Volume 5, Number 1, Spring 2021

Editorial

Guest Editorial

I am delighted that this collaboration between *Retskraft – Copenhagen Journal of Legal Studies* and Artificial Intelligence and Legal Disruption (AI-LeD)[1] has come to fruition. In my mind, Retskraft fills the important role in any Law Faculty worth its salt by giving its students experience with regard to the academic publishing process: as writers and contributors, and as editors. Beyond this, however, I think that a student-edited law journal such as Retskraft provides an important platform for publishing the best work produced by its student body, especially where that work may provide relevant and timely inputs to the public policy debates.

As the convenor of the AI-LeD master's elective,[2] I have been simultaneously impressed by the quality of some of the final essays that my students submitted, and frustrated by the fact that I was essentially the only person who would get to read this work. These final essays often developed original ideas or wove disparate concepts together in intricate ways, and the thought often crossed my mind that there were several papers from each class which could pass peer-review.

The "grading" model, in which the professor-assessors grade the paper in front of them in private, may make sense in more orthodox courses that seek to assess the student's comprehension of an established legal field and where the emphasis is placed upon the student to demonstrate mastery.[3] A different way of putting this might be that there is no need to consider publication of student work in the context of "orthodox" education because the student is not pushed to undertake original research. If the model of education does not envisage original work, then concomitantly there is no need to contemplate publication.

In AI-LeD, however, the aim of the final written assessment is to engage the student in actual research (or failing that, at least research-integrated work), and

[1] https://jura.ku.dk/english/ai-led/ or https://jura.ku.dk/ai-led-dansk/

[2] https://kurser.ku.dk/course/jjua55235u/

[3] Akin to the old apprenticeship model where the apprentice produces a "masterpiece" of a sufficiently high standard to attain membership in a guild or academy.

not "just" education, so the possibility for subsequent publication of good work is only fitting. With this in mind, and once Retskraft was up and running, I ran a pilot for this Special Issue with a former student with his article 'Artificial Intelligence in Court' in an earlier volume of this Journal.[4] The success of that pilot project suggested that we would build up the momentum, both in terms of pushing the students in the course towards more risk-taking research papers, and with Retskraft for providing a platform for such work.

This in turn changed the nature and orientation of the AI-LeD course, which can be maddeningly research-integrated (anecdotally at least, from the student's perspective), and probably research-obsessed: there is complete free-rein within the wide parameters of the course for students to define their own topic, approach, and execution of their final written assessments. Contextualise this broad latitude for the paper within a problem-*finding* orientation[5] that informs the course, where students are expected to proactively explore the potential policy problem space opened up or revealed by AI, and we have a recipe for confusion, uncertainty, creativity, and criticality. Indeed, if legal education was a pizzeria most courses would simply take your order, make your pizza, and then bring it to you. AI-LeD, on the other hand, would invite you behind the counter, present you with the ingredients and the wood-fired oven, show you a few throwing techniques, and leave you to it (with a few pointers for those who want them). My hope in setting such loose parameters is to set the students up to potentially produce some fresh and original work, some of which is showcased in this Special Issue.[6]

Beyond the specific context of the course, the AI-LeD course is also embedded within a Research Group at the Faculty of Law of the same name, as well as a burgeoning approach to the law, regulation, and governance relating to

---

[4] Thomas Buocz, 'Artificial Intelligence in Court' (2018) 2(1) Retskraft – Copenhagen Journal of Legal Studies 41.

[5] Hin-Yan Liu and Matthijs M Maas, '"Solving for X?": Towards a Problem-Finding Framework That Grounds Long-Term Governance Strategies for Artificial Intelligence' (2021) 126 Futures 102672.

[6] With the exception of Karen M. Richmond's contribution, 'AI, Machine Learning, and International Criminal Investigations', the contributions all originated from final written assessments submitted for the class.

AI.[7] The attraction and difficulty with AI-LeD is that it combines the two complex moving parts: Artificial Intelligence and Legal Disruption. As we have sketched out the AI-LeD model[8] there are benefits to adopting "legal disruption"[9] as the focal point for law and policy, and to deploy AI as the driver and lens through which to identify these trends and meet their challenges. As an approach, a framework, or a model, AI-LeD thus provides some structure and direction when attempting the problem-*finding* work that we had advocated for.[10]

We have only begun to scratch the surface of both AI and legal disruption, but the publication of this Special Issue I think signals an important milestone in this endeavour to make sense of the legal, regulatory, and governance implications raised by AI and its applications.

*Hin-Yan Liu**

Editorial

The Editorial Board would like to extend its thanks to Dr. Liu for proposing this special issue on *Artificial Intelligence and Legal Disruption*, and all the authors who have contributed articles to it. The free reign of the AI-LeD course described by Dr. Liu in his guest editorial is evident from the breadth of topics covered by the articles, and we hope that there will be something of interest to any reader with an interest in the interplay between artificial intelligence and the law.

—

In the inaugural editorial of *Retskraft*, the Editorial Board explained how Danish legal education 'was focused on craftmanship rather than scientific production',

---

[7] Hin-Yan Liu and others, 'Artificial Intelligence and Legal Disruption: A New Model for Analysis' (2020) 12 Law, Innovation and Technology 205.

[8] ibid.

[9] See also, Roger Brownsword, 'Law Disrupted, Law Re-Imagined, Law Re-Invented' [2019] Technology and Regulation 10; Roger Brownsword, *Law, Technology and Society: Reimagining the Regulatory Environment* (Routledge 2019).

[10] Liu and Maas (n 5).

* Associate Professor, and Coordinator of the Artificial Intelligence and Legal Disruption Research Group, Faculty of Law, University of Copenhagen [hin-yan.liu@jur.ku.dk]

and that the Journal was founded with an explicit goal of fostering a scientific approach to law among students and encouraging contributions investigating 'how the law is produced, [and] how it operates and impacts society.'[1] This issue reflects this scientific promise – while perhaps flipping the script in describing how *the law is impacted* by AI, and not the other way around – and goes somewhat beyond the typical boundaries of legal research with the problem-finding as opposed to problem-solving approach described by Dr. Liu.[2]

The advent of this special issue also touches upon another aspect of the founding of *Retskraft* which is the idea that scientific inquiry into law can take many different forms and utilize a plurality of theories and methodologies.[3] While traditional doctrinal legal scholarship – the definition of which is in itself disputed – remains a central part of legal education, students should be exposed to other ways of examining the law and its effects as part of their education. The contributions to this issue contain both descriptive and normative elements, questions regarding concrete rules and more philosophical and principled questions, showing how courses like AI-LeD, and other research and methods-oriented courses, have an important part of play in this area. It is our hope that *Retskraft* will remain an attractive avenue for publishing varied legal scholarship.

——

[1] 'Editorial' (2017) 1(1) Retskraft – Copenhagen Journal of Legal Studies 1, 3. It is sometimes pointed out that the use of the English words 'science' and 'scientific' when discussing law can be misleading due to those words having connotations related to the natural sciences or quantitative social science. In this context, it is used in the sense of the Danish *videnskab* or the German *wissenschaft*, which also encompass the other disciplines in academia. See Jakob vH Holtermann and Mikael Rask Madsen, 'European New Legal Realism: Towards a Basic Science of Law' in Shauhin Talesh, Elizabeth Mertz and Heinz Klug (eds), *Research Handbook on Modern Legal Realism* (Edward Elgar Publishing 2021) 68.

[2] While the problem-finding/problem-solving distinction used by Dr. Liu has a particular definition, parallels can be drawn to other critiques of legal scholarship. Cf Hin-Yan Liu and Matthijs M Maas, '"Solving for X?": Towards a Problem-Finding Framework That Grounds Long-Term Governance Strategies for Artificial Intelligence' (2021) 126 Futures 102672, pt 2.1.4.; Rob van Gestel and Hans-Wolfgang Micklitz, 'Why Methods Matter in European Legal Scholarship' (2014) 20 European Law Journal 292, 302.

[3] 'Editorial' (n 1).

This issue marks the first time that we have published an issue dedicated exclusively to a specific topic. As readers of volume 4, issue 1 will know, the subsequent issue will also be on a specific topic, *EU Law & Politics*.

The process of working on a special issue has not been markedly different than that of a regular issue as far as the articles go. The standard article screening and selection procedure was followed, with the exception that we now had to determine whether an article was 'within scope' of the special issue. The most noticeable difference was the difficulty in finding subject matter experts to conduct the peer reviews. Finding, say, a scholar of general criminal law who has enough time to conduct a peer review can be difficult enough, but when one needs to find someone who is both knowledgeable about artificial intelligence and a particular legal subfield, the process becomes more arduous. *Retskraft*, like most scholarly journals that use a peer-review system, is dependent on volunteer reviewers to evaluate the quality of articles, and we are extremely thankful toward the reviewers who have given their time and expertise for this issue.

Given the positive experience we have had working on this issue, we will continue to host themed contributions in the future. In order to allow for regular publishing of non-thematic articles, we will most likely opt for a symposium model, where collections of subject-specific articles can be published alongside regular articles.[4] Once COVID-19 restrictions have been lifted, this could be combined with conferences where students present and discuss each other's work. Students at the Faculty of Law at the University of Copenhagen with an interest in organizing such events should feel free to get in contact with the Editorial Board.

—

The present issue contains five articles, which, despite the common topic of artificial intelligence and legal disruption, span a wide range of issues.

First, Robbe van Rossem uses the issues that arise when proxies for protected characteristics exist in the datasets used by AI, to critically examine the limits of discrimination law.

[4] See, eg, (2019) 10 Journal of International Humanitarian Legal Studies 77–202; (2020) 31 European Journal of International Law 489–619.

Second, Karen M. Richmond uses the history of national litigation concerning probabilistic genotyping in DNA analysis to examine questions of opacity that might arise in the use of forensic artificial intelligence, with a focus on these questions as they relate to international criminal justice.

Third, Laure Helene Prevignano examines how the use of artificial intelligence might blur the public/private law distinction central to most legal systems.

Fourth, Anna Kirby examines how artificial intelligence will affect the field of international diplomatic law.

Finally, Caroline Serbanescu examines whether manipulation enabled by artificial intelligence will disrupt, and therefore threaten, the concept of democracy.

We once again thank the authors for their contributions, and Dr. Liu for proposing the special issue, and hope that you enjoy reading the issue.

Volume 5, Number 1, Spring 2021

Proxy Discrimination and Legal Disruption

The disruptive power of reality

Robbe van Rossem[*]

Reality is mirrored in the many algorithmic systems that are increasingly embedded in our everyday life. When data that refers to real-world phenomena is used in algorithmic systems, an insightful reconstruction of reality is generated. This image of reality becomes more complete as greater amounts of data are involved and as this data is interpreted more intelligently. The trend of a greater mirroring of reality can, however, also trigger a legal disruption, as the law can be confronted with a reality alternative to the one it implies itself. This risk exists particularly in the context of discrimination. In its application to the algorithmic context, non-discrimination law has to examine the very systems that generate a mirroring of reality. This paper investigates the disruptive effects such a confrontation with reality can have for the law in the particular case of proxy discrimination. The features of discriminatory proxies are namely highly descriptive of the structural inequality and discrimination that characterizes society. When theories critical of the limits of non-discrimination law are subsequently confirmed by the reflections in the data, the law faces increased pressure to justify its current scope of and approach to illegal discrimination. While a true disruption depends on the willingness of the law to take on an position of self-reflection, it is argued that any distortion arising from the reflections in the data can hardly be called technological in nature.

[*] University of Copenhagen – University of Ghent [robbevanrossem@gmail.com]

1. Introduction

Reality is reflected in the data our world generates. In our so-called information society, little of this reality is free from being captured in digital form. Vast amounts of data are imported from analogue collections, captured in our online behaviour, collected through the observations of scientific research, etc. The recording and collection of all this data is, however, not without its purpose – and certainly not without its use. Big Data has proven that the consideration of great amounts of information can be extremely valuable for i.a. making decisions or predictions. Great interest thus exists to subject large proportions of data to processes of interpretation such as data mining. The results can be astounding as the data can reveal more than what we thought to know about our world. As human curiosity – or simply the desire for efficiency, knows little to no limit, also more intelligent technology like artificial intelligence has been put to the task to get the most out of our data. As a result, reality is increasingly being mirrored in the systems we use to parse it.

A 'boxed-in' overview of reality can be very enlightening. It is, however, the question whether the law is capable of dealing with the revelations that come with this increased understanding of the world we live in. The reflections of reality that can be found in systems subject to the law's control could easily prove themselves to be overwhelming to the law, and as a result be disruptive. This is potentially the case in the context of discrimination. Algorithmic systems have been plagued by discriminatory results. While algorithmic discrimination always has caused a variety of difficulties for the application of non-discrimination law, the trend of an increased insight in the reality of discrimination could be especially problematic in this regard. After all, the legal frameworks that exist to protect the right to non-discrimination are often criticized for their blindness regarding reality and the discrimination that occurs in it. Now this reality is reflected in the systems that are subjected to the law's examination, non-discrimination law's claimed ignorance towards certain aspects of discrimination is again challenged.

The paper explores the possibility of such a disruption in the particular case of proxy discrimination. This particular form of algorithmic discrimination occurs when information on protected, discriminatory-sensitive characteristics is hidden in other, seemingly neutral data that is used by an algorithm. The

protected characteristics are themselves not included in the data, yet highly correlate with information that is fed to the algorithm – which are called their 'proxies'. When this data is used as an input for the algorithm, the protected characteristics can indirectly influence the algorithm's output, and as a result place the members of the protected group or class at a possibly illegal disadvantage. The discriminatory effect thus occurs because of the algorithm's reliance on information that also happens to be indicative for a protected class. A postal code can, for example, function as a proxy for the protected characteristic race, considering neighbourhoods can have racial profiles due to the ethnicity of their inhabitants. Subsequently, when a decision is based on subjects' postal code, inhabitants of historically racialized neighbourhoods can be discriminated, as the decision will indirectly be based on their race or ethnicity, even though the feature was not directly included in the decision-making process.

The paper commences with an explanation of the occurrence of proxies in datasets and the discrimination that can come from that (2). Next, the capability of an, at least partial, recreation of reality is demonstrated by the means of proxy discrimination's features (3). After a concise look at certain critiques of non-discrimination law's ignorance to the world it operates in (4), the paper discusses the disruption faced by the law and the unique nature that characterizes it (5).

2. Proxy discrimination

2.1 Proxies

Describing reality – for instance human beings – involves comparing corresponding features and adding significant values to as many as possible. Some of the features are independent and fully complementary (e.g. a first name and last name, a postal code and telephone number, …). Others are more or less related to each other (birth date and age, body weight and clothing size, …). When these features overlap to the extent that their correlation can cause them

to refer to the same information, they are regarded to be proxies to one another.[1] For example, your body mass index score (BMI) can indicate whether you are overweight. Similarly, the fact that your clothing size is XL or higher can bear the same information. As a result, a high BMI score and clothing size XXL are proxies for overweight but also for one another. From the perspective of the feature both information points relate to, their coexistence can thus be characterized by redundancy, as the pieces of data can easily substitute each other in a dataset while their mutual information remains intact. Whether this is favourable depends on the information that is reflected and the situation in which it is used.

2.2 Popularity of Proxies

The presence of proxies in data has increased significantly in the last years. In times when storing data was still cumbersome and expensive, redundant information such as proxies was always carefully avoided. In the age of Big Data such concerns are long gone. The capability to store vast amounts of data very cheaply has facilitated the trend of connecting and copying databases without any concern for identical information. The fact that these databases were developed from different perspectives actually adds information to the entire system, allowing for more patterns and conclusions to be found by Big Data tools. In a way, proxies have changed from being a nuisance to serving as a commodity. The popularity of proxies does, however, not necessarily solely relate to the coexistence of multiple substituting information points within the used datasets.

Alternatively, proxies can also be useful precisely when their counterpart is missing from a dataset. They are an efficient tool to include information in a dataset that itself is difficult to observe, unavailable or simply not allowed to be used. It can, for example, be very difficult and costly to determine someone's driving style. The observation would require multiple tests, interviews, field trials, etc. If, however, general test results would be available that indicate that

[1] Solon Barocas and Andrew D Selbst, 'Big Data's Disparate Impact' (2016) 104 California Law Review 671, 691.

male drivers predominately adopt an aggressive driving style, it is tempting to use the easily observable characteristic of gender as a proxy for someone's potential behaviour on the road.[2] Similarly, in the infamous practice of redlining financial institutions used postal codes in their decision to provide certain services such as granting loans.[3] Although areas can coincide with particular levels of income, it has been established how this choice was based in racial animus and prejudice.[4] In this way, geographic information functioned as a 'masked' replacement for an applicant's ethnicity or race, which is of course an illegal basis for differentiation.[5]

2.3 Proxy discrimination defined

The presence of proxies clearly cannot be considered to be desirable in all instances. When referring to certain sensitive characteristics, they can significantly add to the problem of algorithmic discrimination. Of course, algorithms are bound to discriminate in a technical sense; they are designed to tap into the vast amounts of data our 'scored society' generates for the exact purpose of evaluating, ranking, classifying, … subjects in a manner that exceeds human cognition and fatigue, which naturally implies differentiation.[6] While many of these differentiations are considered to be acceptable, illegal discrimination arises when they infringe the rules of non-discrimination law. For most frameworks of non-discrimination law, this implies that a differentiation was based on one of the societally important characteristics the law has rewarded

[2] Toon Calders and Indre Zliobaite, 'Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures' in Toon Calders and others (eds), *Discrimination and Privacy in the Information Society* (Springer 2013) 52–53.

[3] Hunt Bradford, 'Redlining', *Encyclopedia of Chicago* (2005) <http://www.encyclopedia.chicagohistory.org/pages/1050.html> accessed 4 January 2020.

[4] Barocas and Selbst (n 1) 689.

[5] Andrea Romei and Salvatore Ruggieri, 'Discrimination Data Analysis: A Multi-Disciplinary Bibliography', in Calders and others (n 2) 121.

[6] Claude Castellucia and Daniel Le Métayer, *Understanding Algorithmic Decision-Making: Opportunities and Challenges* (STOA 2019) 7.

a special legal protection.[7] Attributes commonly included in these 'protected characteristics' are race, gender, sexuality, religion, etc.[8] When one of these characteristics is used as a direct input or ground for a decision, the illegal discriminatory nature of the output is blatantly clear.[9] Simply excluding these characteristics from the model does, however, not always suffice to prevent a discriminatory result. Proxies for protected characteristics may namely be lurking in the data, allowing the prohibited characteristics to have a continuing influence on the output of the algorithm. It is this indirect effect protected characteristics can have through their proxies, that causes proxy discrimination.

In its simplest form, proxy discrimination can be defined as a differentiation based on facially-neutral characteristics that significantly correlate with membership to a protected class.[10] Although the protected characteristics are not directly involved in e.g. the decision-making process, they can have a similar discriminatory impact when they are represented by proxies that happen to be present in the data. The facially absent protected characteristics can thus be so-called 'redundantly encoded' in the dataset.[11] This is the case when 'a particular piece of data or certain values for that piece of data are highly correlated with membership in specific protected classes.'[12] Present by representation, the legally-prohibited characteristics continue to impact the output of the algorithm,

---

[7] Raphaële Xenidis and Linda Senden, 'EU Non-Discrimination Law in the Era of Artificial Intelligence: Mapping the Challenges of Algorithmic Discrimination' in Ulf Bernitz and others (eds), *General Principles of EU law and the EU Digital Order* (Kluwer Law International 2020) 5.

[8] Dagmar Schiek, Lisa Waddington and Mark Bell, *Cases, Materials and Text on National, Supranational and International Non-Discrimination Law* (Hart Publishing 2007) 510; Christopher McCrudden and Sacha Prechal, 'The Concepts of Equality and Non-Discrimination in Europe: A Practical Approach' (European Network of Legal Experts in the field of Gender Equality 2010) 60, 23.

[9] Xenidis and Senden (n 7) 19.

[10] Barocas and Selbst (n 1) 691–692. See also Anya Prince and Daniel Schwarcz, 'Proxy Discrimination in the Age of Artificial Intelligence and Big Data' (2019) 105 Iowa Law Review 1257, 1266 (who clarify that proxy discrimination relates more specifically to 'scenarios in which an algorithm uses a variable whose predictive power derives from its correlation with membership in the suspect class').

[11] Barocas and Selbst (n 1) 691.

[12] ibid 691–692.

and as a result place the members of a protected class at a possibly illegal disadvantage when they are subjected to the discretion of such an algorithm. In the classic example of redlining, for example, the decision to grant a loan based on a subject's postal code does not directly involve a protected characteristic. It can, however, indirectly amount to a proxy discrimination when areas and neighbourhoods highly correlate with racial profiles, as the subject's postal code would act as a proxy for their race or ethnicity. As proxy discrimination occurs in the form of a practice that facially appears to be neutral, yet disproportionally harms members of a protected class, it is often regarded as a specific subcategory of indirect discrimination.[13]

3. Mapping discrimination

Proxies add a great deal to the persistence of the problem of algorithmic discrimination. Their existence, however, also touches upon something more fundamental concerning the notion of discrimination itself. The redundant encodings offer a lens through which to observe discrimination not only as it appears in the algorithm, but also how it occurs in the real world. After all, it has to be reminded that data is a reflection of reality. In a way, an intelligent processing of data merely offers a cartography of the world we live in. The conclusions derived from the use of the data are only relevant given their analogy with what exists in the real world. Similarly, the information discriminatory proxies reflect and the relations they imply facilitate a 'mapping of discrimination'. This capability of proxy discriminations to map reality can be found in two of its features which coincidentally are of great importance in a judicial review on the illegal discriminatory nature of an algorithmic output.[14] This paper discusses the trade-off between fairness and utility proxies impose on the designers of algorithmic models (3.1) and the endless amounts of proxies that are redundantly encoded in the data (3.2) to conclude on the harsh truth both features bring (3.3).

[13] Prince and Schwarcz (n 10) 1260.

[14] See infra 5.1. on the legal relevance of these features.

3.1 Trade-off

Redundant encodings have proven to be a difficult problem to solve. In case they could be detected, their deletion or exclusion from the model is not always a viable option. The information that doubles as a proxy for membership to a protected class is often 'genuinely relevant in making rational and well-informed decisions'.[15] As is mentioned above, the use of geographical information like postal codes can, for example, lead to an illegal discriminatory effect for certain groups as neighbourhoods can have different racial profiles.[16] An individual's address can, however, be highly relevant in a job related context, as the distance between home and workplace is a strong indicator for employee engagement.[17] This confronts designers of algorithmic models with a difficult trade-off between fairness and utility.[18] While withholding proxies from the data could seem beneficial in an attempt to secure a non-discriminatory result, their exclusion implies a high cost for the overall accuracy of the model as meaningful information would be missing from the decision-making or prediction process.[19]

Although a difficult balancing exercise for the designers of algorithmic systems, the utility-fairness trade-off also showcases how a deeper look into proxies can provide a meaningful addition to our perception of the discrimination faced by certain groups. Through the correlation between sensitive characteristics on the one hand and attributes that are relevant for a rational and well-informed decision on the other, the trade-off indicates how class membership can impactfully condition which traits an individual possesses. After all, one of the main reasons why members of certain classes are systematically discriminated against when 'objective' target variables are used, is

[15] Barocas and Selbst (n 1) 691.
[16] Romei and Ruggieri (n 5) 121.
[17] Don Peck, 'They're Watching You at Work' (*The Atlantic*, December 2013) 72 <https://www.theatlantic.com/magazine/archive/2013/12/theyre-watching-you-at-work/354681/> accessed 6 January 2020.
[18] Philipp Hacker, 'Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies against Algorithmic Discrimination under EU Law' (2018) 55 Common Market Law Review 1143, 1150.
[19] Barocas and Selbst (n 1) 721.

that these relevant criteria happen to be possessed by classes at different rates.[20] This is of course no claim for superiority or inferiority of certain classes. Instead the phenomenon shines a light on the structural and systemic nature of discrimination.[21] The trade-off originates from the wish to avoid a discriminatory output. What risks such an outcome, is the translation of existing inequality in the disposition connected to class which is reflected in the data. By revealing the disposition and the permeating effects class membership can have, proxy discrimination forces the observer to place instances of discrimination in a broader context. Notably, discrimination can not only be the cause of inequality, it can also very well be the result of it.

