (2011).

Ryan Hass & Zach Balin. "US-China Relations in the Age of Artificial Intelligence." *The Brookings Institution*. (January 2019), https://www.brookings.edu/research/us-china-relations-in-the-age-of-artificial-intelligence/;

Samm Sacks & Justin Sherman, "Global Data Governance: Concepts, Obstacles, and Prospects" *New America* (December 2019). https://www.newamerica.org/cybersecurity-initiative/reports/global-data-governance/

Scott Lash & John Urry. *The End of Organized Capitalism*. (Madison: University of Wisconsin Press,1987), pp. 196-209.

Sean McDonald & Xiao Mina, "The War-Torn Web" *Foreign Policy*. (December 2018).

Shanhong Liu, "Big Data-Statistics & Facts" *Statista* (October 2019). https://www.statista.com/topics/1464/big-data/

Stephen McNamee & Michael Glasser. "The Power Concept in Sociology: A Theoretical Assessment." *Humboldt Journal of Social Relations*, Vol. 15, No. 1. (1987), pp. 79-104.

Stephen P. Mulligan. "Cross-Border Data Sharing Under the CLOUD Act." R45173 *Congressional Research Service*. (April 2018).

Steve MacFeely & Nour Barnat. "Statistical Capacity Building for Sustainable Development: Developing the Fundamental Pillars Necessary for Modern National Statistical Systems." *United Nations Economic Commission for Europe* (November 2017), pp. 2-4.

Steven Lukes. *Power: A Radical View* (London: Macmillan,1974).

Stuart Fowler et al. "Developing Cyber-Security Policies that Penetrate Australian Defense Acquisitions." *Australian Defense Force Journal*. (2017)

Susan Aaronson, "The Turn to Trade Agreements to Regulate the Internet," in Jean-Baptise Velut et al. in *Understanding Mega-Free Trade Agreements: The Political and Economic Governance of New Cross-Regionalism*. (Oxford: Routledge, September 2017);

Steven Feldstein. "The Global Expansion of AI Surveillance." *Carnegie Endowment for International Peace*. (September 2019).

Tessaleno Devezas et al. *Industry 4.0.: Entrepreneurship and Structural Change in the New Digital Landscape*. (New York: Springer, 2016);

Thomas A Singlehurst et al. "ePrivacy and Data Protection" *CitiGroup*. (March 2017).

Thio Tse Gan, "Data and privacy protection in ASEAN – what does it mean for business in the region?" *Deloitte*. (2018);

Tom Forester. *The Information Technology Revolution*. (Cambridge: MIT Press,1985);

Will Edwards, "North Korea as a Cyber Threat," *The Cypher Brief*, (July 2016).

William A. Carter & William D. Crumpler. "Smart Money on Chinese Advances in AI." *Center for Strategic and International Studies* (September 2019).

"Artificial Intelligence Index Report 2019." *Stanford Institute for Human-Centered Artificial Intelligence* (2019), p. 5, https://hai.stanford.edu/sites/default/files/ai_index_2019_report.pdf

"Commodification of Cyber Capabilities: A Grand Cyber Arms Bazaar." *2019 Public-Private Analytic Exchange Program.* (2019), pp. 14-18.

"Confronting the Crisis of Global Governance." *Report of the Commission on Global Security,Justice and Governance*, (June 2015).

"Data Localization: A Challenge to Global Commerce and the Free Flow of Information." *Albright Stonebridge Group* (September 2015), pp. 12-15. https://www.albrightstonebridge.com/files/ASG%20Data%20Localization%20Report%20-%20September%202015.pdf

"Digital Globalization: The New Era of Global Flows." *McKinsey Institute.* (March 2016). https://www.the-digital-insurer.com/wp-content/uploads/2016/06/709-mgi_digital_globalization.pdf

"Global Lessons from Durban's Climate Change Challenges." *The New Humanitarian*. (May 2011). http://www.thenewhumanitarian.org/news/2011/05/24/global-lessons-durban-s-climate-change-challenges

"Global Talent 2021: How the new geography of talent will transform human resource strategies." *Oxford Economics.* (2020), https://www.oxfordeconomics.com/Media/Default/Thought%20Leadership/global-talent-2021.pdf;

"Global Talent Risk – Seven Responses." *World Economic Forum* (2011), p. 9. http://www3.weforum.org/docs/PS_WEF_GlobalTalentRisk_Report_2011.pdf

"Global Telecommunications Study 2019" *Ernst & Young* (September 2019), p. 4. https://www.ey.com/Publication/vwLUAssets/ey-accelerating-the-intelligent-enterprise/$FILE/ey-accelerating-the-intelligent-enterprise.pdf

"Globalization in Transition: The Future of Trade and Value Chains." *McKinsey Global Institute*. (January 2019), p. 14. https://www.mckinsey.com/~/media/McKinsey/Featured%20Insights/Innovation/Globalization%20in%20transition%20The%20future%20of%20trade%20and%20value%20chains/MGI-Globalization-in-transition-The-future-of-trade-and-value-chains-In-Brief.pdf

"Government Artificial Intelligence Readiness Index 2019," *Oxford Insights* (2019), https://www.oxfordinsights.com/ai-readiness2019

"Human Rights in the Age of Artificial Intelligence," *Access Now* (November 2018)

"How to Prevent Discriminatory Outcomes in Machine Learning" White Paper, *World Economic Forum* (March 2018), http://www3.weforum.org/docs/WEF_40065_White_Paper_How_to_Prevent_Discriminatory_Outcomes_in_Machine_Learning.pdf

"International Privacy Standards." *The Electronic Frontier Foundation.* (2018). https://www.eff.org/issues/international-privacy-standards

"Measuring the Digital Transformation: A Roadmap for the Future." *OECD* (March 2019), p. 20. https://www.oecd-ilibrary.org/sites/9789264311992 en/index.html?itemId=/content/publication/9789264311992-en

"National Strategy for Artificial Intelligence #AIForAll." *Niti Aayog*. (June 2018), https://niti.gov.in/writereaddata/files/document_publication/NationalStrategy-for-AI-Discussion-Paper.pdf

"State of Privacy in Mexico." *Privacy International* (January 2019). https://privacyinternational.org/state-privacy/1006/state-privacy-mexico

"The Data Protection Regime in China." *European Parliament's Directorate General for International Policies*. (2015).

"The Keys to Data Protection." *Privacy International*. (August 2018), https://privacyinternational.org/sites/default/files/2018-09/Data%20Protection%20COMPLETE.pdf;

"Unlocking the Digital Economy Potential of the Kingdom of Saudi Arabia." *Ernst & Young*, (2019), https://www.ey.com/Publication/vwLUAssets/ey-unlocking-the-digital-economy-potential-of-the-kingdom-of-saudi-arabia/$File/ey-unlocking-the-digital-economy-potential-of-the-kingdom-of-saudi-arabia.pdf

"Unlocking the Potential of India's Data Economy. *Omidyar Network India*. (September 2019), https://www.omidyarnetwork.in/insights/unlocking-the-potential-of-indias-data-economy-practices-privacy-and-governance