3.2 Lines of proxies

In the event that a proxy is detected and the designer indeed sacrifices predictive accuracy by excluding it from the model, this decision can still be futile as there may be many more proxies for the same protected characteristic encoded in the data.[22] This possibility naturally increases as the amount of input data grows. In rich enough datasets, the chance for the redundant encoding of protected characteristics not only reaches near certainty, but often also presents itself in a way that the encodings are redundant to each other.[23] When a proxy is excluded for the purpose of a non-discriminatory output, other proxies for the same protected characteristic will simply continue the discriminatory effect.[24] In these instances 'endless lines of proxies' can be observed as the proxies can easily replace each other.[25] As a result, the attempt to exclude all proxies would have you block information at zero.[26] Even if it would be possible to design a system

[20] Sandra Mayson, 'Bias In, Bias Out' (2018) 128 The Yale Law Journal 2218, 2257–2259; Romei and Ruggieri (n 5) 130.
[21] Barocas and Selbst (n 1) 691.
[22] Ignacio Cofone, 'Algorithmic Discrimination Is an Information Problem' (2019) 70 Hastings Law Journal 1389, 1416.
[23] Barocas and Selbst (n 1) 695; ibid 1414.
[24] Cofone (n 22) 1414.
[25] Cofone (n 22) 1416.
[26] ibid 1414.

with such an objective, its purpose would quickly be defeated as the lack of remaining information would reduce the results to mere randomness.[27]

An important contribution to this obstinacy of proxy occurrence is the fact that proxies do not always present themselves in the form of clear, single substitutes for the protected characteristic that is aimed to be excluded from the model.[28] A data particle might also only slightly correlate with a protected characteristic, to a degree that it seems to be completely neutral when observed individually.[29] In aggregation, however, the correlations of the different information points could cluster into a proper proxy.[30]As a data particle's potential to contribute to the formation of dispersed proxies may only be revealed in aggregation, each data point can theoretically be suspected to hold such a dormant potential for a discriminatory output when it would be combined with the corresponding data points. Illustrative for such dispersed proxy formations are the various kinds of personal traits and attributes that can be observed through someone's activity on social media. A single like on a social networking site such as Facebook is unlikely to reveal the user's sexuality or political views. The accumulation of likes, however, allows social media platforms to observe precisely such highly sensitive personal attributes of their users.[31]

The seemingly infinite chain of proxies that can be observed in large data collections builds on the previous feature of the trade-off to allow for a mapping of discrimination. Where the utility-fairness trade-off highlights how attributes can be distributed unequally between classes, the proxy lines show how many

---

[27] ibid.
[28] ibid.
[29] ibid 1413.
[30] ibid 1414.
[31] Michal Kosinski, David Stillwell and Thore Graepel, 'Private Traits and Attributes Are Predictable from Digital Records of Human Behavior' (2013) 110 Proceedings of the National Academy of Sciences 5802, 5802; Solon Barocas, Moritz Hardt and Arvind Narayanan, Fairness and Machine Learning: Limitations and Opportunities (2019) ch 2 <https://fairmlbook.org/classification.html> accessed 26 October 2020 ('Several features that are slightly predictive of the sensitive attribute can be used to build high accuracy classifiers for that attribute').

attributes can actually be connected to the disadvantaged position membership to a certain class brings forth. Correlation per correlation, the inspection of a data collection for proxies of a protected characteristic shows how wide the impact of class membership can be. In a sense, the proxy lines unearth the branches of inequality and discrimination. Naturally the clarity of this image increases proportional to the data volume. The fact that the reflections of inequality are hardly inescapable in large data collections, that discriminatory potential can be luring even in the smallest things in life, and that this is sometimes only observable when the respective data points are placed in the right constellation of data, is similarly to the trade-off revealing with regards to the nature of discrimination. Namely, rather than the consequence of a particular decision to discriminate, the proxy discrimination seems to be an expression of the systems and environment we live in. It is exactly these processes of disadvantage that can be observed through their exclusionary consequences that are recorded in the data in the form of proxies for protected characteristics.

3.3 An (in)convenient truth

The features of discriminatory proxies unveil that it is reality that produces discriminatory practices, not the machine. The discriminatory results that roll out of an algorithm are not to be reduced to purely virtual phenomena. Their discriminatory nature stems from the real world, whose inherent inequality resonates in the data the machine is being fed.[32] The idea that the origins of algorithmic discrimination can also lay outside the algorithm is, however, not too shocking. Historical biases have been illustrative in this regard as they show that when data that reflects a discriminatory past is fed to an algorithm, the algorithm will reproduce similar discriminatory practices.[33] Sandra Mayson's play on the old computer-science adage 'garbage in, garbage out' wittily summarizes this as 'bias in, bias out.'[34] Thus, to the extent that the data actually

[32] Xenidis and Senden (n 7) 7.
[33] Barocas and Selbst (n 1) 681.
[34] Mayson (n 20) 2224.

represents reality, the algorithm is bound to perpetuate that reality to the same, unequal image.[35]

Proxies offer, however, a more impactful realisation than that inequality and discrimination are a product of the real world. The discussed features show that proxy discriminations also shine a light on the nature and the construction of discrimination. The use of its lessons thus exceeds the algorithmic context, as the revelations can also be insightful for the analysis of discrimination in general. Guided along the many proxies present in a model, the data reveals the effect membership to a class can have on the life of an individual and subsequently the decisions he or she is subjected to. In a way, proxy discriminations illustrate the interaction between inequality and discrimination and hint at the structural nature of both.[36] As a result, the exploration of discriminatory proxies places discriminatory practices in a broader context. It can, however, be questioned whether the occurring image of discrimination is compatible with the notion of discrimination held by the law. After all, after a walk along the 'contours of inequality', the current focus of non-discrimination law on individual cases of discrimination suddenly seems to be extremely narrow if not naïve.[37]

4. Non-discrimination law

Proxy discrimination provides us with a broader picture of discrimination than the single discriminatory acts the law tends to focus on. Naturally this can be of great advantage for the fight against illegal discrimination. The insight offered by the use of algorithms and AI with regard to the construction of inequality and discrimination can be used for technological, socio-political and possibly even legal progress.[38] As regards the legal dimension, it can, however, be questioned

[35] Xenidis and Senden (n 7) 7.

[36] Cf ibid 7–9 ('Structural discrimination, which is the product of past discrimination institutionalised over time and now reflected in many ways in the organisation of society, is mirrored in data').

[37] For 'contours of inequality' see Barocas and Selbst (n 1) 721.

[38] Cf Mayson (n 20) 2284 ('Because predictive algorithms transparently reflect inequality in the data from which they are built, they can also be deployed in reverse: as diagnostic tools to identify sites and causes of racial disparity in criminal justice').

whether the law can parse the many revelations a 'mapping of discrimination' brings forth. Many critiques concerning the established limits of non-discrimination law are based on the claim that the law is blind to certain aspects of discrimination as presented in reality. When these theories are confirmed in the reflections of reality that are captured in the data, it can be argued that non-discrimination law experiences an increased pressure to justify its scope if it does not change its approach. After all, with the implications of opposing theories lurking in the very systems it has to assess, the law faces greater difficulty in maintaining its particular conception of discrimination. The following part discusses a number of critiques on the limits of non-discrimination law relevant for the revelations of i.a. proxy discrimination.

4.1 Intersectionality

An often called upon limit of non-discrimination law is its tendency to address a discriminatory act from the perspective of only one characteristic.[39] This single-axis approach to discrimination is central to the critique formulated in the literature on intersectionality.[40] Coined by Kimberlé Crenshaw in her 1989 feminist critique of the US antidiscrimination doctrine, the concept of intersectionality denotes the various ways in which personal characteristics interact with each other and as a result shape unique experiences for those residing in their overlap.[41] When standing on an actual intersection, you could be hit by traffic coming not only from one direction, but from each direction, and possibly even at the same time.[42] Similarly can a person be discriminated against simultaneously on the basis of his or her gender, race, religion, etc. Crenshaw explains, however, that this does not necessarily result in situations of 'additive' discrimination, where a differentiation is based on the combination of

[39] Anna Lauren Hoffmann, 'Where Fairness Fails : Data , Algorithms , and the Limits of Antidiscrimination Discourse' (2019) 22 Information, Communication & Society 900, 905.
[40] ibid.
[41] Kimberle Crenshaw, 'Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory, and Antiracist Politics' [1989] University of Chicago Legal Forum 141.
[42] ibid 149.

multiple, yet still distinguishable grounds.[43] Characteristics can also interact in a way that their combination can no longer be disentangled.[44] The discrimination experienced by a black woman, for example, is not per se based on her gender or her race, nor necessarily on the accumulation of both grounds, but can instead be focused on her black womanhood in particular.[45]

Ignorance of these intragroup differences comes at a great risk. Crenshaw's critique of intersectionality sought more than a more accurate mapping of identity categories.[46] Central to her thinking was non-discrimination law's role in the reproduction of social hierarchy and inequality. She argued that by reducing experiences of discrimination to a single characteristic, the law banishes those whose experience cannot fully be grasped by one of the protected characteristics, to a permanent stay in the 'basement' of society.[47] This metaphorical basement will at one point host all disadvantaged people. Nevertheless, it reproduces the hierarchy that exists above ground.[48] A relative privilege is namely given to those whose experience can actually be fully addressed by one of the protected characteristics, as only they can claim their rise to the 'ground level'.[49] The other inhabitants of the basement can try to demand their own rise to equality using the same claims, and in this way strengthen the demands of the relatively privileged, but will at least partially be bound to stay in the basement.[50] For example, a black woman will support the fight against singular gender or race discrimination by using the corresponding characteristics to inaccurately address her own experience, yet cannot use these same handles to claim her own rise to a state of non-discrimination.[51]

This risk for reproduction of a social hierarchy by non-discrimination law through the mobilization of a socio-legal privilege remains existent today. Although extremely insightful for the experience of the discriminatee, the law

[43] Schiek, Waddington and Bell (n 8) 171.
[44] Crenshaw (n 41) 149.
[45] ibid.
[46] Anna Carastathis, 'Basements and Intersections' (2013) 28 Hypatia 698, 699.
[47] Crenshaw (n 41) 151.
[48] Carastathis (n 46) 710.
[49] Crenshaw (n 41) 151.
[50] ibid.
[51] ibid.

has yet to adopt the theories of intersectionality.[52] As a consequence, aspects of the discriminatory experience not addressed by the chosen characteristic are still rendered invisible. At the same time, many facets of discrimination that the law ignores are now recorded in the data used by the algorithm. Provided that the transparency of the algorithm is not obstructed by the complexity of the system, the use of algorithms allows, for example, for a more detailed determination of which grounds actually played a role in the result of a discriminatory output.[53] This also means that the intersectional nature of discriminatory practices becomes more visible, placing non-discrimination law's position under increased pressure. Moreover, as the amount of relevant information is increased, one could imagine a situation where the retainment of its single-axis approach could cause non-discrimination law to be inapplicable to any experience of discrimination, as none of the protected characteristics has enough of an impact on the output to amount to an illegal discrimination.

4.2 Protected characteristics

The conclusions of intersectionality are only more troubling when one looks at the narrow set of characteristics that are granted explicit legal protection. Most statutes within the framework of non-discrimination law operate on a limited list of grounds on which the discrimination has to be based in order to be considered illegal. Similar to how non-discrimination law can disadvantage victims of discrimination whose experience is only partially covered by one of the protected characteristics, the lack of recognition in any of the protected characteristics can render a discriminatory experience completely invisible to the law. This is not bizarre, as not every differentiation is a discrimination. Grounds commonly included are race, gender, religion, sexuality, disability and age.[54]

---

[52] Mieke Verloo, 'Multiple Inequalities, Intersectionality and the European Union' (2006) 13 European Journal of Women's Studies 211, 211.

[53] Talia B Gillis and Jann L Spiess, 'Big Data and Discrimination.' (2019) 86 University of Chicago Law Review 459, 474.

[54] American College of Emergency Physicians, 'Non-Discrimination,' (2006) 47 Annals of Emergency Medicine 510; McCrudden and Prechal (n 8) 1–60, 23.

Meanwhile differentiations on, for example, the basis of beauty[55], financial status[56] or vegan preference are currently not deemed discriminatory from the perspective of non-discrimination law.

Which characteristics are included in the list reveals to a certain degree the ruling definition of discrimination within the legal regime at hand. While many statutes share a considerable amount of characteristics, various theories exist on the rightful basis of the inclusion of these attributes. A popular foundation for legislators' reasoning of a list of protected characteristics is the idea that the grounds for illegal discrimination should track existing social categories worthy of protection.[57] This still leaves enormous room for discrepancy between legal frameworks, as concepts such as 'social category' or 'social group' are rather open and dynamic.[58] It can, for example, be debated what degree of saliency is required of the social group[59], whether its members must have experienced a form of subordination due to a power balance,[60] or whether the societal context should even play a role at all.[61] To increase the potential diversity, each of these orientations allows for multiple perspectives. A grouping attribute might, for example, be considered to be defining by the broader public while it does not play a significant role in the subject's perspective on its own identity, and vice versa.[62]

Whichever position is adopted with regard to the defining determinant for rewarding legal protection to characteristics, this choice will increasingly have to

---

[55] William R Corbett, 'Hotness Discrimination: Appearance Discrimination as a Mirror for Reflecting on the Body of Employment-Discrimination Law' (2011) 60 Catholic University Law Review 615.

[56] Frederik Zuiderveen Borgesius, 'Discrimination, Artificial Intelligence, and Algorithmic Decision-Making' (Council of Europe 2018) 35.

[57] Natalie Stoljar, 'Discrimination and Intersectionality' in Kasper Lippert-Rasmussen (ed), *The Routledge Handbook of the Ethics of Discrimination* (Routledge 2018) 72–78.

[58] ibid 68.

[59] ibid.

[60] Patrick Shin, 'Discrimination and Race' in Lippert-Rasmussen (n 57) 203.

[61] Deborah Hellman, 'Discrimination and social meaning' in Lippert-Rasmussen (n 57) 97.

[62] Tal Zarsky, 'An Analytic Challenge: Discrimination Theory in the Age of Predictive Analytics' (2017–18) 14 I/S: A Journal of Law and Policy for the Information Society 11, 16.

be justified as our image of the unfair differentiations made in society becomes more clear. Most of the positions discussed above are at least partly based on a moral appreciation of what *ought* to be, or more fitting what *should not* be. The relevance of reality for the adopted theory for the selection of a particular set of characteristics is, however, not to be underestimated. For example, when legislators embrace the idea that people suffer discrimination as a member of an identifiable social group, it can be argued that they should be aware of the social tags that dominate daily life.[63] Therefore, as long as the adopted theory is based on contingent social factors rather than purely on preconceived moral notions, it can be expected that when confronted with reality the protected characteristics indeed appear to be relevant, at least in the context of the chosen theory for legal protection. Now the use of algorithmic systems increasingly reveals the relevance of non-protected characteristics for unfair outcomes, the chosen theories are increasingly tested on their justification for the inclusion of only a few characteristics.

4.3 Bad actor frame

Ultimately, non-discrimination law's inability to address or perceive the 'full picture' of discrimination may be criticized in reference to the law's focus on the misaligned conduct of individual perpetrators.[64] With the neutralization of the actions of perpetrators as its main concern, the law seems to ignore important systemic and social issues.[65] From this perspective, discrimination is namely seen as being caused by atomistic, discrete events that operate outside a social fabric or historical continuity.[66] Important structural aspects may thus be overlooked as the discrimination is viewed as a particular wrongdoing rather than a social phenomenon.[67] This individualistic approach is most visible when non-

[63] Stoljar (n 57) 72, 78.
[64] Hoffmann (n 39) 904.
[65] ibid; Alan David Freeman, 'Legitimizing Racial Discrimination Through Antidiscrimination Law: A Critical Review of Supreme Court Doctrine' (1978) 62 Minnesota Law Review 1049, 1049.
[66] Neil Gotanda, 'A Critique of "Our Constitution is Color-Blind"' (1991) 44 Stanford Law Review 1, 44.
[67] Freeman (n 65) 1054.

discrimination law openly revolves around intent and a narrow conception of causation, as the requirement of such a fault easily reveals the hunt for 'blameworthy' perpetrators.[68] De-emphasizing these aspects, however, for example, through the incorporation of disparate impact or unintentional discrimination, does not seem to widen the law's gaze too much as the focus remains firmly on discrete sources.[69]

Much of non-discrimination law's discrete source mentality can be traced back to the core mechanism of its design. Alan David Freeman, for example, explains how the core concept of 'violation' leads to such a narrow view on discrimination by inherently siding with the perspective of the perpetrator.[70] He points out that discrimination could instead be approached from the perspective of the victim.[71] From this perspective, discrimination describes the conditions of social existence as a member of the particular group (e.g. employment, housing, education, the psychological effects of being perceived as a member of a group rather than as an individual, etc.). Here, the eradication of discrimination would imply the detection of all the contributing conditions associated with discrimination and consequently their elimination.[72] From the perspective of the perpetrator, however, discrimination is conceived purely as the actions inflicted on the victim by that perpetrator.[73] Therefore, the remedy does not involve an overall improvement of the conditions of the victim's life, but instead limits itself to the neutralization of the misaligned conduct.[74] It is on this basis that Freeman claims that by limiting its remedy to the 'violation' by the perpetrator, the law is hopelessly indifferent to the social, systemic context of discrimination as is reflected in the condition of the victim.[75] And let it be exactly the latter that is to be found in the data upon inspection of e.g. the features of proxy discrimination.

---

[68] Hoffmann (n 39) 905.
[69] ibid.
[70] Freeman (n 65).
[71] ibid 1053.
[72] ibid.
[73] ibid.
[74] ibid.
[75] ibid 1054.

5. Legal disruption

The presence of proxies of protected characteristics in datasets shows itself to be ambiguous. On the one hand, their presence in the data contributes significantly to the obstinacy of discriminatory outputs in algorithmic systems and prevents designers from finding a simple solution for these events. On the other hand, they allow for a more accurate mapping of inequality and discrimination itself. While the latter can be helpful in the fight against discrimination, it could also lead to uncomfortable conclusions for the law. Ultimately, supported by legal literature critical of modern non-discrimination law, the conception of discrimination implied by the reflections in the data appears to be incompatible with the current limits of non-discrimination law. This paper argues that although a confrontation with the image of discrimination evoked by proxies is inevitable for the law (5.1) and this potentially could be disruptive on a fundamental level (5.2) an actual legal disruption depends on the degree to which the law is willing to look itself in the mirror (5.3). In any case, the possible disruption resulting from non-discrimination law's confrontation with the reality reflected in the data is, notwithstanding its many technological requirements, not to be regarded as technological in nature (5.4).

5.1 An inevitable confrontation

An encounter with the 'reality of discrimination' seems unavoidable in the judicial examination of proxy discrimination. The application of non-discrimination law in cases of algorithmic discrimination is flawed in many ways, causing many to contemplate the optimal route to be taken in this context.[76] At all events, however, it has to be proven that the algorithmic system indeed is or is not discriminatory.[77] Whether this is established under direct or indirect discrimination, the confrontation with the implications of the features of proxy discrimination is bound to occur in the subsequent assessment of the justification of differentiation. After all, most non-discrimination statutes deem a differentiation as justified when i.a. requirements of necessity and

[76] See eg Barocas and Selbst (n 1); Hacker (n 18); Xenidis and Senden (n 7).
[77] Xenidis and Senden (n 7) 21.

proportionality are fulfilled.[78] In the case of algorithmic discrimination these tests are likely to come down to the evaluation of the trade-off between efficiency and non-discrimination made by the developers of the system.[79] In such instances, judges are not only confronted with the ground truth of the inequal distribution of goods, skills, etc., as implied by the trade-off itself, but they also have to interact with the second feature of proxy discrimination discussed in this paper. After all, in order to evaluate a particular balance in the exclusion and preservation of proxies, one should at least have a superficial idea of the amount of proxies present in the data. As a result, the adjudicating body is forced to follow the 'lines of proxies' to a point where their inconvenient truth can no longer be avoided.

5.2 A fundamental disruption

Once observed by a court in its analysis of a case of supposed algorithmic discrimination, the unveiled 'reality of discrimination' shows itself to be disruptive for non-discrimination law. The discriminatory proxies found in the data reflect an image of discrimination which is incompatible with the conception currently held by the law. Supported by legal theory critical of the current demarcations of non-discrimination law, the features of proxy discrimination imply the necessity to consider i.a. contextual, structural and systemic aspects of discrimination, and overall demand a broader and more nuanced approach to events of illegal differentiation.[80] Reminded that data merely reflects the reality it applies to, the discrepancy between law and what is mirrored quickly leads to an alarming conclusion: non-discrimination law is in

[78] See eg Council Directive 2000/43/EC of 29 June 2000 implementing the principle of equal treatment between persons irrespective of racial or ethnic origin [2000] OJ L180/22, art 2(b); Council Directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services [2004] OJ L373/37, art 2(b); Directive of the European Parliament and of the Council 2006/54/EC of 5 July 2006 on the implementation of the principle of equal opportunity and equal treatment of men and women in matters of employment and occupation [2006] OJ L204/23, art 2(1)(b).

[79] Xenidis and Senden (n 7) 22.

[80] See supra part 3.

its current form incapable of fully considering discrimination as it presents itself in reality. This deficiency may hamper the doctrine's effectiveness for bringing about attempted positive change.[81] At least, this is the case if one agrees with the popular opinion in academic literature that non-discrimination law finds its goal in directing 'social change to eliminate group-based status inequalities'.[82]

The fundamental nature of this disruption follows from the adaptations to non-discrimination law required to accommodate the image in the mirror. Unfortunately, it is unlikely that it would suffice to simply broaden the scope of non-discrimination law. The law namely not only overlooks relevant social demarcations, intragroup differences or overall noteworthy experiences of discrimination, but it is also blind to the structural and systemic origins of many exclusionary practises.[83] A blindness that finds its significance in the unobstructed, if not re-entrenched continuation of these structures. An attempt to integrate these realisations in the law arguably implies a great intervention in its construction and its approach to discrimination. Take for example the critique that the law wrongly focusses on the misaligned conduct of faulty perpetrators, as illustrated by Freeman.[84] As this is a commentary on the law's core approach to discrimination, adapting its gaze to this conclusion would be fundamentally disruptive for the law's current shape and limits. The reflections of the data could thus not only require a calibration of non-discrimination law to the projected reality, they could also force it back to the drawing board.

---

[81] Hoffmann (n 39) 901.

[82] For discussions of the so-called antisubordination theory, see Ruth Colker, 'Anti-Subordination above All: Sex, Race, and Equal Protection' (1986) 61 New York University Law Review 1003; Kenneth Karst, 'Why Equality Matters' (1982) 48 Sibley Lecture Series. <https://digitalcommons.law.uga.edu/lectures_pre_arch_lectures_sibley/48/> accessed 15 September 2020; Jack M Balkin and Reva B Siegel, 'The American Civil Rights Tradition: Anticlassification or Antisubordination' (2003) 58 University of Miami Law Review 9; Abigail Nurse, 'Anti-Subordination in the Equal Protection Clause: A Case Study' (2014) 89 New York University Law Review 293; Cass Robert Sunstein, 'The Anticaste Principle' (1994) 92 Michigan Law Review 2410; Samuel R Bagenstos, 'The Structural Turn and the Limits of Antidiscrimination Law' (2006) 94 California Law Review 1.

[83] See supra part 4.

[84] Freeman (n 65).

5.3 Forced self-questioning

Nevertheless, any claim of a legal disruption by proxy discrimination ought to be nuanced by non-discrimination law's own influence on this matter. After all, proxies do not create an immediate obstacle for the application of non-discrimination law, however inescapable or infinite their presence may be. Instead, the disruption stems from the law's confrontation with an awkward image of reality. This image is, however, far from new.[85] The legal critiques mentioned in this paper have been well-established for decades, and undoubtedly must have come to all actors of the law's awareness.[86] Thus, nothing stops the law from continuing this alleged ignorance as before, regardless of the negative implications this may have for the eradication of discrimination. In the end, the law holds a factual monopoly on the decision of what it regards as discriminatory, and could turn a blind eye for the mere reason it does not wish to be disrupted. Furthermore, it has been addressed by others how a more structural approach might demand too much from non-discrimination law, and rather belongs to 'the realm of politics and social change…than to the narrow confines of legal doctrine'.[87] However strikingly diagnostic data's mirroring of reality thus may be, its image only proves to be disruptive where the law allows it to be.

However, the reflections in the data already make a compelling case for the law to embrace their implications. After all, to the degree that the law strives to base itself on the reality it tries to bring order to, it can be highly discreditable to disregard the reality which it is constantly confronted with in its application to e.g. proxy discrimination. Furthermore, the mirrored reality shows the law more than simply the structures and mechanisms of a socially stratified world. It may also confront the law with its own role in the continuation of inequality. Law's blindness to the reality of discrimination does namely not only allow discrimination and inequality to proceed at the same pace, but it can also

---

[85] Mayson (n 20) (who argues that algorithms merely shine a new light on the old problem of racial inequality in risk assessment).

[86] See eg Bagenstos (n 82) (describing a 'structural turn' in academic literature); Verloo (n 52) (documenting a growing body of studies and comments on multiple discrimination and intersectionality).

[87] Bagenstos (n 82) 45.

reproduce, entrench and exacerbate the disadvantage present in society.[88] Finally, the law inevitably observes the reflected image in the data, regardless of whether it later chooses to ignore it. It is thereby wise to make use of the diagnostic capabilities this 'clear mirror' offers, rather than to blindly proceed relying on the 'cloudy mirror' that is inherent to human decision.[89]

5.4 A non-technological disruption

The disruption faced by non-discrimination law as a result of its confrontation with the reflections of reality luring in the data, is not easily situated within the existing literature on legal disruption by technology. First of all, it can be questioned whether technology is directly responsible for the disruption discussed in this paper. The vast amounts of data, the computational power, the assistance of artificial intelligence etc. are of course necessary for the reflections to be shown to non-discrimination law in this particular way. Their role is, however, merely facilitative with regards to the disruption. Contrary to many other discussions concerning algorithmic discrimination, such as the difficulty of the opacity and complexity of certain algorithms, it are not the technical characteristics of the technology involved that create a difficulty for the application of the law.[90] Instead, the disruption is caused by the message these

[88] Barocas and Selbst (n 1) 674 ('Approached without care, data mining can reproduce existing patterns of discrimination, . . . It can even have the perverse result of exacerbating existing inequalities by suggesting that historically disadvantaged groups actually deserve less favorable treatment.'); Elise Boddie, 'Adaptive Discrimination' (2016) 94 North Carolina Law Review 1235, 1266 ('Time does not inevitably lead to improvement if we misunderstand the problem. In fact, if anything, time can exacerbate the problem if we leave the malady untreated'); Crenshaw (n 41) 151.

[89] See Mayson (n 20) 2224 (who explains that subjective prediction by humans reflects the past similarly to algorithmic prediction. Human prediction is, however, based on less reliable anecdotal data. The precise algorithmic mirror should thus not be discarded for the cloudy one).

[90] On the challenges of opaqueness and inexplainability of algorithms Borgesius (n 56) 34; Danielle Keats Citron and Frank Pasquale, 'The Scored Society: Due Process For Automated Predictions' (2014) 89 Washington Law Review 1; Tal Zarsky, 'The Trouble

technological tools bring, not by the medium by which it is delivered. Secondly, it can similarly be questioned whether a need for change comes from a shift in the sociotechnical landscape. Both the disruption itself, as well as the necessary adaptations it requires from the law to overcome it, can hardly be based on the effects newly enhanced technological capabilities have on people's activities or environment.[91] After all, the image reflected in proxy discriminations has not enlightened our society with a new, changed look on discrimination. The only novelty is that the law is now directly confronted with an old truth it was comfortable ignoring for a long time.

6. Conclusion

Reality is mirrored in the data that is used in the various algorithms that increasingly rule our lives. An intelligent processing of this data allows for a mapping of reality, which is accompanied by an increased understanding of the phenomena observed through the data. This development can disrupt non-discrimination law, as also existing inequalities and discriminations are reflected in the data. Observed through the lens of the many discriminatory proxies that lure in algorithmic systems, a broad notion of discrimination imposes itself on the law. As a result the law faces a potential disruption. Confronted with the reflections in the data, it can no longer ignore the world outside its scope, and thus, experiences an increased pressure to justify its limits. Non-discrimination law's position is only more problematized now that many theories and critical literature regarding the current state of non-discrimination law find basis in the data. The proxies in the data thus hold up a mirror to the law, challenging it to examine itself. While modern technology facilitates the mirrored image, there is nothing technological about the reality it depicts.

---

with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making' (2016) 41 Science, Technology & Human Values 118.

[91] See for a more detailed discussion of sociotechnical change: Lyria Bennett Moses, 'Regulating in the Face of Sociotechnical Change' in Roger Brownsword, Eloise Scotford and Karen Yeung (eds), *The Oxford Handbook of Law, Regulation and Technology* (2017).

Volume 5, Number 1, Spring 2021

AI, Machine Learning, and International Criminal Investigations

The lessons from forensic science

Karen M. Richmond*

The evolving field of machine learning and artificial intelligence is frequently presented as a positively disruptive branch of data science whose expansion allows for improvements in the speed, efficiency, and reliability of decision-making, and whose potential is impacting across diverse zones of human activity.[1] A particular focus for development is within the criminal justice sector, and more particularly the field of international criminal justice, where AI is presented as a means to filter evidence from digital media, to perform visual analyses of satellite data, or to conduct textual analyses of judicial reporting datasets. Nonetheless, for all of its myriad potentials, the deployment of forensic machine learning and AI may also generate seemingly insoluble challenges. The critical discourse attendant upon the expansion of automated decision-making, and its social and legal consequences, resolves around two interpenetrating issues; specifically, algorithmic bias, and algorithmic opacity, the latter phenomenon being the focus of this study. It is posited that the seemingly intractable evidential challenges associated with the introduction of opaque computational machine learning algorithms, though global in nature, are neither novel nor unfamiliar. Indeed, throughout the past decade and across a multitude of jurisdictions, criminal justice systems have been required to respond to the implementation of opaque forensic algorithms, particularly in relation to complex DNA mixture analysis. Therefore, with the objective of highlighting the potential avenues of challenge which may follow from the introduction of forensic AI, this

* Postdoctoral Research Fellow, Copenhagen University [karen.richmond@jur.ku.dk]

[1] Rhiannon Jackson and Maria McAreavey, 'Black-Box Medicine: Protecting Patient Privacy Without Preventing Innovation' (2019) 3(1) Retskraft – Copenhagen Journal of Legal Studies 68.

study focusses on the prior experience of litigating, and regulating, probabilistic genotyping algorithms within the forensic science and criminal justice fields. Crucially, the study proposes that machine learning opacity constitutes an enhanced form of algorithmic opacity. Therefore, the challenges to rational fact-finding generated through the use of probabilistic genotyping software may be encountered anew, and exacerbated, through the introduction of forensic AI. In anticipating these challenges, the paper explores the distinct categories of opacity, and suggests collaborative solutions which may empower contemporary legal academics – and both legal and forensic practitioners – to set more rigorous and usable standards. The paper concludes by considering the ways in which academics, forensic scientists, and legal practitioners, particularly those working in the field of international criminal justice, might re-conceptualise these opaque technologies, opening a new field of critique and analysis. Using findings from case analyses, overarching regulatory guidance, and data drawn from empirical research interviews, this article addresses the validity, transparency, and interpretability problems, leading to a comprehensive assessment of the current challenges facing the introduction of forensic AI. It builds upon work undertaken at the Nuffield Council on Bioethics *Horizon Scanning Workshop: The future of science in crime and security* (5th July 2019, London).

1. Introduction

Technologies, writes Zuboff, 'define the horizon of our material world, as they shape the limit of what is possible and what is barely imaginable.' Their usage connotes neither neutrality nor objectivity, but rather a contingency that is 'brimming with valence and specificity in the opportunities that it creates and forecloses.'[2] Zuboff's definition encapsulates the contemporary challenges generated by the requirement to standardise, and to regulate, novel forms of machine learning (ML), and artificial intelligence (AI), both of which are the subject of sustained attention from academics, and associated institutional

[2] Shoshana Zuboff, 'Automate/Informate: The Two Faces of Intelligent Technology' (1985) 14(2) Organizational Dynamics 5, 5.

agents.[3] Thus, despite its myriad potentials, this emergent field of data science is characterised as inherently disruptive, and capable of presenting novel, and seemingly insoluble, challenges simultaneously across diverse fields. However, a review of the relevant academic literature suggests that researchers have thus far omitted to consider whether a proportion of the seemingly intractable challenges associated with the introduction of AI are as novel and unfamiliar as is frequently perceived. This article therefore addresses the omission, focusing on the forensic science and legal fields, both of which have been at the forefront of scientific development. The study considers the degree (if any), to which the courts' prior experience of standardising, and regulating, forensic algorithms within the criminal justice system, may generate insights which can aid contemporary legal academics and forensic practitioners in their efforts to set more rigorous and practical standards, with respect of this latest wave of 'disruptive' technology.[4]

The objective of the article is to highlight the implications for rational legal fact-finding, and adjudication, pursuant to the implementation of forensic and investigatory forms of AI within the criminal justice field, consequently its introduction to the international criminal courtroom by way of expert opinion evidence.[5] Whilst the potentials of AI are being explored across diverse national jurisdictions, and in heterogeneous fields such as law enforcement, forensic science, and academic research, it is posited that the international criminal justice arena represents a particularly engaging arena of analysis, given that this sector may invite investigation at a scale most suited to the mobilization of AI-driven

[3] For the purposes of this article, Artificial Intelligence is used to denote all forms of machine learning, utilising artificial neural nets (ANNs) and other forms of algorithmic computation. Machine learning thus forms a subset of artificial intelligence, as commonly understood.

[4] Thomas Buocz, 'Artificial Intelligence in Court: Legitimacy Problems of AI Assistance in the Judiciary' (2018) 2(1) Retskraft – Copenhagen Journal of Legal Studies 41.

[5] See, for example, 'Scientists Developing AI to Spot Paedophiles Just From Images of Their Hands' (*The Week*, 28 February 2020) <https://www.theweek.in/news/sci-tech/2020/02/28/Scientists-developing-AI-to-spot-pedophiles-just-from-images-of-their-hands.html> accessed 27 December 2020.

efficiencies.[6] The receptivity of the international criminal justice (ICJ) sector is further enhanced by both responsiveness of the courts, when presented with evidence drawn from 'open source' data,[7] and the relative lack of procedural safeguards, particularly the absence of a gatekeeping mechanism for expert opinion evidence.[8] Thus, it is posited that the concomitant challenges associated with the deployment of AI may prove particularly impactful in the international justice arena. Nonetheless, the instant study demonstrates that such obstacles constitute a mere extension of those first encountered by national courts in relation to the use of algorithmic DNA analysis software.[9] Further, that the global challenges generated by forensic AI may be resolved in a similar fashion to those generated by the introduction of probabilistic genotyping software, through rigorous validation processes guided by overarching guidelines and regulations.

Whilst the introduction of algorithmically-derived evidence has required the mobilization of diverse bodies of expertise, in both common law and civilian jurisdictions, this study focusses on the comparatively developed and rigorous common law jurisprudence encountered in the United States and United Kingdom, in addition to those regulatory responses and guidelines published by

---

[6] Examples include the use of AI to analyse satellite data to detect the destruction of human settlements, Milena Marin, Freddie Kalaitzis and Buffy Price, 'Using Artificial Intelligence to Scale Up Human Rights Research: a Case Study on Darfur' (*Citizen Evidence Lab*, 6 July 2020) <https://citizenevidence.org/2020/07/06/using-artificial-intelligence-to-scale-up-human-rights-research-a-case-study-on-darfur/> accessed 27 December 2020. A further example is the use of AI to filter evidence from large repositories of open source data, Abishek Kumar, 'Digital Evidence and the Use of Artificial Intelligence' (*International Criminal Court Forum*, 31 May 2020) <https://iccforum.com/forum/permalink/122/33560> accessed 27 December 2020.

[7] Lindsay Freeman, 'Digital Evidence and War Crimes Prosecutions: The Impact of Digital Technologies on International Criminal Investigations and Trials' (2018) 41 Fordham Int'l LJ 283, 283–328; Sam Dubberley, Alexa Koenig and Daragh Murray (eds), *Digital Witness: Using Open Source Methods for Human Rights Investigations, Advocacy and Accountability* (Oxford University Press 2020).

[8] See n 49.

[9] Julia Gasston and others, 'An Examination of Aspects of the Probabilistic Genotyping Tool: Forensic Statistical Tool' (2020) 2 WIREs Forensic Science e1362.

the European Network of Forensic Science Institutions (ENFSI),[10] the regulatory guidelines published by the Forensic Science Regulator for England and Wales,[11] and the reports of both the United States' Executive Office of the President's Council of Advisors on Science and Technology,[12] and the House of Lords' Science and Technology Select Committee.[13]

Theoretically, this article founds upon scientific theories of evidence interpretation, specifically the the Rationalist Model of Adjudication, as proposed by John Henry Wigmore, and elaborated by William Twining, its most notable contemporary proponent. According to this model,[14] the direct end of adjectival law is rectitude of decision-making through the correct application of valid law, and the accurate determination of true past facts, proved to specified standards, on the basis of careful and rational weighing of reliable evidence, presented to impartial decision-makers. This rigorous formulation forms the backdrop to a careful review of law's instrumentalisation of DNA mixtures analysis software, in its efforts to present information to the court which is beyond the common experience of the trier-of-fact. The review and analysis thus demonstrate the ways in which the introduction of computer-driven probabilistic genotyping methods in 2010 – despite having initially appeared to resolve issues generated by the increased sensitivity of DNA profiling methods – generated significant juridical challenges related to opacity and methodological validity. To add further depth to the analysis, the study draws on qualitative interview data drawn from a study of the perspectives of forensic

[10] European Network of Forensic Science Institutes, 'Guideline for Evaluative Reporting in Forensic Science' (2015) <http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdfEuropean> accessed 27 December 2020.

[11] Forensic Science Regulator, 'Software Validation For DNA Mixture Interpretation' (FSR-G-223 Issue 2, 2020) <https://www.gov.uk/government/publications/software-validation-for-dna-mixture-interpretation-fsr-g-223> accessed 27 December 2020.

[12] President's Council of Advisors on Science and Technology, 'Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods' (2016) <https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensic_science_report_final.pdf> accessed 27 December 2020.

[13] Science and Technology Select Committee, Forensic Science and the Criminal Justice System: a Blueprint for Change (HL 2017–19, 333).

[14] William Twining, *Rethinking Evidence: Exploratory Essays* (2nd edn, Cambridge University Press 2006) 72.

scientists operating within the forensic market in England and Wales. The article maintains a specific focus on the legal challenges mobilised against evidence derived from probabilistic genotyping (PG) software packages, converging on the two related methodological categories of concern; the absence of acceptable standards of validation (both developmental and internal), and the underlying lack of transparency.[15] The study demonstrates how the courts' growing appreciation of these methodological weaknesses necessitated the introduction of novel procedures, and validation protocols. Further, that in a number of instances probabilistic genotyping evidence derived from opaque algorithmic processes was ruled as wholly inadmissible in criminal trials. In substantive terms, the instant paper thus seeks to demonstrate how, and to what extent, problems traceable to a lack of foundational validity, and a lack of transparency, may re-emerge in a heightened form with the proposed implementation of AI within the forensic field. Further, that such evidential problems may become critical, particularly in relation to the deployment of 'opaque AI', since the program's algorithmic base may be manipulated recursively in order for the AI to learn, develop, and build efficiency and accuracy, through a process of trial-and-error. Crucially, this process of manipulation and change occurs beyond the threshold of human perception and control, obstructing reproducibility. When such technologies are introduced into the forensic sphere, as is currently planned, it is posited that their use may present potentially insoluble evidential problems, given that transparency and interpretability are central procedural and legal requirements, necessary in order to establish the validity of novel technologies, and expert opinions, within the courtroom.

2. Opacity

Computational algorithms are now harnessed across all sectors of human endeavour. Their capacity for efficient discrimination, and classification, has enabled them to proliferate in an environment rich in personal and trace data. Algorithms may play either a central or peripheral role, acting singly, or jointly

[15] See, for example, *Commonwealth v Foley* 38 A 3d 882, 2012 Pa Super 31 (Pa Super Ct 2012).

with other algorithms. They enable routine tasks to be performed efficiently, and serve as the engine for mundane data management tasks such as filtering 'spam' and performing internet searches. Algorithms also assume socially consequential roles, where their predictive capacities enable them to make onerous decisions, such as on an applicant's suitability for employment, or ability to repay a loan. In their most advanced iterations, computational algorithms form the cognitive drivers for machine learning systems, as utilized in facial recognition programs, or the autonomous AI of self-driving cars. So too are they deployed throughout the criminal justice sector, where the ability to make accurate categorisations is at a premium. The discriminatory capacities of computational algorithms thus form the basis for a number of forensic technologies, all of which converge around biometric discrimination. The US Government Accountability Office reports that,

> Federal law enforcement agencies … are primarily using three types of forensic algorithms to help assess whether or not evidence collected in a criminal investigation may have originated from an individual: probabilistic genotyping, latent print analysis, and face recognition.[16]

Nonetheless, the harnessing of these technologies has not been unproblematic. Concerns have arisen regarding the potential for algorithms and machine learning systems to exhibit 'algorithmic bias', or to entrench socio-economic and racial inequalities.[17] These analyses view algorithmic decision-making as a distillation of human decision-making. As such, the influence of social inequalities and biases which afflict human decision-making translate to – and are visibly encoded within – the algorithmic system, mediating its outputs. Similar concerns have similarly been raised around the propensity for algorithmic systems to exhibit behaviours which display significant deficiencies

---

[16] See United States Government Accountability Office, 'Forensic Technology: Algorithms Used in Federal Law Enforcement' (GAO-20-479SP, 12 May 2020) <https://www.gao.gov/assets/710/706849.pdf> accessed 28 December 2020.

[17] See Alexander Babuta, Marion Oswald and Christine Rinik, 'Machine Learning Algorithms and Police Decision-Making: Legal, Ethical and Regulatory Challenges' (RUSI Whitehall Report 3-18, Royal United Services Institute, September 2018).

with regard to discernibility, predictability, and tractability. These crystallise around the concept of 'algorithmic opacity.' As defined by Burrell, algorithms are opaque to the extent that '…if one is a recipient of the output of the algorithm (the classification decision), rarely does one have any concrete sense of how or why a particular classification has been arrived at from inputs.'[18] In terms of rational adjudication, these phenomena are not thereby consonant with the requirement for efficacy and reliability in relation to expert opinion evidence.

Furthermore, those algorithmic inputs may themselves be opaque, or undefined, particularly in relation to that subset of machine learning systems which manipulate their own algorithmic substructure. Opacity is thus often contraposed with the concept of 'algorithmic transparency,' and with calls for the introduction of non-proprietary 'open source' systems. These epistemological issues assume a particular significance within the field of forensic science, and criminal justice, where the 'black-boxing' of algorithmic classifications may require the trier-of-fact to accept expert assertions, absent of meaningful examination and evaluation, whilst simultaneously concealing problems relating to the foundational validity of novel scientific methods. As will be posited in the critique and analysis below, to the extent that these problems remain unaddressed, they threaten to disrupt, or subvert, fundamental principles of the law of evidence, the *ipse dixit* rule, and the overarching right to a fair trial. However, the concept of algorithmic opacity first requires elaboration, alongside an illustrative elaboration of algorithmic typology and mathematical design since, as Burrell contends, 'recognising distinct forms of opacity…is key to determining which of a variety of technical and non-technical solutions could help to prevent harm.'[19] Therefore, in the following section, discussion turns to Burrell's tripartite classification of algorithmic opacity, placing the diverse forms in a rationalist adjudicatory context.

The first category of opacity encountered is 'intentional opacity', designed into the system as a form of proprietary protection, thus intended to help maintain a market position within a competitive field, and to better enable the developer to protect 'trade secrets.' This primary variant of intentional opacity

[18] Jenna Burrell, 'How the Machine 'Thinks:' Understanding Opacity in Machine Learning Algorithms' (2016) 3 Big Data & Society, 1.
[19] ibid.

has been encountered within marketised segments of the criminal justice sector, and occupies a long-standing area of contention in criminal litigation, particularly in relation to privately developed probabilistic genotyping algorithms. A variant of intentional opacity comprises those covert forms designed to conceal the internal logics of computational algorithms, and deployed as a means to obscure 'sidestepped regulations, the manipulation of consumers, and/or patterns of discrimination.'[20] The deliberate 'black-boxing' of decision-making processes, for commercial interests, militates not only against the rationalist approach to adjudication, and the need for transparency in matters of logical inference: such obfuscation also impacts significantly on the rights of the accused and upon the principle of the equality of arms, the preservation of the latter being paramount wherever technical solutions are deployed in answer to evidentiary challenges.[21] However, the foregoing instances of 'remediable incomprehensibility'- it will be suggested – may be remedied, by the implementation of 'open source' forensic systems even if, as will be demonstrated *infra*, such a solution may offer only partial mitigation.

The secondary variant of algorithmic opacity is 'technical opacity', generated as a by-product of the high degree of specialisation and technical expertise required to design integrated computational systems. The ability to read, and write, computer code clearly requires advanced literacy in programming languages alongside a familiarity with software engineering. Translated to either the national, or international, criminal justice system, it is questionable to what extent many defence practitioners may routinely marshal the necessary skills. Proactive examples will be cited of efforts to reverse-engineer proprietary probabilistic genotyping algorithms using expert programming analysts. However these are the exception, and it is debatable to what degree such expertise is diffused across the criminal justice system. The corollary of the foregoing discussion is that the absence of diffuse expertise may potentially limit

[20] ibid 4.

[21] The Grand Chamber of the ECtHR summarized the principle of 'equality of arms' in *Edwards and Lewis v United Kingdom* (2005) 40 EHRR 24: 'It is in any event a fundamental aspect of the right to a fair trial that criminal proceedings, including the elements of such proceedings which relate to procedure, should be adversarial and that there should be equality of arms between the prosecution and defence'.

the mitigating influence of open-source solutions.[22] A more comprehensive solution may therefore be reached through transparent validation processes or, in the case of commercial suppliers, the commissioning of an independent and confidential review by an external expert.[23] The establishment of foundational validity, or failure thereof, should be the central criterion for courts to determine the reliability of expert scientific opinion, consonant with the need for rectitude of decision-making.[24]

The third variant of algorithmic opacity is 'inherent opacity', which appears as a function of the internal features and operational dynamics of algorithmic systems. It may be otiose to highlight that a number of machine learning systems operate at a scale, and a level of complexity, which renders their overall operations opaque even to those who design the discrete components incorporated within the system.[25] However, it is not the scalar element of machine learning and AI systems which generates the greatest challenges to evidential transparency. Whilst an inability to effectively limn the contours of multi-component systems presents significant obstacles to achieving 'equality of arms', the greatest challenge to tractability derives from the fundamental divergence of human, and machine, logics. Thus, the following critique and analysis must attempt to distinguish between distinct classes of algorithms, and the forms of machine logic particular to each. The first illustration focusses on a visual recognition task using a neural network. The computational algorithms used to perform these 'pattern-matching' tasks display a degree of mimesis with a human neural network, such that a number of input nodes are linked to a central set of nodes called the 'hidden layer', thence to a corresponding set of output nodes. The lines connecting the nodes are ascribed a quantitative value (or weight), and – through a rapid process of trial and error – the machine learns the optimal value for the conjoined matrix of linear weights.

---

[22] Burrell (n 18) 4.

[23] Forensic Science Regulator (n 11) 26.

[24] See, for example, the US Supreme Court Rule 702 (as amended); the English *Criminal Practice Directions 2015* [2015] EWCA Crim 1567, [2015] All ER (D) 134 (Sep), Rule 19A.5.

[25] The prime example is the 'Google' search engine.

However, when set a simple practical task, such as recognizing handwritten numerals, the most salient feature is the marked difference between the dynamics of machine logic and the ways in which human actors might disaggregate the task into a set of intelligible sub-tasks. This fundamental incommensurability between the logic of the 'hidden layer', and human cognition, 'arises from the very notion of computational 'learning.' Machine learning is applied to the sorts of problems for which encoding an explicit logic of decision-making functions very poorly.'[26] Indeed, whilst basic algorithms must be written in a way that is understandable, and logically explicable to those whose task is to develop or maintain the system, the step to machine learning may collapses that division, since the inherent feature of advanced ML and AI is the ability of learning systems to manipulate their algorithmic base. The challenges to transparency are further compounded by a secondary learning process known as back-propagation: '[back-propagation] tweaks the calculations of individual neurons in a way that lets the network learn to produce a desired output.'[27] Clearly, for the ML system, or autonomous AI, explicability – or even intelligibility to human actors – is not a concern. And it is relatively straightforward to discern the central problem: while overarching efficiencies of machine learning may be readily transposed to the criminal justice system, and in particular the forensic science field, it is clear that the inherent opacity of those machine logics may begin to generate unassailable explanatory barriers when implemented in an investigative, classificatory, or evaluative capacity.

Burrell cites a second example of the inherent opacity of machine learning systems, in this instance programs tasked with filtering 'spam' messages.[28] This model utilises algorithmic modules known as Support Vector Machines (SVMs) in order to differentiate 'spam' messages from 'non-spam', through a linear regression process. The training module learns a set of words and ascribes a weighting to each. Once again, however, it is the incommensurability of the machine logic, when performing these protocols, which generates inherent

[26] ibid 6.
[27] Will Knight, 'The Dark Secret at the Heart of AI' (*MIT Technology Review*, 11 April 2017)<https://www.technologyreview.com/2017/04/11/5113/the-dark-secret-at-the-heart-of-ai/> accessed 28 December 2020.
[28] Burrell (n 18) 7.

opacity and diverges from human norms, since the computational algorithm is blind to any natural semiotic configuration between words, phrases, and narratives. Further, the ML does not attempt to reason with regard to the presence or absence of certain words, but rather aggregates the weightings associated with all of the words contained in a given sample. This relatively simple example once more demonstrates the counterintuitive nature of machine logic, whose inherent opacity may impact not only on our ability to explain classifications when applied to practical tasks within the legal and forensic fields, but potentially circumscribe legal and forensic research based upon discourse, and narrative, analyses. These challenges increase exponentially when opaque algorithms are incorporated into a multi-dimensional model working across a multitude of features. It is posited that Burrell's tripartite classification, as developed above, serves as a useful typology with which to analyse specific extensions of algorithmic computation, particularly the use of machine learning and AI in international criminal investigations. However, discussion first turns to the use of proprietary forensic DNA profiling algorithms, and the challenges which these generated, in order to discern ascertain whether the solutions arrived at by the courts – and allied institutional agents – may offer practical insights, whose application might reduce those risks associated with the use of opaque ML and AI systems.

3. DNA Profiling and the Criminal Justice System

The criminal justice system has been one of the foremost sectors willing to embrace the efficiencies of algorithmic and machine learning classification. Indeed, the forensic science field has, for the past decade, been at the forefront of testing and adapting innovative methods, in an effort to harness the discriminatory potentials of automated computation. One area of rapid development involves the automated interpretation and evaluation of complex DNA profiles, including DNA mixtures, degraded DNA, and trace samples. This contentious area has generated a body of criminal litigation and a rich seam of academic comment. It is posited that the creative tensions between the legal and forensic science fields, which emerged in relation to the issue of probabilistic genotyping, form a cogent base for further discussion regarding algorithmic

opacity, and the potentials of forensic AI, and machine learning. However, before proceeding with this wider critique it is first necessary to establish the underlying conceptual foundations relative to DNA profiling and analysis.

It is generally accepted that the palette of forensic techniques which together go under the term 'forensic science' do not all enjoy equal merit, exhibit similar levels of foundational validity, or are accorded comparative scientific status. Of all of these techniques – ballistics, fingerprinting, and the like – DNA profiling alone has been accorded the epistemic status of research science, a standing acknowledged by forensic scientists, academic commentators,[29] and members of the public alike.[30] Indeed, the US National Academy of Science (NAS) committee, when delineating the ambit of their 2009 study, and explaining the absence of DNA profiling within their review, noted that forensic DNA had previously been subject to two landmark studies, which had settled 'the DNA wars' and had firmly established the pedigree of forensic DNA profiling.[31] As Murphy observes, running counter to the ascendancy of DNA profiling,

---

[29] A review of the literature demonstrates that, beyond the core-set of forensic-scientific practitioners (and associated institutional actors), DNA-profiling techniques have been accorded an exceptional – if not unassailable – epistemological status. Evidence derived from DNA-profiling has been described by defence lawyers as 'infallible', or as furnishing 'irrefutable proof' [see Barry C Scheck, 'Preventing the Execution of the Innocent: Testimony Before the Senate Judiciary Committee' (2001) 29 Hofstra Law Review 1165]; by judges as a 'truth machine', or 'revelation machine' [Helena Machado and Rafaela Granja, 'Police Epistemic Culture and Boundary Work with Judicial Authorities and Forensic Scientists: the Case of Transnational DNA Data Exchange in the EU' (2019) 38 New Genetics and Society 289]; and by a prison inmate as 'God's signature'; [Michael Lynch, 'God's Signature: DNA Profiling, the New Gold Standard in Forensic Science' (2003) 27 Endeavour 93]. Such epistemic exceptionalism is not uncommon amongst the academic literature, and associated publications, devoted to forensic DNA profiling.

[30] The epistemological privileging of knowledge claims derived from such techniques is not limited to the claims of institutional actors. Prison inmate Loyd, E-J., is quoted as stating that 'DNA – deoxyribonucleic acid – is God's signature. God's signature is never a forgery.' See Jodi Wilgorin, 'Confession Had His Signature; DNA Did Not' *New York Times* (New York, 26 August 2002) A 1.

[31] Committee on Identifying the Needs of the Forensic Sciences Community, 'Strengthening Forensic Science in the United States: A Path Forward' (National Academy of Sciences 2009).

> … the traditional forensic disciplines that had long served as the backbone of scientific evidence in the courtroom, and continued to make up the majority of the scientific evidence in criminal cases, went largely ignored despite loud pleas from a dedicated coterie within the scholarly and scientific community.[32]

Thus, forensic DNA was presented as the paradigm forensic technique, uniquely scientific, the benchmark forensic science discipline, and the purpose of the NAS report was therefore to provide the groundwork for the residuary categories of forensic techniques to meet the scientific standards set by DNA, in order that they might establish similarly robust epistemic credentials. Murphy rightly highlights the difference between 'first generation' pattern-matching techniques, and 'second generation' bio-identification sciences, and sheds light on the way in which DNA became to be regarded as a '*sine qua non*'. With regard to single source DNA, this is a convincing analysis. However, when probabilistic genotyping of mixed samples is factored into this analysis, the picture changes. Absent from Murphy's critique as presented here (though the subject of trenchant analysis throughout her work) is the conception that DNA may itself be fallible, affected by technological developments, or influenced by alterations to overarching governance structures. Indeed, it is necessary to stress that later iterations of DNA profiling techniques must continue to establish a basic foundational validity which meets legal standards and the overarching objectives of the NAS Report.

4. Mixtures and Low Template DNA

At this stage, it should be re-iterated that the basic DNA profiling protocols, on which the above perceptions are based, had been subject to thorough validation and accreditation procedures, and had established reliable scientific underpinnings. In contrast, even though pattern-matching techniques present their conclusions in terms of a 'match/non-match', such unique categorisations

[32] Erin Murphy, 'What "Strengthening Forensic Science" Today Means for Tomorrow: DNA Exceptionalism and the 2009 NAS Report' (2010) 9 Law, Probability and Risk 7.

lack a scientific basis, being non-probabilistic, open to significant bias, and unable to articulate established error rates. The reason that 'single-source and simple-mixture sample analyses are considered highly reliable [is] because each of the steps involved in the analysis is 'repeatable, reproducible, and accurate.' This trio of requirements is referred to as 'foundational validity.'[33] However, the same foundational validity, based on a high degree of trust in the accuracy of results, is neither exhibited by first generation techniques, nor capable of extension to more complex processes, such as those involving minute traces of 'low template' DNA, or degraded DNA, especially where these involve the interpretation of 'DNA mixtures' drawn from a number of individuals.

The occurrence of DNA mixtures has risen sharply since the introduction of sensitive testing protocols (such as DNA-17 and Globafiler-24, both of which replaced the less sensitive SGM Plus system).[34] These protocols are now capable of picking up trace amounts of 'low template' DNA, their use leading to the routine reporting of mixed DNA profiles. Complex mixtures undergo the same forms of processing as simple, or single-source DNA samples. In short, the sample is stabilised, and amplified. Scientists then use standardised procedures to count the numbers of Short Tandem Repeats (STRs are polymorphisms, or areas which exhibit a high degree of variation) at a number of loci, or sites, on the DNA. A graphical output displays each loci as a peak whose height is a product of the number of STRs at that site. Together these peaks create a DNA profile which can be rendered numerically, for statistical analysis against background population data.

However, in the case of DNA mixtures, these require deconvolution, and the interpretation of the results may display significant levels of variation, not least as the set of superimposed peaks require to be carefully evaluated in order to

[33] Katherine Kwong, 'The Algorithm Says You Did It: The Use of Black Box Algorithms to Analyze Complex DNA Evidence' (2017) 31 Harv JL & Tech 275, 277.

[34] See, for example, Matthew J Ludeman and others, 'Developmental Validation of GlobalFiler™ PCR Amplification Kit: A 6-Dye Multiplex Assay Designed for Amplification of Casework Samples' (2018) 132 International Journal of Legal Medicine 1555.

determine whether a suspect profile is included.[35] This can be achieved manually, and mathematically. Alternatively, probabilistic genotyping (PG) programs may be utilised. These computerised mathematical models and simulations estimate the likelihood that a particular individual's DNA is part of the mixture present in the sample. Although the preponderance of PG systems (and subsequent cases cited) emanate from the United States, it should be noted that the issues raised affect forensic practice in a multitude of jurisdictions. For example, empirical research in the UK revealed similar concerns regarding the use of probabilistic genotyping algorithms to de-convolute mixed DNA profiles as those raised in the literature, particularly with regard to validation.

> There are two different types. Cellmark uses David Balding's [open source LikeLTD] system. LGC developed LiRA. These systems can deal with two or more people, though for a while Balding's system wasn't validated – it is now. There are differences between the systems but the same system can deliver different answers depending on how the question is formed.
>
> (Interview with Lead Scientist: Oxford, 2015)

This typical response (drawn from 33 semi-structured interviews with DNA profiling scientists and allied institutional agents), supports the claim of levels of scepticism amongst groups of experts with regard to the scientific validity and operational dynamics of algorithmic forms of probabilistic genotyping. Such scepticism also focusses on the need to establish foundational validity within the courtroom. Further, to ensure that the operator inputs – including the framing of propositions – are explicitly noted in order to facilitate transparency and reproducibility.[36] The following section elaborates on these concerns, analysing

---

[35] See Rich Press, 'DNA Mixtures: A Forensic Science Explainer' (National Institute of Standards and Technology, 3 April 2019) <https://www.nist.gov/featured-stories/dna-mixtures-forensic-science-explainer> accessed 28 December 2020.

[36] Whilst a variation in output consequent to a variation in input is hardly problematic, within the forensic and legal context, the propositions on which probabilistic

the use of PG software in the courtroom with reference to a number of case studies, and utilising Burrell's tripartite classification in order to discern the forms of opacity encountered therein. It goes on to evaluate the implications of the generation of particular forms of opacity for the exercise of rational fact-finding and legal adjudication.

5. Probabilistic Genotyping Software Case Studies

The first example of forensic-algorithmic opacity focusses on the use of a probabilistic genotyping package known as the Forensic Statistical Tool (FST). This software system was developed by the New York City Office of the Chief Medical Examiner (OCME). Introduced in 2010, the OCME began to routinely use the FST in tandem with high sensitivity testing (HST) in cases which involved mixed, trace, and/or degraded, samples. Indeed, the laboratory stated that it had used High Sensitivity Testing (HST) in 3450 cases between 2006, and 2017. Further, that it had used the Forensic Statistical Tool in 1350 cases between 2011 and 2017. However, for nearly six years, between 2010 and 2016, defense requests to conduct independent expert witness reviews of this in-house proprietary software (including the source code, supporting development material, and executable software versions) were denied, even where the request involved an audit under protective order. When, in 2016, the source code was first reviewed, several problems were encountered, not least a previously undisclosed data-dropping function which discarded evidence of potential value to the defence. In later studies, which focused on the quantitative impact of the undisclosed function on the original validation data of 439 samples, it was found that the data-drop was triggered in 23.7% of cases (104 samples). The overall

calculations proceed must be addressed carefully in order to elicit an accurate answer to the particular question which is being asked in relation to the evidence eg whether the DNA sample was deposited by a particular source, as opposed to through a particular activity. That process must meet the same requirement for transparency as that pertaining to the calculation itself. See Forensic Science Regulator (n 11) 17.

effect was 'to skew results towards false inclusion for individuals whose DNA was not present in the evidence sample.'[37]

A landmark case involving the FST followed an assault on an individual in Brooklyn, New York, in 2013.[38] In the wake of a brawl in a Hasidic Jewish district,[39] during which an African American male was seriously injured by a number of assailants, a shoe was recovered, and sent to the NYC Medical Examiner's office for testing. When an area of the shoe was swabbed, a mixed DNA sample from two individuals was recovered. The sample size was 97.9 picograms, which was below the lower limit for standard DNA processing (100pg).[40] Therefore the sample was also subjected to high-sensitivity testing (HST), which extrapolated the size of the sample by reproducing it. Ordinarily samples underwent 28 cycles of amplification. However, HST samples underwent 31 cycles. This boosted the sample size but also served to amplify any latent defects and artefacts. The resulting sample was then subjected to probabilistic genotyping, analysed using the FST. The OCME stated that the two-person mixture contained both the victim's DNA, and that of the accused, with an attendant probabilistic determination of 133 to 1. The accused was convicted but the verdict was overturned on appeal, the evidence from the FST being described as 'less than convincing.' The reasoning was based on the OCME's combining two testing methods which both lacked foundational validity. Further, the unsuitability of the FST calculations when applied to a suspect drawn from a genetically homogenous population. Thirdly, due to the fact that the technician had altered the testing parameters. For the purposes if the instant study, it should be noted that throughout this case the OCME

---

[37] Jeanna Matthews and others, 'The Right to Confront Your Accusers: Opening the Black Box of Forensic DNA Software' in American Association for Artificial Intelligence and Association for Computing Machinery, *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (United States Association for Computing Machinery 2019) 321.

[38] *People v Herskovic* 2018 NY Slip Op 06763.

[39] The ethnicity of the victim and accused is an important consideration, when attempting to derive a statistical output from a DNA profile measured against a population database.

[40] A picogram (pg) is one trillionth of a gram, or 0.000000000000001 kilogram (SI unit).

vigorously opposed examination of its FST source code. Nonetheless, a comprehensive code audit was later conducted, which unearthed significant problematic features.

The cases involving the FST, and the opaque features exposed by the subsequent quantitative code audit, exhibit the first, second, and third, categories of algorithmic opacity, relating respectively to intentional, technical, and inherent opacity. Firstly, while it must be noted that the OCME was not operating within a competitive market, and had no commercial proprietary interest in the FST, significant efforts were made to avoid regulatory, and legal, oversight. That regulatory oversight would have required the independent validation and adversarial testing of the software (and development material) and publication of results. Next, the FST cases provide an example of technical opacity, deriving – initially – from the comparative lack of technical awareness and literacy amongst defendants, and public defenders, compounded with a dearth of resources necessary to address these issues. Lastly, the FST case displayed a form of inherent opacity. This related to a data-discard function which had been introduced during development, as an improvised solution to resolve other software issues, and contravened both the published methodology of the FST, and that promulgated in oral evidence.

These themes, involving lack of validation and opposition to oversight, would recur in subsequent cases involving commercial PG software packages, as detailed below. However, it is first necessary to place the foregoing analysis in a legal and regulatory context. As stated, *supra*, this analysis gauges the purported validity of PG software variants (and prospective forensic AI developments) in correspondence with rationalist evidentiary norms, instantiated through the comprehensive regulatory requirements laid down by the US PCAST report, the ENFSI '*Guidelines for Evaluative Reporting in Forensic Science'* and the UK Forensic Science Regulator's *'Guidance on Software Validation for DNA Mixture Interpretation.*'[41] The FSR guidance[42] offers a number of solutions aimed at ensuring that the development, validation, and use, of proprietary forensic software conforms to the highest standards. The guidance now requires oversight

[41] Forensic Science Regulator (n 11).

[42] The FSR guidance is itself based upon the preceding PCAST report, see n 12.

involving routine operating quality checks and addresses data input considerations. Thus, minimum standards are now specified for a DNA profile to be considered suitable for interpretation, and criteria for reports now requires that all relevant information used in the calculations be included, in addition to 'the alternative scenarios considered to facilitate checking, auditing and defence review, and the reproduction of results.'[43] Further, the population genetic issues which surfaced in the *Herskovic* case have been addressed, the guidance stating that, '…in relation to population genetic issues, the ability to specify a range of ethnic databases is essential.'[44] In procedural terms, this stipulation answers the need to provide comprehensive background data in relation to those variables which may influence the result of a particular forensic calculation. In summary, these technical requirements together constitute a quality management framework which embeds transparency into all stages. Further, it ensures that technical opacity is addressed through stringent reporting requirements which, also known error rates. Discussion now turns to the legal and regulatory responses triggered by the paradigm example of intentional opacity in proprietary forensic software.

The zenith of protection of proprietary interest protectionism was reached in the case of *Commonwealth v Foley*,[45] the first case to challenge the foundational validity and scientific pedigree of a commercial PG software system. This case involved the assault and murder of a dentist at his home. A mixed sample of DNA from two individuals – presumably the victim and the murderer – was recovered from under the victim's fingernails. Three experts testified that the DNA was consistent with that of the accused, a state trooper who had been living with the victim's estranged wife. However, the experts' probabilistic determinations differed by several orders of magnitude, ranging from 1 in 13,000 to 1 in 189 billion. The latter statistic was arrived at by using a proprietary software package (TrueAllele) marketed by Cybergenetics, a company owned by one of the reporting scientists. The defence challenged the expert's testimony on the grounds that this automated PG approach constituted

[43] Forensic Science Regulator (n 11) 16.
[44] ibid 17.
[45] *Commonwealth v Foley* (n 15)

a novel and unproven method.[46] Further, they requested the release of the source code in order to conduct validation tests. The courts ruled against the *Frye* challenge and denied access to the proprietary algorithms on commercial grounds, stating that, 'TrueAllele is proprietary software. It would not be possible to market TrueAllele if it were available for free.'[47]

Further, the court in *Commonwealth v* Foley, stated that scientists were not in any case prevented from assessing the reliability of a software package absent the release of the source code, accepting the argument proffered by the makers of TrueAllele that the publication of the results of internal validation studies in peer-reviewed journals signaled that the scientific community had debated, and accepted, the scientific foundations of the PG package. Thus, TrueAllele was held to have met the US *Daubert* test[48] for expert scientific evidence. However,

[46] The US courts introduced the Frye standard (*Frye v United States*, 293 F 1013 (DC Cir 1923)) in order to determine the admissibility of expert opinion evidence. This test holds that expert testimony based upon scientific techniques is only admissible when these techniques have become generally accepted within the relevant scientific community. It has now been superseded in the preponderance of US states by the *Daubert* test, discussed *infra.*

[47] *Commonwealth v Foley* (n 15) 889.

[48] Following the judgment in *Daubert v Merrel Dow Pharmaceuticals* 509 US 579 (1994), the Supreme Court amended Rule 702 (regarding the use of expert testimony) to introduce the *Daubert* admissibility test. Within the preponderance of US states, all expert opinion evidence must now meet the *Daubert* standard, measured against five criteria. Daubert requires that, in judging the admissibility of expert evidence, the court must look to the underlying methods used, in order to assess: whether a method can or has been tested; the known or potential rate of error; whether the methods have been subjected to peer review; whether there are standards controlling the technique's operation; and, the general acceptance of the method within the relevant community. Thus, the judge exercises a gate-keeping function, and must now ensure that all expert testimony 'proceeds from scientific knowledge''. It should also be noted that the UK now employs an 'enhanced *Daubert*' test, see Tony Ward, 'An English *Daubert*? Law, Forensic Science and Epistemic Deference' (2015) 15(1) Journal of Philosophy, Science and Law 26. See also, Karen M Richmond, 'The Forensic Regulator Bill: Articulating Normative Standards in a Forensic Market' in K Jakobs and D-H Kim, (eds), *Proceedings of the 25th EURAS Annual Standardisation Conference: Standards for Digital Transformation: Blockchain and Innovation* (Verlag Mainz 2020) 245–59.

as Kwong[49] argues (elaborating upon oblique criticisms contained in a PCAST report),[50]

> … having internal validation studies published in peer-reviewed journals does not mean that the scientific community has debated and accepted the science involved; it merely indicates that the peer reviewers did not identify any disqualifying characteristics of the study as it was described by the paper, such as obvious methodological errors or inaccurate analysis [of the reported results].

Utilising Burrell's typology of algorithmic opacity, the cases involving TrueAllele can be said to exhibit intentional opacity, deployed both to maintain market position, and to avoid legal oversight and review. Indeed, the *Foley* case is most notable for the placing of proprietary interests above the rights of the accused. However, it was far from a sole instance of private interests trumping fundamental rights. As of 2017, all defence requests to view the TrueAllele source code had been defeated, or were being vigorously opposed.[51] As for the inherent opacity of the TrueAllele system, it should be noted that the validation studies for this PG package only accounted for use within narrow, pre-defined parameters. However, the system has subsequently been operated outside the validation parameters. Thus, development, application, and a concomitant extension beyond the validated methodological boundaries can, in this instance, be seen to generate inherent opacity. Further, whilst the designers of TrueAllele have claimed that it is 'impossible' for the package to return a false positive,[52] others have been more circumspect about the possibility of error.[53]

---

[49] Kwong (n 33) 289.
[50] President's Council of Advisors on Science and Technology, 'Forensic Science in Criminal Courts' (n 12).
[51] Kwong (n 33) 292.
[52] See Exec. Office of the President, President's Council of Advisors on Science and Technology, *An addendum to the PCAST Report on Forensic Science in Criminal Courts* 8 (2017) at 8President's Council of Advisors on Science and Technology, 'An Addendum to the PCAST Report on Forensic Science in Criminal Courts' (2017) 8.
[53] Kwong (n 33) 290.

The legal and regulatory guidance specified in relation to the FST, *infra*, remains pertinent. In this instance, the guidance places an onus upon the developer to explicitly acknowledge errors and mistakes, particularly in relation to the overall error rate, and to analytical mistakes, for example: whether the model on which the software is based rests on unjustifiable assumptions; and whether mistakes in software coding result in inaccuracy and unreliability of function.[54]

The requirement of transparency is placed within a framework for end-to-end validation, which encompasses both conceptual, and end-user, validation. The process commences with the requirement to establish conceptual validity which states that, when publishing developmental studies,

> ideally the underlying data on which conclusions are based should also be made available, for example, as supplementary material within the journal or access provided online to downloadable material including all data and a full statistical description. This enables other scientists in the field to inspect it independently and verify the results obtained in order to enable general acceptance of the model concept within the scientific community. Such transparency is essential for any software used within the CJS, for which there can be no 'secret science'.[55]

At the other extreme, the guidance requires end-user validation from the court reporting officers, who need to be satisfied, through the provision of full validation documentation – plus formal assessment and authorisation by their respective organisations – that the software they are relying upon to provide expert opinion is fit for purpose and will not result in misdirection of the court.[56] Indeed, some developers of proprietary software systems have striven to meet the required levels for transparency, and to address known errors in their source code. A notable example occurred in relation to STRMix (a proprietary software package designed by New Zealand's Crown Research Institute, in collaboration

[54] Forensic Science Regulator (n 11) 24.
[55] ibid 26.
[56] This requirement is encapsulated in the Criminal Practice Directions Rule 19A.6(b), and the Federal Rules of Evidence Rule 702.

with Forensic Science South Australia), whose makers drew attention to two coding errors, the inclusion of which had affected the results of DNA analyses in a significant proportion of criminal cases.[57] Further, STRMix has released its source code to defense teams for inspection subject to a confidentiality agreement. Whilst this provides a rare instance of intentional transparency, it nonetheless supports the apprehension of inherent opacity, as endemic to complex algorithmic systems. In the final section, discussion turns to the legal implications of such algorithmic opacity, and discusses the implications for forensic AI packages.

6. Legal Implications and Solutions

As demonstrated above, the introduction of 'black-boxed' algorithmic decision-making systems have given rise to a number of inter-related legal issues. These crystallise around one question, appositely framed by Jeanna Matthews; 'In a society that purports to guarantee defendants the right to face their accusers and confront the evidence against them, what then is the role of black-box forensic software systems in…decision-making in forensic science?'[58] The question surfaces the inherent tensions between resort to algorithmic efficiency, and the paramount importance of established legal principles: the right to a fair and public trial; the rights of accused persons to review and confront the evidence against them; and the right to equal justice under the law. It is argued that there are few circumstances which might be envisaged in which the former should supercede the latter. Indeed, as has been demonstrated, such supersession may run counter not just to legal principle, but to the procedural rules of evidence. As Murphy argued in relation to the courts' protection of proprietary interests in the TrueAllele cases, 'courts would not accept opinions from witnesses not shown to have the qualifications as an expert, so, too, should courts not accept opinions from digital 'experts' without probing the 'qualifications' of the technology.'[59] It may be further argued that the true issue extends beyond the

[57] Kwong (n 33) 292.
[58] Matthews and others (n 37) 321
[59] Erin Murphy, *Inside the Cell: The Dark Side of Forensic DNA* (Nation Books 2015).

'qualifications' of digital experts, which may have been widely accepted within the scientific community, and whose use may have been uncontroversial, at least to the extent that they remained unchallenged. Rather, in relation to AI and advanced machine learning systems, the 'opinions' of algorithmic experts are a direct product of their opaque underlying methodologies. As such these outputs constitute a 'digital *ipse dixit*.' The *ipse dixit* rule, a prohibition of arguments from authority and unsupported expert opinion, extends across multiple legal systems and domains, restricting experts from offering unsupported opinion evidence.[60] The principle focusses neither on the expertise, nor the experience, of the witness but rather on the underlying methodology on which the expert claims are based. Thus, claims from expertise and experience may be validly proffered, provided that such claims are supported by a clear explanation of how experience leads to conclusion; why experience is a sufficient basis for such testimony; and how said experience may be reliably applied to the facts.[61] In the context of algorithmic decision-making, and forensic AI, it is posited that this elementary duty to provide support for an assertion cannot be discharged, or avoided, absent of the rigorous validation processes detailed, *infra*.

It remains to consider the implications, and possible solutions, for forensic ML, and AI, applications. Given the above, it is clear that the introduction of machine learning processes within the forensic, and (international) criminal justice fields, may compound the problems already posed by tertiary forms of algorithmic opacity. This applies to both pattern-matching techniques, which lack the foundational validity of DNA profiling, and to attempts to use quantitative analyses, or visual recognition, in order to process mixed DNA profiles, or to filter open source data. A number of solutions legal and technical solutions present themselves. First, the use of such approaches may be controlled by way of legislative intervention, aimed at limiting or regulating their use.

---

[60] Michael J Saks, 'Banishing *Ipse Dixit*: The Impact of *Kumho Tire* on Forensic Identification Science' (2000) 57 Wash & Lee L Rev 879.

[61] See *United States v Frazier* 387 F 3d, 1244 (11th Circuit 2004) (*en banc*), in which scientific opinion evidence was excluded, the forensic specialist having failed to establish the methodological reliability of his opinion.

Indeed, the European Commission White Paper on Artificial Intelligence[62] makes a number of recommendations in this area. These recommendations reflect seven key requirements listed by the High-Level Expert Group.[63] Of these seven, the Commission identifies a lack of transparency in AI as a particular risk, positing that existing EU, and national, legislative frameworks could be improved in order to address the current lack of oversight in this area. The Commission expressed particular concerns over the use of opaque AI in the private sphere, stating that,

> The lack of transparency (opaqueness of AI) makes it difficult to identify and prove possible breaches of laws, including legal provisions that protect fundamental rights, attribute liability and meet the conditions to claim compensation. Therefore, in order to ensure an effective application and enforcement, it may be necessary to adjust or clarify legislation in certain areas.[64]

The Commission uses the term 'high-risk AI systems' when addressing those systems whose capabilities, functional protocols, and limitations are not explicitly articulated.[65] It is proposed that the legal response may be extended to the international criminal justice arena. However, as discussed, *supra*, softer legal and regulatory responses have been promulgated, such as the use of software audits, and open source systems. However, these solutions may be limited by a lack of requisite expertise, and a lack of diffuse experience across legal jurisdictions. In addition, more general developments in legal and forensic training, might serve to address the need for improved interdisciplinary communication, and the need to compass the normative and epistemological

---

[62] Commission, 'White Paper on Artificial Intelligence: A European Approach to Excellence and Trust' COM (2020) 65 final.

[63] The 2019 experts group lists seven key requirements under the following heads: Human agency and oversight; Technical robustness and safety; Privacy and data governance; Transparency; Diversity, non-discrimination and fairness; Societal and environmental wellbeing, and; Accountability.

[64] COM (2020) 65 final (n 62) 14.

[65] See Riikka Koulu, 'Human Control over Automation: EU Policy and AI Ethics' (2020) 12(1) European Journal of Legal Studies 9.

requirements of allied fields.[66] Technological 'solutions' – for example a resort to 'constrained AI'[67] – may be attempted. However, these involve a significant compromise in efficiency whilst failing to eliminate the risks exposited above. In conclusion, none of these solutions should be approached in isolation. Indeed, Matthews recommends that 'both in research and in casework, an emphasis should be placed on comparisons between multiple reasonable systems' evaluations of the same input data.'[68]

The comparative lack of diffuse expertise within the international criminal justice sector may cause further complications. In relation to evidence handling, the ICJ sector is notable for a marked spatial and temporal divergence separating evidence collection, stabilization, evaluation, and reporting. In a domestic jurisdiction these processes are approached holistically, through the joint efforts of forensic experts and allied institutional agents, who together shape the evidential trajectory. However, in the context of alleged international crimes there exists a fundamental bifurcation between collection and stabilisation of evidence – particularly in relation to open source data collected and filtered by members of the public and NGOs – and its subsequent evaluation and reporting by prosecution experts. Whilst proponents of open source investigation may highlight the potentials of emergent open source data collection and processing systems to furnish the international courts with evidence, in light of the foregoing discussion it may be stated with relative certainty that by placing forensic AI systems in the hands of uncertified volunteers, their functions are comparatively less amenable to control. Therefore, to conform with regulatory

[66] See Chris Lawless, 'A Curious Reconstruction? The Shaping of "Marketized" Forensic Science' (2010) CARR Discussion Paper 63; Christopher James Lawless, 'Policing Markets: The Contested Shaping of Neo-Liberal Forensic Science' (2011) 51 British Journal of Criminology 671; Sally F Kelty, Roberta Julian and Alastair Ross, 'Dismantling the Justice Silos: Avoiding the Pitfalls and Reaping the Benefits of Information-Sharing between Forensic Science, Medicine and Law' (2013) 230 Forensic Science International 8; The Rt Hon the Lord Thomas of Cwmgiedd, 'The Legal Framework for More Robust Forensic Science Evidence' (2015) 370 Philosophical Transactions of the Royal Society B 20140258, 1.

[67] Constrained AI founds on paramaterised algorithms operating within limits set by the operator. These are utilized in an attempt to increase the tractability of machine learning and AI processes.

[68] Matthews and others (n 37) 322.

guidance, levels of access should be imposed such that only the input variables can be defined by the operator, 'whilst access to files that define the analytical parameters would require a higher level of authorisation. System access logs, settings changes and parameters used for past tests should be auditable.'[69]

This leads to a broader issue, which goes beyond the fundamental need for transparency and accuracy in forensic reporting In the context of a criminal investigation, a calculation should proceed only if the software is capable of aiding a meaningful interpretation. It should be borne in mind that while the efficiencies offered by machine learning may prove increasingly attractive to researchers and practitioners, academics have aptly demonstrated that efficiencies drawn from mathematical expertise and human endeavor are still capable of delivering the most accurate and transparent efficiencies.[70] Thus, the international criminal justice system should be particularly circumspect in its engagement with novel but opaque technologies whose underlying methodologies resist exegesis. In the allied fields of international criminal justice, legal research, and forensic science – where the interpretability of results, and the explicability of propositional foundations, are at a premium – the utilisation of machine learning, and AI systems, should be exercised with caution, particularly in respect of thee more complex, and comparatively opaque, instantiations. The efficient processing of data must be tempered by 'healthy skepticism about the design, development, and use of complex software systems used in criminal justice.'[71] Otherwise, the established principles of rational inference, rectitude of adjudication, and legal order, negotiated collectively over centuries, could be fatally undermined by the introduction of automated systems whose logics cannot be explained.

---

[69] Forensic Science Regulator (n 11) 19.

[70] Therese Graversen and Steffen Lauritzen, 'Computational Aspects of DNA Mixture Analysis' (2015) 25 Statistics and Computing 527. See Faculty of Science, 'Danish DNA Detective Helps English Police in Homicide Cases' (*University of Copenhagen*, 23 May 2018) <https://www.science.ku.dk/english/press/news/2018/danish-dna-detective-helps-english-police-in-homicide-cases/> accessed 17 January 2021.

[71] Matthews and others (n 37) 322.

Volume 5, Number 1, Spring 2021

The Impact of AI on the Law

Rethinking the public-private dichotomy in the age of algorithms

Laure Helene Prevignano*

We are now more than half a century into the digital revolution. However, in recent years, our societies have made rapid progress toward a higher level of digital maturity, particularly with regard to the developments of Artificial Intelligence (AI)[1], one of the most pivotal phenomena of digital advancement. Thus, any serious long-term prognosis concerning the future shape of societies and their legal framework runs the risk of becoming whimsical[2]. However, some thoughts might be of interest.

In this context, this paper aims to examine how AI might blur the already murky boundary line separating the public and private powers within the legal system, thus making most legal systems relatively inadequate to the reality they aim to apprehend. Qualms about the mounting confusion surrounding the public-private divide are not novel. In 1957 already, scholars were wondering what legal factors impeded a reassessment of the relation between the State and group power[3]. Nowadays, similar concerns are voiced, for instance about the growing influence exerted by private entities without being subjected to some

* [laurehelene.prevignano@unifr.ch]

[1] Ryan Calo, 'Artificial Intelligence Policy: A Primer and Roadmap' (2017) 51 UC Davis Law Review 399, 404–35, ici 405. Whereas AI can be regarded as an umbrella term entailing many technologies, it will nevertheless be referred to within this paper for clarity purposes.

[2] Gudula Deipenbrock, 'Is the Law Ready to Face the Progressing Digital Revolution? – General Policy Issues and Selected Aspects in the Realm of Financial Markets from the International, European Union and German Perspective' (2019) 118 Zeitschrift für Vergleichende Rechtswissenschaft 285, 286.

[3] Wolfgang G Friedmann, 'Corporate Power, Government by Private Groups, and the Law' (1957) 57 Columbia Law Review 155.

guarantees regarded as proper to the State[4], or about the incremental tendency of the State to allow public domain to land in private hands[5]. Expressions of this confusion may be found under various question marks. Should healthcare be public or private? Should human rights generate obligations for private entities? Should, and more specifically *how* should transnational corporations be made accountable, considering the enormous impact they have on individuals[6]? Should the role of the State be redefined[7]? The list goes on.

Interestingly, those interrogations all seem to arise from the fact our legal system strongly and structurally revolves around the divide between public and private entities, each endorsing specific right and duties, to the point where this model is hardly ever challenged *per se*. However, beyond legal roles attributed in accordance with this basic legal dichotomy, shouldn't also the dichotomy in itself be examined more closely, as well as the impact AI will have on it? In effect, AI and the prospects it brings might exacerbate the fragmented character of the division and lead to the emergence of new forms of centralized entities, ultimately deeply disrupting our legal landscape.

After concisely examining the notion of the State as a central public entity, its history, role, as well as the evolution of the influence of private entities (1), it will be interesting to delve into the specificities of AI as a technology and the peculiar impact they may have on the relation between public and private entities (2). Then, some specific angles from which the

---

[4] Gary Younge, 'Who's in Control – Nation States or Global Corporations?' *The Guardian* (London, 2 June 2014) <https://www.theguardian.com/commentisfree/2014/jun/02/control-nation-states-corporations-autonomy-neoliberalism> accessed 17 January 2021.

[5] There is indeed a growing phenomenon of privatization, which will be briefly discussed further in this paper.

[6] Michael Goodhart, 'Democratic Accountability in Global Politics: Norms, not Agents' (2011) 73 The Journal of Politics 45.

[7] Jean-Pierre Raffarin, Emmanuelle Auriol and Augustin de Romanet, '« 2019, la fin d'un monde ? » : faut-il redéfinir le rôle de l'Etat ?' *Le Monde* (Paris, 23 March 2019) <https://www.lemonde.fr/economie/video/2019/03/23/2019-la-fin-d-un-monde-faut-il-redefinir-le-role-de-l-etat_5440198_3234.html> accessed 17 January 2021; Lukas van den Berge, 'Rethinking the Public-Private Law Divide in the Age of Governmentality and Network Governance: A Comparative Analysis of French, English and Dutch Law' (2018) 5 European Journal of Comparative Law and Governance 119, 122.

figure of the State may be weakened, thus increasing the inadequate character of the public-private dichotomy, will be discussed (3). Further on, this paper will consider how the possible fragmentation of this segregation might be the mere expression of the erosion of the rule of law as a whole, or prove to be part of a distinct phenomenon (4). Some critics and perspectives will be explored (5), before allowing for a brief conclusion.

1. The State: history, role and powershifts

1.1 Emergence of the State and modern role

The divide between public and private within the law seems difficult to apprehend properly without examining the notion of the State in which it is rooted. However, since this paper does not aim to discuss historical or societal questions, it will be succinct on this – fascinating – topic. Moreover, given that the concept of public and private is traditionally regarded as being antagonistic, any reflection on the role of the State necessary mirrors the aforementioned dichotomy. This angle of approach has thus mainly been chosen for this analysis.

Intriguingly, even though we seem to live in a '*world of states*'[8], it has not always been the case. The dominant institutional forms have evolved over time, successively taking various shapes and colours, and such a shift may be happening again. Indeed, the contemporary governance structure might be undergoing some transformation, as it already has in the past – one might think of central powers embodied in the figures of empires, feudal states or cities – or even revolutionised.

Broadly speaking, public law was developed as a response to the feudal system, in which public and private law were not differentiated. The *State* was thus incrementally considered as an entity having to pursue general interest instead of individual ones, and thus guided by principles serving the common good. While the State is far from being the only actor within the legal system, and, *a fortiori* politics, his role is largely recognised as having an enormous impact

[8] Idiom notably used by J. D. B. Miller in his book, JDB Miller, *The World of States: Connected Essays* (Croom Helm 1981).

on individuals living under its yoke. While public law was envisioned as vertical, handling the relations between an individual and the State, private law was depicted as horizontal, that is, regulating the relations between individuals[9]. Consequently, the bodies of public and private law have developed with their own principles and procedures. Gradually, the State has assumed more and more tasks and responsibilities[10].

Nowadays, the notion of the State can be defined in many ways. One commonly accepted definition within the field of political sciences is that given by *Max Weber* who refers to the State as a human community that successfully claims the monopoly of the legitimate use of force within a given territory[11]. The legal field mainly addresses this delicate definition through the lens of international law, which apprehends the State as an entity presenting the following features: it should possess a permanent population, a defined territory, a government, and have the capacity to enter into relations with other States[12]. Both of these definitions of the State, and as a result the corresponding notion of private entities, may have become unsuitable for the reality of our structures, and *a fortiori* of their influence and power.

1.2 Balance of power and legitimacy: the end of Rousseau contrat social?

To express these powershifts more concretely, it is worthwhile to consider some facts. Google's parent company, *Alphabet*, out-earned Puerto Rico in 2017,

---

[9] van Den Berge (n 7) 121 ff.

[10] Indeed, the State has incrementally penetrated into society, mostly during the 19th and 20th century, because of economic and social developments, progressively becoming the 'welfare state', Chris Renwick, 'Why We Need the Welfare State More Than Ever' *The Guardian* (London, 21 September 2017) <https://www.theguardian.com/news/2017/sep/21/why-we-need-the-welfare-state-more-than-ever> accessed 17 January 2021.

[11] *Encyclopedia of Power* (2011) 400 ff.

[12] When defining the State within international law, the Montevideo Convention is usually referred to. Montevideo Convention on the Rights and Duties of States, signed at Montevideo, 26 December 1934.

reporting earnings that surpassed the entire GDP of the country[13]. In the context of our postmodern world where policy making and implementation powers shift ever faster from single states to larger supranational entities and global regulatory apparatuses, the financial power of those tech giants is even magnified. Indeed, it is simpler and more affordable than ever for those companies to extend their reach globally, as facing a more centralized legislative framework implies fewer expansion costs. Scandals like '*Cambridge analytica'* which is not unheard of but unprecedented both in its scope, its reach and the depth of its influence, have also emphasized the enormous influence large corporations have on individuals' daily lives and caused considerable turmoil among civil society, shedding light on the inadequacy of the current system. While the State remains *the* primary democratic entity on paper, due to globalisation and the power of financial capital lying in the hands of private entities, it would no longer be up to this role[14]. Governments struggle more and more to pursue and enforce national agendas, which haven't been endorsed by international capital first. It has been argued that the recent nationalist wave spreading across Europe and reflected by the European parliamentary elections would be an expression of this situation. Whereas it has readily and willingly been described as xenophobic, it would rather incarnate the fear that the system we evolve in is shaped and controlled by diffuse and fuzzy private forces[15]. Thus, concern has been expressed about the way our legal system currently (*does not*) reflect(s) those developments in a satisfying manner, notably with regard to accountability of private action, as mentioned, but also with regard to the potentially insufficient transparency and efficiency of the public one. Regarding this last point, the lack of action taken and resources mobilized by governments to tackle the climate crises provides a clear example of the critic according to which the current model of the State is not the best to address global challenges. In addition to climate change, terrorism or pandemics could also be mentioned. All these phenomena raise the question of private and public entities' legitimacy; power does not seem to be correlated

[13] Fernando Belinchón and Qayyah, '25 Companies That Are Bigger Than Entire Countries' (*Business Insider*, 25 July 2018) <https://www.businessinsider.com/25-giant-companies-that-earn-more-than-entire-countries-2018-7> accessed 17 January 2021.
[14] Younge (n 4).
[15] ibid.

with the will of the greatest number of people anymore[16], and the current legal structures and apparatus seem not to have been able to keep pace with an increasingly global and digitalized environment.

2. General impact of AI peculiarities on the existing legal landscape

2.1 A fear of the unknown like any other? Mind the gap

While those concerns and challenges have been brought to light a while ago, the magnitude of the effects brought about by the stupendous development of AI for our legal system might be unprecedented, and thus, highlight the unsuitability of the structural distinction between public and private fields. In other words, as if the fine line between public and private powers was not blurry enough, AI's peculiarities as a technology may exacerbate this confusion.

One could argue that other technologies regarded as revolutionary, like electricity or nuclear power also brought about profound societal and legal changes without affecting the fundamental division between private and public powers. They may have contributed to the blurring surrounding this division, notably by exacerbating the powershifts mentioned above[17], but only relatively, and in any case not sufficiently to question legal structures. The fundamental division between public and private powers has been challenged in the past. However, the control private corporations have nowadays over the information and communication systems is unprecedented. Thus, AI could potentially lead to a certain reconfiguration of this dichotomy, without challenging it entirely. However, this comes down to assuming AI does not significantly differ from those technologies, whereas its specific traits may potentially generate a particularly important impact on the legal landscape. Therefore, it might be useful to consider some distinctive characteristics of AI with regard to other

[16] This conception of legitimacy largely impregnated our legal culture: one might think of the influence of '*Le contrat social*' de Rousseau.

[17] Without entering into any specifics, this is notably due to the more powerful role thus endorsed by private entities mastering those technologies.

technological developments, and see how they may underline the inadequacy of the public-private distinction in order to apprehend the influence and power shift at stake.

First, it could be noted that there is a considerable technology gap governments should mind. It will be difficult to pass laws without the necessary knowledge bound to it[18], as discussed further in this paper. Although it is common practice for legislative bodies to involve the private sector in the process of establishing a legal and regulatory framework, the broad and complex nature of AI further restrains the ability of central powers to make informed policy decisions and elaborate corresponding regulation. As a result, the ability for the central power to make informed policy decisions and elaborate corresponding regulation might be restrained. While a certain lack of expertise from the government is not exclusive to AI as a technology, the complexity of such systems may be unmatched. Delegating some policy-making tasks and responsibilities might compensate some lack of technical knowledge. Such an outsourcing is not something that has never been done before. However, the extent of this externalisation may be unprecedented in the case of AI. This reality will seemingly require from the public sector to rely on the expertise of the private one[19], far more than what was the case for other technical advancements, since the involvement of the State in the case of AI may incrementally evaporate. Such a degree of dependence, in addition to its regulatory implications as discussed below, raises serious questions about the figure of the public institution, and therefore, about the divide between public and private within the law.

2.2 Volatility or versatility

Another peculiarity of AI systems rests in the fact that the players in the global technology industry which constitute the main driving force behind AI advancements spread across the globe, since one single person does not require the same amount of resources and infrastructure a large company would in order

[18] Matthew U Scherer, 'Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies and Strategies' (2016) 29 Harvard Journal of Law and Technology 353, 380.

[19] Namely to elaborate appropriate laws.

to write computer code and engage in the AI-related venture[20]. These individual actors may not even be part of any kind of organization and their activities may prove to be very delicate for a central institution to regulate. Progressively, it might lead to a considerable loss of influence of governments.

Plus, the actors mastering this technology may be more difficult to identify than the ones mastering previous ones. For instance, nuclear weapons are expensive to elaborate and demand components that are difficult to obtain. In other words, private entities dealing with such technology have to be large enough in order to do so, and thus are easily identifiable. Consequently, they are also more likely to be apprehended, controlled and regulated. By comparison, AI applications may be relatively inexpensive and affordable to produce, even mass-produced. The loss of governments' central power thus induced by the creation and use of AI might be tremendous because of this expanded affordability. What is more, the impact of an actor may be inversely proportional to its size: sophisticated software can be designed as much from a slum as from the golden glasshouse of a billionaire corporate. Consequently, AI can be regarded as a technology with a different and far broader impact than the technologies that have emerged so far.

In addition to this, even if the State manages to identify the players, any rule may be hard to enforce, since any software may be developed in any country worldwide without difficulty. This may pose a supplementary challenge for the notion of jurisdiction and for the laws a State traditionally enforces within its own territory and boundaries[21]. In addition to this, participants in the AI-related venture may easily relocate in another country with more lax laws. Considering the relatively low cost of infrastructure discussed above, and the tiny physical footprint needed to develop such a technology[22], attempts by States to regulate and embrace their citizens' activities may prove to be ineffective. As a result, central institutions will probably be deeply challenged by the specific nature of AI as a technology, due to its volatile nature.

---

[20] Scherer (n 18) 370.
[21] ibid 372.
[22] ibid.

3. The State figure – concrete flaws and illustration of an incremental fragmentation

3.1 Data as a precious resource, jeopardizing of security systems and use of 'legitimate violence'

After having discussed a few reasons why AI may impact the law and regulatory environment in general, it may be of interest to delve deeper into the question of how the very notion of "State" might be specifically challenged by AI. Once viewed as omnipotent, the concept and relevance of the State today may be undermined in some particular ways. Is the *Leviathan* as *Hobbes* described it disappearing? According to the renowned futurist *Yuval Noah Harari*, the mere idea of a coherent nation-state is now threatened[23], and this is only one voice among others.

First, attention should be paid to one fundamental feature of AI in this conversation: *data*. Indeed, artificial intelligence is ultimately tied to and thrives on data. What started as a discussion about the control of individuals over their personal data translated into a discussion about the power of data and private data collection in general.[24] Citizens seem to be bound to become consumers, giving up their data in order to access whatever they need to, be it a public service, a pharmaceutical product or a leisure service. The increasingly fuzzy distinction between citizens and consumers seems to match the growing confusion surrounding the dichotomy of public and private within the law. As a result, citizen-consumers may not fully understand the implications related to the sharing of their personal data. This can become especially tricky when considering the relatively recent measures on data-sharing imposed by both national and supranational authorities in response to emerging security threats. Due to this growing confusion, citizen-consumers may encounter some difficulty to realize the consequences of them sharing personal information in

[23] Helen Lewis, '21 Lessons for the 21st Century by Yuval Noah Harari review – A Guru for Our Times? *The Guardian* (London, 15 August 2018) <https://www.theguardian.com/books/2018/aug/15/21-lessons-for-the-21st-century-by-yuval-noah-harari-review> accessed 17 January 2021.

[24] Calo (n 1) 420.

function of the context: public requirements or private ones. Even considering the existence of culturally different perspectives on the concept of privacy and personal data, an increasing amount of private information is given away beyond the full awareness of their theoretical owners. As a matter of fact, the mere purpose of AI is to spot and detect patterns a single person cannot distinguish[25]. Consequently, a dizzying and ever-increasing amount of data is being handed to private entities, offering them a fundamental resource and advantage as compared to a State in the landscape of AI. Thus, beyond possession of financial means, which already offers an enormous power in setting the various policies through lobbying as briefly examined earlier, the possession of data might *de facto* place private actors in a position where they exert even more influence on policies and rule-making, to the point where one could wonder which side actually exerts influence on which (see 4.1). Even though this phenomenon is not novel *per se*, the shift in the balance of power between the public and private sectors is expected to accelerate, driven by AI's developments. It is difficult to see how the central power could not lose at least some legitimacy without denying the importance of such resources in setting agendas.

Furthermore, the possible weakening of the State might be due to security issues. Indeed, not only might it be more than delicate for a government to control the players of this new game, but it might also prove to be extremely challenging to play and defend against them if they breach the rules. While challenges posed by private entities to the central power are certainly not something new, their dimension and scale risks being of a different magnitude. In fact, non-state actors playing in the AI field will probably also be able to conduct more attacks against the central power, with less time, funds, or manpower. Plus, those possible nefarious actions may be precisely targeted, very effective and almost impossible to assign to someone because of their volatility[26]. In addition to this, they can also effortlessly be performed anonymously[27]. This technology is thus very different from previous ones in the sense that it can

[25] ibid 421.

[26] Paige Young, 'Artificial Intelligence: A Non-State Actor's New Best Friend' (*Over the Horizon*, 1 May 2019) <https://othjournal.com/2019/05/01/artificial-intelligence-a-non-state-actors-new-best-friend/> accessed 24 January 2021.

[27] Scherer (n 18) 370.

directly be used to hack and seriously affect the central institution. Even without any actual threat of cyber attack paralyzing central institutions, the vulnerability to which they are exposed may weaken the model of an omnipotent State and further undermine the dual model most legal systems revolve around.

3.2 Privatization and 'de facto regalian function'

Another issue to be addressed has to do with the growing privatization taking place in our societies, and as a result, in our legal systems. While already occurring in the past, the gradual shift from the fulfillment of tasks considered as public from the government to private entities made the legal distinction between public and private law more and more difficult given the complex nature of AI [28].. This phenomenon may prove even more difficult to address in our ever-growing technological world. However, the privatization of tasks that were historically considered to be the responsibility of some public authority should not be confused with the public sector using AI itself. This might pose a different set of difficulties. Thus, after briefly discussing the use of AI within the public sector, the privatization of the public sector in general will be examined.

Regarding the use of AI within the public sector, it should be noted that an increasing number of public tasks are automatized. While theoretically remaining in public hands, automation is shaking up the State to its core, challenging some basic assumptions we make when considering the guarantees offered by the State. Examples of the use of AI in public administrations stem from diverse areas, for instance in the fields of predictive policing, court proceedings or control of traffic[29]. Specificities of the use of AI in the public sector[30] may challenge some public guarantees, such as the right to a fair process,

[28] van den Berge (n 7) 133.

[29] Such are the suggestions put forward by Nadja Braun Binder. Nadja Braun Binder, 'Künstliche Intelligenz und automatisierte Entscheidungen in der öffentlichen Verwaltung' [2019] Schweizerische Juristen-Zeitung 467, 470 ff.

[30] According to Nadja Braun Binder, specificities that would challenge traditional public guarantees are mainly three. Decisions resulting from algorithms are not easily comprehensible (at least with regard to machine learning procedures), machine-learning procedures must be trained before they can be used, and a huge amount of data is processed.

and thus call into question public law as an exceptional set of rules within the legal system. Some argue that the State could still be able to perform its tasks properly by following certain rules and standards. However, these concerns seem to attest to the fact that the use of AI in the public sector may indeed challenge its mere nature. Not only are private actors incrementally assuming public tasks, but the public one also seems to function more and more like a private entity[31], fuelling confusion and interdependence. An illustration of the will to mitigate risks associated with the use of AI in the public sector is the *European Ethical Charter in the Use of Artificial Intelligence in Judicial Systems and their environment*, elaborated in 2018[32].

Regarding the privatization of the public sector in general, one should note that the public sector largely relies on the private one for the use of AI. Thus, the question of privatization, while having been discussed since decades, may take another dimension in the coming years. Indeed, as the industry is assuming a leading role in the development of AI[33], the State is increasingly forced to rely on their services. While privatization has been considered as a mean of rendering the State more efficient, it may now become an absolute necessity, leaving the realm of convenience.

However, even more noteworthy are the somewhat insidious effects this trend may have. As a matter of fact, often, with sovereign tasks come sovereign rights. Anecdotally speaking, as Facebook announced its intention to issue its own digital currency, it was interesting to note that, despite an initial surprise coming with such a statement, most of the reactions were then focussed on security issues and soon translated into (legitimate) concerns of possible

---

[31] An interesting light is shed on some '*private practices*' of the State by Mariana Mazzucato: Mariana Mazzucato, *The Entrepreneurial State: Debunking Public vs. Private Sector Myths* (Anthem Press 2013).

[32] European Commission for the Efficiency of Justice (CEPEJ), 'European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment' (3–4 December 2018) 14 <https://bit.ly/2G18u8x> accessed 24 January 2021.

[33] Calo (n 1) 406.

hacking[34]. However, the fact that a private entity is practically about to endorse a role once thought to be *regalian* does not appear to be shocking. Or at least, not as shocking as it used to be. More generally, one could argue that *regalian* privileges and rights are progressively given to private entities without causing a tremendous turmoil, because they enjoy *de facto* an enormous financial and technological power. With this trend probably increasing with AI, the confusion surrounding the question of who should endorse which role might considerably intensify.

4. Erosion of the State, or erosion of the law?

4.1 Solely some old-fashioned lobbying

Until now, critics have voiced concerns about the influence private entities have on regulation and policy setting in general. As a matter of fact, it is pretty safe to assume that the access to greater financial resources translates into a growing capability to influence policy and law-making altogether[35]. This issue is neither foreign nor recent and the increasing influence exerted by lobbyists over national and international governing bodies is generating entirely legitimate concerns. In the case of the emergence and rapid development of AI, this phenomenon may intensify, with the private sector exercising its openly large influence to impact the regulation. However, once again, the paradigm might be sifting. With AI, the risk may not materialize in an intensive lobbying from the private sector to influence existing rules or standards, but rather in the mere absence of regulation coming from the public one.

There are many reasons why the State may not assume its role of rule-maker and leave regulation behind. One may think of a definite lack of expertise, but also of the incredibly smaller amount of resources injected by central governments into AI research, development and formation. Until now, the

[34] Mike Orcutt, 'Critics Say Facebook's Libra Threatens America's Power. Zuck Says They've Got It All Wrong' (*MIT Technology Review*, 24 October 2019) <https://www.technologyreview.com/f/614621/critics-say-facebooks-libra-threatens-americas-power-zuck-says-theyve-got-it-all-wrong/> accessed 24 January 2021.
[35] Scherer (n 18) 377.

private sector has been using AI way more frequently and intensively than the public one. Even though many governments embrace the idea of progressively introducing this technology to facilitate a broad set of tasks, this is happening at a much slower pace and a more confined scale than in the private sector[36]. In addition to this, regulatory competition among States may contribute to the reluctance to regulate. In fact, despite the fact that various States have set AI as a priority within their policy agenda, they are also conscious that investors won't be attracted to their jurisdiction if they put sharp regulation forward[37]. As a result, regulation, when elaborated, may still be kept to a minimum. Therefore, the centre of gravity of the conversation would not be limited to lobbying: private entities might well be led to regulate this field themselves.

4.2 Expansion of self-regulation

This phenomenon could eventually result in the expansion of self-regulation. Indeed, the non-government sector dramatically needs some predictability and legal framework to embrace the use of AI. In fact, the regulatory power might *de facto* change hands, since the private sector will need to set rules. Many leaders from this industry have indeed voiced concerns and called for more regulation. Beyond tech entrepreneurs and futurists, various academics also seem to agree that *ex ante* action is highly needed to ensure that AI remains under human control and aligned with people's interests[38]. According to them, difficulties regarding supervision and control of AI are likely to materialize and the legal system should be able to mitigate them.

Consequently, our traditional view of regulation stemming from governments might not be adapted to the rapid evolution of AI and its use. Growing self-regulation may also be induced by the State incrementally relying on the private sector's expertise to do so. Thus, the regulatory power might

---

[36] Tod Newcombe, 'Is Government Ready for AI?' (*Government Technology*, July/August 2018) <https://www.govtech.com/products/Is-Government-Ready-for-AI.html> accessed 24 January 2021.

[37] John Armour and Horst Eidenmüller, 'Selbstfahrende Kapitalgesellschaften?' (2019) 183 Zeitschrift für das gesamte Handelsrecht und Wirtschaftsrecht 169, 186.

[38] Scherer (n 18) 368.

progressively shift from the public to the private sector. This has already happened in some fields: for example, the financial sector in Switzerland is largely regulated through its actors (mostly banks, but also private insurance institutions) themselves. Once again, this phenomenon is not new *per se*. It proves to be quite known in industries such as finance, financial services, and banking. In that case, it has been regarded as pretty successful. Indeed, this method seems to allow more flexibility and technical knowledge necessary to draft such rules. However, the reach and scope of self-regulation in the case of AI combined with the tremendous impact on individuals' daily lives of such a technology might provide enormous power to the private sector. A power that used to be conceived as having to lie in public hands.

To sum up, while an erosion of the rule of law might be witnessed due to the use of AI, one could also argue that a shift of the regulatory power will rather be observed, illustrated by the thrive of self-regulation.

5. Critics and perspectives

5.1 Need for the figure of the State

It could be argued that the importance of the State as a legal model will outweigh its - in some aspects at least – desuetude and prevent the complete blur of the public-private dichotomy. In effect, the State appears to embody fundamental features and guarantees. Certainly, some public tasks could be assumed by private entities in the future, and common good policies might even stem from the greater influence, which the private sector exerts on our society. Such a possibility seems worthy of discussion, and AI might well emphasize and underline the need for such a global conversation. However, one of the essential characteristics of public entities as they are framed in our legal systems is the notion of territoriality. It appears to remain one of the fundamental features of the notion of the State and of traditional public law, according to both international law and political science, as previously discussed, and this trait seems a hard one to transfer to large global corporations. The State as a figure of proximity gives room for differentiation, experimentation, diversity, cultures and habits. The probable legal homogeneity possibly induced by growing self-

regulation may lead to a backlash and a valorisation of the State as a central power with a strong local dimension. As mentioned earlier, political colours of recent elections in many european States have emphasized the will for a strong central institution. Consequently, the divide between public and private law may even be strengthened, with public entities eventually acting as a shield against globalization and paying tribute to local voices. The question whether the State as conceived today is able to fulfill these tasks nowadays, and thus if a whole rethinking of our legal system would not be preferable, remains open.

Another interesting perspective is offered by the legal transformations taking place in China. Without delving into details, one can argue that this example stands at odds with the one given by the US, where the approach to AI is rather driven by the industry, and not by the government as it is in the Chinese case[39]. In such circumstances, far from incarnating the figure of a State promoting proximity and individualization within the collective, it also shows how a coercive State figure may be reinforced by AI, also standing far from its original features and duties. More generally, it should be noted that, whereas the elusion of the divide between public and private notably and largely stems from globalization, and might be intensified through AI development, a powerful counter-current might on the contrary reinforce the State, and, as a result, the divide between public and private within the law.

5.2 Alternative perspective: a merge

The reality is that only few corporations have the resources, such as financial means and data, to take the lead in the AI industry. It seems that this technology might thus lead to some kind of centralization and monopoly, public or private[40]. For instance, large companies generate a huge amount of data themselves, and thus have an important strategic advantage in comparison with smaller platforms, or even with some States[41]. Some smaller firms may even encounter growing difficulties to enter the market. Since the use of AI

[39] Børge Lindberg and others, *An AI Nation: Harnessing the Opportunity of Artificial Intelligence in Denmark* (McKinsey & Company and Innovationsfonden 2019) 17.
[40] Calo (n 1) 424.
[41] Armour and Eidenmüller (n 37) 175.

progressively spreads across fields and sectors, it is likely that the control of those important entities will largely exceed the boarders of AI and challenge the legal system more broadly.

Nevertheless, the scenario of mastodons playing above the rules might not materialize as such. As previously suggested, there will most likely be a need for rules, and any central entity, government or company, is likely to require some to function properly, even in the hegemonic way AI may open. Rather than disperse power, AI may centralize it, but neither in the form of the State as currently conceived, neither in the form of a purely private company as we envisage it today. For example, an entity could work like a company, but have a goal set for and pursued by the algorithms to suit better the interests of the shareholders, thus being more representative, even, in a sense, *democratic*. Thus, it could lead to new legal models simply not fitting the legal categories generally referred to as models at this point. As a comparison, the field of international law, whose models and actors were once more defined, has seen the birth of new entities which did more or less break into its once pretty rigid framework; international organisations are taking a major role on the international scene, especially the European Union, which is considered to be '*sui generis*': neither a State, nor an international organisation *stricto sensu*, and nevertheless influencing the international scene more and more.

So, why not imagine the emergence of some kind of Corpo-government, or even of Govern-poration, as a response to the possible obsolescence of the public-private divide in the AI era?

6. Conclusion

It should be noted that public interest, as a legal concept, does not seem to have lost of its value. It is still a valid point of reference[42]. Indeed, even if the fine lines between public and private have been and will be challenged by the development of AI, it does not mean that this differentiation has lost its value *per se*. However, the conception of public law as an exceptional regime within the legal system might prove to be obsolete. More specifically, the peculiar duties it involves

[42] van den Berge (n 7) 135.

should not be the sole remit of the State, but also bind private entities, at least to some extent. Similarly, public actors should eventually be bound to pursuing public purposes in a strict(er) manner.

Even if prognosis do seem ambitious, as mentioned in the early lines of this paper, it seems worthwhile to question the segregation between public and private within the legal landscape, exacerbated by the fundamental transformations induced by AI. Indeed, the answer to some current difficulties might not be solved by asking how to regulate private or public entities, but rather by asking how to create the legal conditions to embrace the fundamental transformation of the actors and power structures the law traditionally aims to regulate. Since most legal systems currently revolve around the progressively fading dichotomy between public and private law, entities and sectors, our legal system might ultimately be profoundly disrupted, in its most ancient and intimate confines.

A final remark might touch upon the fact that this paper was written before the COVID-19 pandemic. Interestingly, from several angles, this health crisis has highlighted various difficulties in distinguishing between public and private within the law, for example with regard to health resources or tracing applications, particularly concerning the collection of personal data. These complicated discussions may prove to be an illustration of this delicate distinction. Thus and finally, AI may also provide a much needed and unique opportunity to rethink the public-private dichotomy as just one way of conceiving the legal landscape among others, perhaps better suited to the era of algorithms.

Volume 5, Number 1, Spring 2021

A 'New Technology World Order'?

Will the Impact of Artificial Intelligence on International Diplomatic Practice Render Existing Diplomatic Law Obsolete?

Anna Kirby*

The legal commentary on artificial intelligence tends to focus on specific practical issues such as liability or security. While these are not futile endeavours, this focus on the implications of the direct implementation of AI within society diverts attention away from the less conspicuous and equally imminent effects of the technology. Development within this field will have considerable effect on international diplomatic practice: through changing the nature of communication itself and transforming the global landscape within which it takes place. The concurrent demise of the nation state and rise of big tech means that many powerful global non-state actors operate outside the sphere of existing international diplomatic law. This illustrates a legal void within which tech corporations act increasingly divergent to state practice, with potentially disastrous consequences for the future of AI development, as ethics are traded-off against profit. An interdisciplinary and multi-stakeholder approach is crucial to develop a governance framework for AI that balances public and private centred interests. In an era of globalisation and digitalisation, there will always still be a need for traditional diplomacy; AI will disrupt the channels through which it is conducted, and it is the contention of this article that while existing International Diplomatic Law requires reform, it is not obsolete.

* LL.B., LL.M. Student in International Law of Global Security, Peace and Development, University of Glasgow [anna@kirby-family.net]

1. Introduction

The impact of artificial intelligence (AI) on the world as we know it is not a novel contemplation. For some, the future of development in this field is fraught with negative connotations and visions of killer robots, technological unemployment; the end of humanity.[1] Others believe advancements in AI could hold the solution to societal problems such as social care, and even global issues like climate change.[2] While these visceral perceptions of its potential effects may turn out to be valid predictions, they are extreme. Discussion in this sphere is largely based around the direct effects of AI technology, and the threats posed by it, on everyday life. However, the less conspicuous consequences may take effect sooner, and we must be prepared. Described as the next general-purpose technology,[3] AI is defined as "the science and engineering of making intelligent machines."[4] Advancement in this domain is often compared to past industrial revolutions, except that it will be both faster and larger.[5] Unlike the steam engine or electricity, AI has the capacity to transcend and alter all aspects of society, and therefore the threshold that must be met for AI to become 'globally disruptive' is much lower than that of general-purpose technologies in the past.[6] Likely long before we see the humanoid robots characterised by science-fiction movies walking our streets, AI would already have a profound impact in international relations and diplomatic practice.

---

[1] Mark Bryant, 'Artificial Intelligence Could Kill Us All. Meet the Man Who Takes that Risk Seriously' (*The Next Web,* 8 March 2014) <https://thenextweb.com/insider/2014/03/08/ai-could-kill-all-meet-man-takes-risk-seriously/?fromcat=all#!zpEzt:> accessed 18 January 2021.

[2] David Rolnick and others, 'Tackling Climate Change with Machine Learning' (2019) Cornell University: Computers and Science <arXiv:1906.05433v2>.

[3] Kai-Fu Lee, 'The AI World Order.' (*Kai-Fu Lee,* 2018) <https://kaifulee.medium.com> accessed 18 January 2021.

[4] John McCarthy, 'What Is AI? / Basic Questions.' (jmc.Stanford.Edu., 12 November 2007) <http://jmc.stanford.edu/artificial-intelligence/index.html> accessed 17 January 2021.

[5] Klaus Schwab, *The Fourth Industrial Revolution* (Currency Publishing 2017).

[6] Matthijs Maas, 'International Law Does Not Compute: Artificial Intelligence and the Development, Displacement or Destruction of the Global Legal Order' (2019) 20(1) Melbourne Journal of International Law 29.

As vocalised by Stephen Hawking, "[s]uccess in creating effective AI, could be the biggest event in the history of our civilisation"[7] and Vladimir Putin, "the one who becomes the leader in this sphere will be the ruler of the world,"[8] the power of artificial intelligence is immense. From both the scientific and political realms, there is agreement that it will have significant influence on world order and power relations; AI as a topic on the international agenda is one that cannot be ignored. This is reflected in the numerous recent global initiatives that have been introduced to tackle the risks associated with AI: for example, the Council of Europe's ad hoc Committee on Artificial Intelligence (CAHAI)[9] and the Global Partnership on AI.[10] Rapid development in technology such as autonomous vehicles and weapons brings issues of security and ethics to the forefront, and governments must address their implications both domestically and internationally. The digitalisation of diplomacy is one representation of how traditional practice has evolved over the years, and the AI revolution means that it will continue to do so. With the growing influence of non-state actors (NSAs) and the unpredictability of the future capabilities of AI, it is unclear whether existing international diplomatic law is sufficient, let alone relevant. It is difficult to separate the issue of AI's impact on diplomatic practice, and the issue of AI as an international policy concern. Everyday diplomatic practice will undoubtedly be affected, but so will the broad landscape in which diplomacy takes place. Therefore, in order to effectively assess the adequacy of current legal protection, one must examine both AI's effect on diplomacy and how foreign ministries respond to this and influence its future.

There is much discussion generally on the ability of current law to accommodate for the changes brought by AI. As is the case with a lot of

---

[7] 'AI and the future of diplomacy: What's in store?' (Internet Governance Forum, 13 November 2018) <https://www.intgovforum.org/multilingual/content/igf-2018-ws-423-ai-and-the-future-of-diplomacy-what's-in-store> accessed 18 January 2021.

[8] 'Putin: Leader in artificial intelligence will rule world' AP News (Moscow, 1 September 2017) <https://apnews.com/bb5628f2a7424a10b3e38b07f4eb90d4> accessed 19 January 2021.

[9] 'Artificial Intelligence' (*Council of Europe)* <https://www.coe.int/en/web/artificial-intelligence/home> accessed 18 January 2021.

[10] 'Home.' (*The Global Partnership on Artificial Intelligence*) <https://gpai.ai> accessed 18 January 2021.

legislation, the principal legal authority for diplomatic relations, the Vienna Convention on Diplomatic Relations (VCDR), was drafted in 1961, before the extent of technological development could be envisaged. The convention does not extend protection to actors who fall outside of the traditional definition of a nation state. With the emergence of powerful multinational corporations and organisations, this strict interpretation is no longer a true reflection of global players in the diplomatic field. Furthermore, the very substance of diplomatic relations - communication - has changed significantly with the advent of the smartphone, and the constant generation of vast amounts of new information.

This paper examines the impact of AI on international diplomatic practice and whether the changes it brings will be so material that existing law is rendered obsolete. The discussion will be divided into two main themes: AI and International Diplomatic Law, and AI and global power. It is important that these issues are studied in conjunction because of the way in which they interact; International Diplomatic Law is vital in regulating power relations, and as AI influences global power, this has an impact on diplomacy. Firstly, this paper considers the time-sensitive nature of this issue and why it is so important. It will then explore the evolution of diplomacy as a result of technological development, digital and cyber diplomacy, before looking at the current state of the law in this area. After analysing both the wider shift in power relations and the direct impact on diplomatic practice that will be brought by AI, and assessing the legal implications of these, it will be concluded that there must be a balance between traditional and new methods of diplomacy. Thus, it is argued that the transformation of diplomatic practice is such that the existing legal framework is outdated. However, it is not obsolete. There will always be the need for 'old-fashioned' face-to-face diplomacy, which can be aided through the practical use of AI. Foreign ministries must cooperate with multinational tech companies to define what the desired future is to look like, and promote initiatives such as TechPlomacy,[11] elevating emerging technologies to the forefront of foreign and security policy. There must be legal reform but also the construction of

[11] 'About TechPlomacy' (*Office of Denmark's Tech Ambassador)* <http://techamb.um.dk/en/techplomacy/> accessed 19 January 2021.

supplementary soft-law to create an inclusive framework that can adapt to future developments.

2. Artificial Intelligence on the International Agenda

Despite its origins dating back to the 1950s, discussion about AI is largely absent from foreign policy agendas. General transformative technologies, "interrupt and accelerate the normal march of economic progress,"[12] and falling under this definition, AI demands immediate attention as a matter of universal importance. On the back of a digital revolution, the emergence of AI promises a strikingly greater transformation than that seen in the past; it facilitates the mechanisation of skilled as well as physical labour, meaning tasks previously requiring human cognitive ability may now be undertaken by machines.[13] Furthermore, its capabilities are not confined to industry. While the steam train was the driving force of the industrial revolution, its technological competency was limited to industry. AI systems can be implemented across a broad range of tasks, in virtually every realm, resulting in unprecedented disruption at a societal and global level.

Undoubtedly the more deeply that AI is embedded into society, the bigger the transformation of diplomacy. Universally, governments must acknowledge this and engage in a discourse about how they want AI to impact their states. The relationship between global actors and AI is reciprocal, in that the changing technological landscape will undoubtedly impact both domestic and foreign affairs but simultaneously, certain policies could also shape AI's progress. By carefully formulating policies for development and choosing how best to govern it, states can manage how AI affects not just their own territory but how other states utilise it too. Aside from the desire to be ahead of the game for economic reasons, the mass of possible new security risks mean that there is also a need for states to actively participate in this discussion, for their own safety. Furthermore,

---

[12] Lee (n 3).

[13] McKinsey Global Institute, 'Digitization, AI, and the Future of Work: Imperatives for Europe' *(Mckinsey & Company*, September 2017) <https://www.mckinsey.com/featured-insights/europe/ten-imperatives-for-europe-in-the-age-of-ai-and-automation> accessed 18 January 2021.

these risks are less likely to be addressed by the market than the opportunities.[14] If states dismiss AI as being purely technological and better dealt with by the corporations who produce AI, this neglect will be to everyone's detriment as speed and advancement could be traded off against safety.

Foreign policy itself must be distinguished from diplomacy, the former composed by governments while the latter is performed by diplomats. However, foreign ministries must also play a part in the formulation of policy and in this respect, AI as a topic on the international agenda has a direct effect on diplomatic practice. A diplomat's work crucially involves the observation and communication of developments in other states that they are based in, as well as protecting the interests of their own nationals.[15] Taking these functions into consideration, as well as the transformation of the global landscape in which diplomacy takes place, AI should be at the forefront of diplomatic practices today.

The term AI incorporates numerous processes and techniques. For the purposes of this paper, the AI referred to that will be used directly within diplomatic relations involves simple algorithmic techniques, such as those found in smartphones. Yet, speaking on a broader scale, it is the entire AI industry, and all that falls within it, that will disrupt international diplomacy in a redistribution of global power, by way of an already emerging AI arms race. The exact rate of development is uncertain and difficult to calculate. While technological evolution generally tends to be gradual, many are of the view that the AI revolution will happen much faster.[16] With one breakthrough, there could be rapid progress across a broad range of functions. Each technological advancement empowers many others, unlocking new capabilities in a sort of multi-directional chain reaction. Danzig asserts that "technology often functions as an intensifier"[17] and that the entire process of invention is simplified and

---

[14] Allan Dafoe,'AI Governance: A Research Agenda' (Future of Humanity Institute, University of Oxford, 27 August 2018) <https://www.fhi.ox.ac.uk/wp-content/uploads/GovAIAgenda.pdf> accessed 18 January 2021.

[15] Vienna Convention on Diplomatic Relations (adopted 14 April 1961, entered into force 24 April 1964) 500 UNTS 95 (VCDR) article 3.

[16] Dafoe (n 14).

[17] Richard Danzig, 'An Irresistible Force Meets a Moveable Object: the Technology Tsunami and the Liberal World Order' (2017) 5(1) Lawfare Research Paper Series.

accelerated through other technologies. Communication of new techniques and dissemination of designs can now be done instantaneously, and therefore the rate of further development will only continue to rise. Currently, AI remains 'narrow' in the sense that a system can be trained only to complete the specific task in hand. Advances in machine learning technology, a subset of AI, mean that through one common capability, a system can learn other closely linked activities. One breakthrough in this area could potentially unlock a level of general intelligence (AGI), triggering rapid universal progress in a multi-directional chain reaction. Generally associated as signifying the start of the post-human era and the concept of 'singularity,' Bostrom asserts that AGI may result in a positive feedback loop, allowing AI systems to construct other, more advanced AIs.[18]

AI technology will continue to progress at an exponential rate for which, as of yet, there is no evident limit: there is nothing to suggest that AGI will not surpass human-level intelligence.[19] Furthermore, aside from the concerns surrounding AGI, development in the field of narrow AI continues to be dramatic. It is natural human tendency that incremental change often goes unnoticed, and issues associated with current use of the technology already affect us considerably. Former President of the Supreme Court, David Neuberger, contends that the future presented by the media, diverts people's attention away from the real changes resulting from AI.[20] Maas and Stix assert that a gap exists between those scholars concerned with the short term impacts of AI and those who focus on the possible long term implications, and that this division hinders progress in the formulation of AI governance.[21] It is important not to get swept away by sensationalist conceptions of AI's potential effects and adopt a pragmatic approach going forward. The ramifications of AI are not simply future concerns,

[18] Nick Bostrom, 'How Long Before Superintelligence?' (1998) 2 International Journal of Future Studies.

[19] Stuart Armstrong, Nick Bostrom, Anders Sandberg, 'Thinking Inside the Box: Controlling and Using an Oracle AI' (2012) 22 Minds and Machines 299.

[20] David Neuberger, 'Foreword' in Jacob Turner, *Robot Rules: Regulating Artificial Intelligence* (Palgrave Macmillan 2019) vii.

[21] Charlotte Stix and Matthijs Maas, 'Bridging the Gap: the Case for an 'Incompletely Theorized Agreement 'on AI Policy' (2021) AI and Ethics <https://doi.org/10.1007/s43681-020-00037-w/> accessed 18 January 2021.

they are a matter of the present, for which international governance is lagging. To emphasise this, Turner uses the analogy of climate change, asserting that if pre-emptive measures of governance were put in place decades ago, the state of the world now could have been very different;[22] "[p]ut starkly, either we will rule the "game" or the "game" will rule us."[23]

3. Artificial Intelligence and International Diplomatic Law

3.1 Existing Legal Framework

When looking to examine the ability of International Diplomatic law to cope with emerging technologies such as AI, it is valuable to consider global governance of AI in general. Much of the legal discussion on AI has been limited in the past to issues of safety and liability, and there has been little tangible progress in AI governance.[24] Developments are beginning to emerge, commonly in the form of ethical principles. Yet while these codes are often centred around the same key trends- such as privacy, transparency and accountability,[25] they remain disparate from one another. Not only is there currently no uniform system of governance, some areas appear to lack any regulation at all. Without an international system of rules, technological development and practice will become so divergent between states that there will inevitably be conflict. However, the concept of one coherent, universal body of AI law is problematic for several reasons. With it pervading so many aspects of society, it is difficult to

[22] Jacob Turner, Robot Rules: Regulating Artificial Intelligence (Palgrave Macmillan 2019) 35.

[23] Joe McNamee, 'Governing the Game Changer - Impacts of Artificial Intelligence Development on Human Rights, Democracy and the Rule of Law' (Council of Europe High Level Conference, Helsinki, 26-27 February 2019) 1 <https://rm.coe.int/conference-report-28march-final-1-/168093bc52> accessed 18 January 2021.

[24] Dafoe (n 14).

[25] Jessica Fjeld and others, 'Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI' (2020) Berkman Klein Center Research Publication No 2020-1.

determine which areas of AI could be protected by existing law and those which fall completely outside of the current framework.

The extremely dynamic nature of AI innovation complicates this differentiation even further, and any attempt to create rigid legal definitions would be futile. Therefore, there is contradiction between the need to procure rules that can be reasonably applied to AI and the complexity of demarcating something so fluid in nature. The general definition of AI given in the introduction can be broken down further, defining 'intelligence' as the "computational part of the ability to achieve goals in the world."[26] This definition is, among others, problematic from a legal standpoint. It is elliptical in the sense that it defines 'intelligence' by way of an equally vague word, meaning that it is difficult to know exactly what is encapsulated by it. Schuett contends that there is no definition for AI that meets the requirements for legal definitions.[27] Instead, the aim is to formulate a "functional definition"[28] that allows for legal regulation but doesn't restrict the scope of protection to allow for future development.

Additionally, there is the issue of who is best placed to make the rules. Governments have the authority to formulate new laws which will be recognised as such, yet given that AI is so technologically complex, they generally lack the scientific knowledge that is required to make an effective system of governance. As Boutin suggests, new technologies do not necessarily require new laws, "legal notions are flexible and abstract enough to adapt to new scenarios",[29] perhaps AI developments can be assimilated into established legal norms. It would be incorrect to contend that all facets of AI can be encapsulated by existing legal frameworks, as it transcends so many sectors, and thus it must be looked at on a sector-specific basis.

---

[26] McCarthy (n 4).

[27] Jonas Schuett, 'A Legal Definition of AI' (2019) Cornell University Computers and Society arXiv:1909.01095v1.

[28] Turner (n 22).

[29] Berenice Boutin, 'Technologies for International Law & International Law for Technologies.' (Groningen Journal of International Law, 22 October 2018) <https://grojil.org/2018/10/22/technologies-for-international-law-international-law-for-technologies/> accessed 20 January 2021.

The main body of law in relation to diplomatic practice is the Vienna Convention on Diplomatic Relations. Despite being drafted in 1961, it has thus far been compatible with technological development. There are 191 state parties to the convention, meaning it is effectively universal, and along with customary international law it provides core protection for all diplomatic missions, premises and communications. Since ancient times, diplomacy has been recognised as the "best means devised by civilisation for preventing international relations from being governed by force alone,"[30] and it remains a fundamental concept, even in a globalised world. Underpinned by the general principles of state sovereignty and equality, the purpose of international diplomatic law is essentially to maintain good relations between states and protect peaceful interactions. While the everyday practice of a diplomat has changed over time, the basic function remains the same. Set out in Article 3 of the VCDR, the list of functions of a diplomatic mission is not exhaustive, meaning it is flexible and able to adapt to new tasks as practice changes. Conduct by diplomatic agents, if it does not come under one of the traditional diplomatic functions, may still be protected by the VCDR if it can reasonably be interpreted as being consistent with the reasoning and purpose of the convention.

At the time of drafting, those at the Vienna conference could not have predicted the progression of modern technology. While the term 'artificial intelligence' was coined in 1956,[31] there could be no comprehension of the sheer magnitude of the AI revolution. However, it is so ingrained into society that it could not be separated from diplomatic practice and thus there are many tasks related to AI that can be described as proper diplomatic functions. In order for diplomatic missions to carry out their functions, there are several fundamental principles of protection. Diplomatic agents are afforded privileges and immunities, the extent of which differ depending on their categorisation as a

[30] Ivor Roberts, 'Diplomacy - a Short History from Pre-Classical Origins to the Fall of the Berlin Wall' in Ivor Roberts (ed), *Satow's Diplomatic Practice* (7th edn, Oxford University Press 2017).

[31] Chris Smith and others, 'The History of Artificial Intelligence' (University of Washington, December 2006) <https://courses.cs.washington.edu/courses/csep590/06au/projects/history-ai.pdf> accessed 20 January 2021.

member of the mission, as defined in Article 1.[32] The premises of the mission are also protected, and this inviolability extends to the property within,[33] all archives and documents[34] and all official correspondence.[35] These provisions are essential in allowing the purposes of the convention to be realised efficiently and therefore they must be compatible with the changes brought by AI if the VCDR is to remain relevant to modern diplomatic practice.

In practice, there are several issues pertaining to the relationship between AI and the Vienna Convention. Firstly, the emergence of non-state actors, namely large tech companies, as major players on the global field. This is problematic as the VCDR does not extend protection to non-state entities or employees of such, and therefore they are not bound by the same obligations as signatories. Accordingly, this means that relations between state and NSAs are not protected in the same way as state-to-state relations. Secondly, the use of technology within daily practice calls into concern the safety of diplomatic communication in the modern day and increasingly blurs the lines between domestic and foreign affairs. Originally published in 2013, the Tallinn Manual[36] is a non-binding study prepared by a group of experts from around the world, examining the application of international law to cyber warfare. Expanding on this analysis, a version 2.0 was released in 2017 which focuses on ordinary, everyday cyber issues: 'cyber operations'. Chapter 7 of the Tallinn Manual 2.0[37] provides rules on the application of diplomatic and consular law in a cyber context. The publication attempts to apply both existing treaty and customary law to issues relating to cyberspace, a sphere which largely overlaps with AI. Although it is not legally binding, it aims to provide a resource for legal advisers across the

[32] VCDR (n 15) article 1.
[33] ibid article 22.
[34] ibid article 24.
[35] ibid article 27.
[36] Michael Schmitt (ed), *Tallinn Manual on the International Law Applicable to Cyber Warfare* (Cambridge University Press 2013).
[37] Michael Schmitt (ed), *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (Cambridge University Press 2017).

globe[38] and can be utilised as a reference for interpreting and expanding subject-specific legislation, such as the VCDR.

On many of the matters discussed in Tallinn it was not possible to reach a consensus between the experts, and the study provides a commentary on all diverging views. This disparity is indicative of the highly controversial nature of cyber and technological issues and demonstrates the difficulty in cooperating on international rules. Wide variation in state practice, technological development and societal values means that reaching agreement can be problematic. Furthermore, due to the advantages to be gained by being first-movers within the AI industry, there is a level of secrecy that is inherent to states' views on operations within this field. In order to assess the relevance of existing international law in a world of increasing AI influence, this paper examines the relevant provisions of the VCDR, with reference to Chapter 7 of the Tallinn Manual 2.0.

3.2 Technology and Diplomacy

Over the decades, although the functions and premise of diplomacy remain mainly unaltered, the context in which it is conducted has undergone several transformations. The word 'diploma' denotes an official document, and accordingly, diplomats are those who deal with these. Inviolability of the agent has long been recognised as a means of ensuring safe and effective communication. With official correspondence delivered physically to the head of another state, there was reluctance to send a delegate through foreign territory unless their safety could be assured. On a basis of reciprocity, states guaranteed safety of passage throughout their territory for envoys carrying official messages. Bilateral agreements concluded between states accorded embassies and official communications protection from invasion and interception, for the same reason of ensuring diplomatic tasks could be carried out efficiently. This network of treaties and customary law was later consolidated into multilateral conventions, the most notable of which being the VCDR.

[38] 'Tallinn Manual 2.0' (*CCDOE*) <ehttps://ccdcoe.org/research/tallinn-manual/> accessed 20 January 2021.

In the modern world, physical presence is no longer required for the communication of official messages. Diplomatic correspondence more commonly takes the form of an email rather than a *Note Verbale*[39] and in many circumstances, communication is of a much more public nature. The sphere of diplomacy has not evaded the era of digitalisation and this conversion has many practical consequences. There are elements of AI already integrated into diplomatic practice; the term AI generally prompts connotations of complex machines acting with some level of human intelligence, yet many features in devices like smartphones also come under the definition. In this sense, through digitalisation, AI has already had a great direct impact on everyday diplomatic practice: most diplomatic agents own smartphones and many embassies utilise technology. In terms of planning for the future of diplomacy and AI, it can be useful to examine the impact of digitalisation and how diplomatic practice responds. While not all of the examples here constitute direct uses of AI, such as social media, they are illustrative of the effect that technology has already had on diplomacy: an effect that will likely be exacerbated by further implementation of AI.

3.2.1 Cybersecurity Diplomacy

In practical terms, the digitalisation of diplomacy has challenged the protection provided by the VCDR. Cybersecurity is an issue relevant to many areas of law, as is demonstrated by the extensive content covered in the Tallinn Manual. It is of concern therefore for governments and policy makers worldwide on a general level, but also in particular in relation to the safety of diplomatic communications. Article 24 of the VCDR provides that "[t]he archives and documents of the mission shall be inviolable at any time and wherever they may be".[40] However the provision gives no further definition of these terms and thus from the Convention alone, it is unclear whether they extend to protect electronic archives and documents. Included in the preamble is the sentence,

[39] Patricio Grané Labat and Naomi Burke, 'The Protection of Diplomatic Correspondence in the Digital Age Time to Revise the Vienna Convention?' in Paul Behrens (ed), *Diplomatic Law in a New Millennium* (Oxford University Press 2017).
[40] VCDR (n 15) article 24.

"[a[ffirming that the rules of customary international law should continue to govern questions not expressly regulated by the provisions of the present Convention".[41] Therefore where there is ambiguity, one should look to applicable customary international law in order to fill the gaps. It is asserted that this inviolability is extended to include electronic archives and documents[42] in international practice and there was consensus in favour of this demonstrated in Rule 41, Chapter 7 of the Tallinn Manual 2.0.[43] Taking into consideration the purpose and object of the treaty, it is reasonable to conclude that they fall within the protection of Article 24, and this was affirmed by the House of Lords in 2013.[44] Furthermore, the latter part of the provision means that archives and documents will be protected even when they are not within the premises of the mission or in the custody of a diplomatic agent. This can be taken to imply that electronic documentation that is stored on a remote server is inviolable, and the experts at Tallinn suggested that archives stored on a private remote server are protected so long as they are intended to be confidential and remain undisclosed to third parties with the consent of the sending state.[45] Accordingly, as soon as information is posted in a public server, it is no longer protected.

In terms of Article 27,[46] it is recognised in customary international law that electronic modes of correspondence are included. Therefore emails, text messages and even social media interactions are all inviolable as long as they constitute official correspondence. Under this provision, the protection goes even further: there is a positive duty imposed on the receiving state to "permit and protect free communication",[47] meaning that not only is the state required to refrain from intercepting the correspondence themselves, but they must also protect it from interference by other states and non-state actors. In addition to

---

[41] ibid.
[42] Grané Labat and Burke (n 39).
[43] Schmitt, *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (n 37).
[44] *R v Secretary of State for Foreign and Commonwealth Affairs, ex parte Bancoult* (No 2) [2008] UKHL 61.
[45] Schmitt, *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (n 37) chapter VII rule 41.
[46] VCDR (n 15) article 27.
[47] ibid article 27(1).

this obligation, Rule 40[48] proposes that the receiving state is also under a special duty to protect the cyber infrastructure on the premises of the diplomatic mission "against intrusion or damage". Neither obligation is regarded as absolute, and only requires the receiving state to take "all appropriate steps" to protect the diplomatic premises and correspondence.

While these provisions go some way in targeting interference with the cyber infrastructure and correspondence of diplomatic missions, it is unrealistic that it will actually prevent it. Cyber-attacks will become increasingly prevalent and more sophisticated with advancement in AI technology. Establishing liability will likely become more difficult as machine learning abilities progress and furthermore, the number of actors with access to the technology grows. This issue, known as the 'many hands' problem,[49] stems from the concept that liability is traditionally understood in terms of individual responsibility; while it is not unique to AI, the numerous components necessarily comprised in an AI system make it a highly relevant concern. The Vienna Convention is a "self-contained regime"[50] and all available remedies for breaches of the convention are prescribed within its provisions. This means however that those who are not party to the convention cannot be held in breach of it. Although the duty to protect the premises and infrastructure therein refers to attacks from any origin, and thus the receiving state is obligated to protect against interference from non-state actors, the non-state actors themselves are not bound by the rules of the VCDR. As the future of AI technology lies largely in the hands of non-state bodies, this is problematic as it leaves gaps that may compromise the confidentiality of diplomatic cables, and ultimately undermines the functioning of diplomacy.

---

[48] Schmitt, *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (n 37) chapter VII.

[49] Karen Yeung, 'Responsibility and AI: A Study of the Implications of Advanced Digital Technologies (Including AI Systems) for the Concept of Responsibility Within a Human Rights Framework' (Council of Europe, 2019) <https://rm.coe.int/responsability-and-ai-en/168097d9c5> accessed 18 January 2021.

[50] Sanderijn Duquet and Jan Wouters, 'Legal Duties of Diplomats Today: The Continuing Relevance of the Vienna Convention' in Paul Behrens (ed), *Diplomatic Law in a New Millennium* (Oxford University Press 2017).

Due to the unpredictable and complex nature of technology, governance in this area has been largely reactive. Various discourses on what 'security' means, reflecting societal principles, and differing levels of willingness to harmonise these ideas have resulted in a "patchwork cyber governance".[51] As previously discussed, the AI revolution is gaining momentum and at an unprecedented speed. The capacity of the VCDR to encapsulate developments in cyber organisations demonstrates its flexibility and how it can be interpreted to protect new modes of diplomatic practice. However, this responsive method of governance will not suffice if we are to be prepared for the impact of AI on international diplomatic law. There must be a much more proactive approach that will take into consideration non-state actors, in order to ensure the continued efficient functioning of worldwide diplomacy.

3.2.2 Digital Diplomacy

Each new piece of technology contributes to the "acceleration of international relations"[52]; from the telegraph in the 1800s, communication became faster and easier, and in general less official. This transformation was even more drastic with the invention of the internet. While the speed of correspondence is not necessarily new, the "ubiquity of information"[53] generated by the internet age is a phenomenon with far reaching consequences worldwide. It is asserted by former Google CEO Eric Schmidt that in the present day, as much information is created every two days, as has been from the beginning of civilisation,[54] and this statistic only continues to grow. Individuals are constantly bombarded with

---

[51] Iiona Stadnik, 'Cybersecurity Diplomacy: Business and Tech Replacing the States?' (ECPR General Conference, Hamburg 2018) <https://www.researchgate.net/publication/327605493_Cybersecurity_Diplomacy_Business_and_Tech_Replacing_the_States> accessed 20 January 2021.

[52] David Paull Nickles, 'Under the Wire: How the Telegraph Changed Diplomacy' (Harvard University Press 2003) 79.

[53] Cristina Archetti, 'The Impact of New Media on Diplomatic Practice: An Evolutionary Model of Change' (2012) 7 The Hague Journal of Diplomacy 181.

[54] MG Siegler, 'Eric Schmidt: Every 2 Days We Create as Much Information as We Did up to 2003'(*TechCrunch,* 5 August 2010) <https://techcrunch.com/2010/08/04/schmidt-data/> accessed 17 January 2021.

new information in all settings of life, meaning that society is becoming more informed while simultaneously being increasingly susceptible to disinformation. People from all over the globe have the ability to group together in communities of interest, creating large information-sharing networks and consequently, gaining an audience is easier than ever.

Digitalisation and globalisation have resulted in a blurring of the lines between foreign and domestic, and diplomats are progressively engaging with populations outside of their own state. Public diplomacy refers to this interaction: effectively the antithesis of traditional diplomacy, where diplomats communicate via public statements and through the media. These expressions address both officials and the general public of the diplomat's home state but also that of other territories and would conventionally be the result of domestic political tension. However, social media has produced a new kind of public diplomacy. With a Twitter account, diplomats and world leaders can communicate directly and instantaneously with millions of people. On one hand this is a powerful tool to gather domestic support for foreign policy in a domestic context. This is particularly pertinent these days, as many challenges faced locally must be tackled on a global scale, such as climate change. Social media platforms can also be utilised to build ties with populations of other territories and diplomatic counterparts, and online interaction may be used to publicly demonstrate cooperation on certain issues.

On the other hand, as asserted by political science professor Adler-Nissen, use of social media within diplomatic practice can be dangerous. Access to social media during the negotiation process and when establishing points of collaboration, tasks that would traditionally be undertaken outside of the public eye, undermines diplomacy's "three foundational pillars".[55] Successful diplomacy is grounded in three elements: time, space and tact. Firstly, the process of negotiation requires time: a solid agreement necessitates back-and-forth proposals of ideas and redrafting. Furthermore, it demands space; there must be distance between the negotiators and also the dialogue itself, so that decisions can be made in confidentiality that best reflect both parties' interests.

[55] Rebecca Adler-Nissen, 'Behind the Scenes of Digital Diplomacy', (*Ted Talk,* 12 June 2017) <http://tedxcopenhagen.dk/talks/behind-scenes-digital-diplomacy> accessed 18 January 2021.

Finally, diplomacy fundamentally involves tact. While the formality of diplomatic communication may be considered gratuitous or outdated, a level of sensitivity and care over the phrasing of correspondence is necessary in order to reach an effective outcome. Negotiating parties often have broad cultural differences and in the context of conflict between states, protocol and tact can facilitate an agreement.

The use of social media during negotiations, and in conducting discussion itself, challenges these foundations of diplomacy. Reacting in real time can compromise the integrity of any agreement reached. This was illustrated during discussions between the EU and Ukraine in 2013 to end violent protests in Kiev, when the Polish Foreign Minister tweeted confirming the brokering of a deal from inside the negotiation room, before it had been confirmed. He was applauded by the public for seemingly consolidating the deal, however, was largely criticised by the other negotiating parties for potentially jeopardising the peace process.[56] With a public audience, the pressure to respond instantly means that the process of negotiation is rushed and likely not the best outcome possible. Moreover, with a maximum character limit to posts on many social media platforms, there is only a certain amount of tact that can be incorporated into messages and thus they can be easily misinterpreted or cause offence.

With such large proportions of states' populations now on social media, vast data sets are created. These are of high value, as data is what AI technology runs on, described as the "new oil".[57] It cannot be overlooked that both state and non-state actors can exploit this digitalisation in ways that undermine democratic and diplomatic processes, as was demonstrated very publicly in the Cambridge Analytica scandal.[58] A further concern of digital diplomacy and increased online presence is the prevalence of disinformation, and this is heightened through the capabilities of AI technology. As our data-driven

---

[56] ibid.

[57] Corneliu Bjola, Jennifer Cassidy and Ilan Manor, 'Public Diplomacy in the Digital Age' (2019) 14 The Hague Journal of Diplomacy 83.

[58] Ash New, 'Brexit: The Uncivil War Showed us how the EU Referendum was Won with Data Science.' (*Towards Data Science,* 11 January 2019. <https://towardsdatascience.com/brexit-the-uncivil-war-showed-us-how-the-eu-referendum-was-won-with-data-science-3d727ee03fc0> accessed 18 January 2021.

interactions increase in frequency, so will algorithm driven engagements and thus the potential for being fed disinformation will grow. AI systems can produce doctored or fake images, videos and online interactions mimicking human characteristics so closely that it can be very difficult to discern what is genuine. This is a challenge for both the public, who wish to be well informed, and the leaders and diplomats disseminating real information. The tangible dangers of disinformation have never been clearer, with Donald Trump's tweets deemed as having a directly causal effect on the violent January 2021 Capitol riots, leading to his account being permanently suspended from the platform.[59] Modern world leaders' fixation with conducting business in the public spotlight and the use of technology such as smartphones is putting pressure on traditional diplomatic practice. Despite the obvious risks, public diplomacy undoubtedly has its merits in building stronger networks of support both domestically and abroad, and technology can facilitate this. Transparency can be beneficial, however, there will always be a need for traditional diplomacy, away from the public eye. AI can be integrated into this more conventional diplomatic practice, and utilised in a way that targets dissemination of disinformation and exploitation of data.

Digital technology can contribute to a number of diplomatic functions. In fact, as the volume of information that must be processed in order to carry out vital tasks increases, use of technology may become essential. Public diplomacy, if utilised correctly, can rally support for diplomatic treaties which in turn may become political support. An example of this in practice is Obama engaging with the American public over Twitter to gather backing for the Iran Nuclear Agreement, resulting in Congress endorsing it.[60] The gathering, sorting and communicating of information is a fundamental function of the mission, and AI technology can be employed to expedite and enhance these tasks. Algorithms can be used to sort through large sets of information, and the data sets to underpin use of this technology within diplomatic tasks already exists in the form of legal texts. This 'text-as-data' approach could be used to both identify existing

[59] 'Twitter "permanently suspends" Trump's account' (*BBC News*, 9 January 2021) <https://www.bbc.co.uk/news/world-us-canada-55597840> accessed 18 January 2021.
[60] Corneliu Bjola and Ilan Manor, 'Revisiting Putnam's Two-Level Game Theory in the Digital Age: Domestic Digital Diplomacy and the Iran Nuclear Deal' (2018) 31 Cambridge Review of International Affairs.

and create new international law, asserts Deeks.[61] Foreign state's customary law can often be difficult to establish, and AI tools could aid in identifying this. In relation to treaty negotiations, machine learning could assist in determining the other negotiating party's preferences and past tendencies, and to predict which terms are likely to be agreed on. Inside the negotiation room, software could also be implemented to facilitate instant translation and emotion recognition. At present, there are several examples of algorithms already being employed in diplomatic practice. The Israeli Ministry of Foreign Affairs use algorithms to "map social-media bubbles" that promote certain narratives about Jewish communities and then engages with members of these online publics, providing them with factual information and building relationships.[62] Another example of this in practice is the crowdsourcing of information by the FCO in relation to the conflict in Syria, where social media was used "to listen to and identify key voices during the Libya crisis and Arab spring, thus serving as an open-source for collecting intelligence, warning of impending developments, and identifying key influencers".[63] As long as the data source is reliable, in the future, algorithms could be programmed to react in a certain way to a given scenario facilitating quick responses in cases of emergencies abroad. The concept of 'virtual diplomacy' has also been proposed, with virtual embassies: the idea holding appeal due to the expense of a proper diplomatic mission as well as the increasing difficulty of organising around changing family dynamics, security issues and diminishing diplomatic privileges in modern times.

While these manifestations of technology mays seem complex, they can be integrated simply into the existing diplomatic toolbox. The digitalisation of diplomacy could give rise to innumerable benefits, including the expedition of negotiation, ultimately leading to stable relations and peace. Yet, it is important to be mindful of the way technology is distributed worldwide among diplomatic actors, so that negotiation outcomes do not favour those with greater resources. On a general level, technology is a good platform to communicate with non-

[61] Ashley Deeks, 'High-Tech International Law.' (2020) 88 George Washington Law Review 575.
[62] Bjola, Cassidy and Manor (n 60).
[63] Jess Pilegaard, 'Virtually Virtual? The New Frontiers of Diplomacy' (2017) 12 The Hague Journal of Diplomacy 316.

state actors and form the coalitions with tech companies that are needed to create a "trusted digital environment".[64]

4. Artificial Intelligence and Global Power

4.1. The Demise of the Nation State and the Rise of Big Tech

In 2018, Apple declared their biggest annual turnover to date of 265.6 billion U.S. dollars,[65] which saw a growth of over 15% since the previous year. This statistic is dwarfed by Amazon, whose net revenue in 2019 was 280.5 billion[66], having doubled over just three years, closely followed by that of Google, Microsoft and Facebook, who along with the other two tech giants have become known as the 'frightful five'.[67] These numbers considerably surpass that of numerous countries' GDP. Amazon is about equivalent financially to Chile.[68] Despite these figures, none of these corporations are recognised by existing law, and the majority of states, as legitimate global diplomatic actors and are thus not protected by the Vienna Convention. Parties to the VCDR are all nation States as defined in Article 1 of the Montevideo Convention on the Rights and Duties of States:[69] they have a permanent population; a defined territory; a government;

---

[64] Bjola, Cassidy and Manor (n 60).

[65] Arne Holst, 'Apple's revenue worldwide from 2004 to 2020' (*Statista,* 4 January 2020) <https://www.statista.com/statistics/265125/total-net-sales-of-apple-since-2004/> accessed 18 January 2021.

[66] Tugba Sabanoglu, 'Annual net sales of Amazon 2004-2019' (*Statista,* 30 November 2020) <https://www.statista.com/statistics/266282/annual-net-revenue-of-amazoncom/> accessed 18 January 2021.

[67] Farhad Manjoo, 'Tech's 'Frightful 5' Will Dominate Digital Life for Foreseeable Future' (*NY Times*, 20 January 2016) <https://www.nytimes.com/2016/01/21/technology/techs-frightful-5-will-dominate-digital-life-for-foreseeable-future.html> accessed 18 January 2021.

[68] 'Projected GDP Ranking' (*StatisticsTimes,* 20 November 2020) <http://statisticstimes.com/economy/projected-world-gdp-ranking.php> accessed 18 January 2021

[69] Convention on Rights and Duties of States adopted by the Seventh International Conference of American States (adopted 26 December 1933, entered into force 26 December 1934) (Montevideo Convention) article 1.

and finally, the capacity to enter into relations with other states. Thus, abiding by this narrow definition, companies and IGOs cannot sign or ratify the VCDR nor enjoy its protection. Relatively recently, supranational organisations such as the UN and the EU have been recognised as international diplomatic actors with legal personality, despite not fulfilling the criteria of a nation state. While they are not covered by the VCDR, their constituent instruments confer privileges and immunities upon the organisation and personnel which are similar but not identical to that of a diplomat. The EU has 'ambassadors' in many third states and other international organisations, meaning it interacts similarly in many ways as a nation state.

Living up to their name, the power held by these tech companies is enormous, and cannot be overstated. Moreover, although this power stems from accumulation of wealth, it extends much further than being purely economic. The tech industry possesses considerable social and political influence, controlling "the infrastructures of public discourse and the digital environment for elections,"[70] and thus the mechanisms that are essential to democracy. Described by scholar Shoshana Zuboff, as "surveillance capitalism,"[71] the algorithmic model originally created to improve targeted advertisement and allowing corporations such as Facebook to gather individuals' personal data in the process, has had potentially catastrophic consequences on democracy. Private sector activities within the AI domain immensely surpass that of nation states; with South Korea's annual investment of 862 million U.S. dollars into the industry[72] completely overshadowed by the funding allocated by the 'big five', who in 2018 invested between 20 and 30 billion U.S. dollars.[73] Many NSAs now wield powers that until now were reserved to nation states. Yet, the two types of

[70] Danzig (n 17).

[71] Soshana Zuboff, *The Age of Surveillance Capitalism* (Profile 2019).

[72] Mark Zastrow, 'South Korea trumpets $860-million AI fund after AlphaGo "shock"' (*Nature*, International Weekly Journal of Science, 23 March 2016) <https://www.nature.com/news/south-korea-trumpets-860-million-ai-fund-after-alphago-shock-1.19595> accessed 19 January 2021.

[73] McKinsey Global Institute, 'Artificial Intelligence. The next Digital Frontier?' (*McKinsey & Company,* June 2017) <https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/how-artificial-intelligence-can-deliver-real-value-to-companies> accessed 18 January 2021.

entity continue to operate on increasingly divergent paths. A maintained separation between states and corporations will be massively detrimental to the future of AI and our ability to govern it in a way that balances the interests of both the public and private sectors.

4.2 A New Technology World Order?

On top of this growth in non-state power, there is an increasing shift in world order. "Inequality that comes from developments in AI and deep learning will not be contained within national borders",[74] those states who are front runners in the AI industry, currently the US and China, will jump even further into the lead. Where technology was formerly nearly exclusively military, the driver of AI is primarily commercial, resulting in a reshuffling of global markets. As it runs on a "cycle of data-driven improvements",[75] each progression accelerates further development and as states accumulate more data, one breakthrough by an actor already at the top of the market could lock in a monopoly. Typically, more socialist market economies like China, where the state has access to vast data sets, are already at a global advantage. In 2017 China announced an AI Development Plan, outlining plans to become "the world's largest economic power" through increased focus and funding given to AI, with the industry valued at $150 billion.[76] Already, the country has invested in numerous small European tech firms and start-ups as a way of "capturing innovation,"[77] as well as offering considerable benefits to employment within the AI sector.[78] China could plausibly "create an ecosystem that the rest of the world depends on,"[79]

[74] Lee (n 3).
[75] ibid.
[76] Abishur Prakash, 'The Geopolitics of Artificial Intelligence: As the U.S. and China Vie for Global Influence, AI will be Central to the Balance of Power' (Scientific American, 11 July 2019) <https://blogs.scientificamerican.com/observations/the-geopolitics-of-artificial-intelligence/> accessed 19 January 2021.
[77] Bjola, Cassidy and Manor (n 60).
[78] Fabian Westerheide, 'China – The First Artificial Intelligence Superpower' (Forbes, 14 January 2020) <https://www.forbes.com/sites/cognitiveworld/2020/01/14/china-artificial-intelligence-superpower/#2557a6b62f05> accessed 18 January 2021.
[79] Prakash (n 76).

accumulating data from abroad through the exportation of products, and channelling it back into the industry. This funnelling of technological aptitude into states that are already at the top of the market intensifies the gap in power to an even greater extent. Within the field of international law itself, Deeks points out that even if international lawyers or diplomats in certain countries remain skeptical about the benefits of utilising AI, they cannot prevent other states implementing such tools to their advantage.[80] Consequently, it is a priority for even the states that do not wish to use it. In largely industrialised countries, there is expected to be large productivity gains, flowing mainly to the capital holders.[81] As human labour value decreases this is likely to result in increased inequality and pressure on social welfare systems. Developing countries whose economies rely on cheap labour will lose this advantage and poverty will escalate further.

If the development of AI continues on this trajectory, becoming a global race, the future looks bleak. Whoever does win the AI race will have considerable influence on what AI regulation will look like, and this must be taken into account when considering which values should underpin it. While the claim that governments do not understand technology is unsubstantiated, states have been arguably naive to the power of big tech. Coupled with the private sector's wariness of centralised governance, this has created a "diplomatic deficit in the old structures of international relations"[82] that does not make sense in the current context of world power. NSAs are showing increasing capacity to take centre-stage within the global order, yet states are not demonstrating the requisite capacity to react to this power. The universal nature of AI demands a multi-stakeholder approach, that transcends national borders as well as conventional approaches to foreign policy. Despite being relatively small, Denmark was the first country worldwide to acknowledge this shift in power by appointing a Tech Ambassador in 2017.[83] This TechPlomacy initiative has a

[80] Deeks (n 61).
[81] Ryan Avent, *The Wealth of Humans: the Future of Work in the Twenty-first Century* (St Martins Press 2016).
[82] Caspar Klynge and others, 'Diplomacy in the Digital Age: Lessons from Denmark's TechPlomacy Initiative' (2020) 15 Hague Journal of Diplomacy 3.
[83] Bjola, Cassidy and Manor (n 60).

global mandate, with offices in Copenhagen, Beijing and Silicon Valley; this presence allowing for direct communication and collaboration with the tech industry. Other countries have since followed suit, with France, Taiwan, and Ireland all developing similar initiatives in recent years. These enterprises promise to bring significant benefits to small states without the technological resources of those countries at the top of the market.

AI makes this dialogue more necessary than ever and of benefit to the individual states themselves, as they can utilise industry ties to shape the future of technology in a way that conforms to their national mandate. "TechPlomacy is about putting democratically elected governments back into the equation,"[84] and offers a practical solution to ensuring the future of AI governance takes a human-centred approach. While viable in theory, this approach has not been without its pitfalls. Denmark's first tech ambassador left the post in early 2020 for a job at Microsoft, confessing that he had found it difficult to instigate "meaningful discussions" with tech corporations.[85] This points to a lack of motivation within Silicon Valley to work with states towards an ethical framework of governance for AI. It is the mandate of more recent initiatives, such as the Global Partnership on AI to facilitate the sharing of multi-stakeholder research and AI concerns, to promote the concept of "trustworthy AI."[86] Launched in 2020, with currently nineteen member states, partnerships like this one bridge the gap between government bodies and industry experts, and could well build the momentum necessary to get tech companies on board.

5. The Future

5.1 Upholding the Principles of Diplomacy

There are several obstacles to the harmonisation of AI and traditional diplomacy that must be overcome. Effective governance going forward and integration into

[84] Klynge and others (n 82) 9.

[85] Christian W, 'Denmark to get new tech ambassador' (*CPH Post Online*, 24 August 2020) <https://cphpost.dk/?p=117711:> accessed 18 January 2021.

[86] 'About' (*The Global Partnership on Artificial Intelligence)* <https://www.gpai.ai/about/:> accessed 18 January 2021.

the existing framework of diplomatic relations requires a fusion of new and old techniques: "new forms of diplomacy remain complementary to traditional diplomacy".[87] However, where new practices challenge the fundamental principles of diplomacy, they must be adapted to ensure the functions of diplomatic practice can be realised. The unprecedented transparency of the digital age must be counterbalanced with the need for confidentiality. While the public deserves a true understanding of the individuals behind negotiations and the accountability that public diplomacy provides, the importance of secret, 'back-channel' diplomacy in finalising agreements cannot be overstated. Use of AI for diplomatic tasks also requires a certain level of transparency, in order to show that its application is ethical and those implementing it must be considerate of any bias that may be produced, either through biased input data or built into the system during development. On a global level, the emergence of non-state actors into the diplomatic field is problematic for several reasons. Firstly, diplomatic relations function largely on a basis of reciprocity and this is difficult to achieve without a physical territory. This can be addressed by establishing Tech ambassadors with a physical presence, somewhat like an embassy, in the same territory as the headquarters of tech companies. Furthermore, the shift in global order resulting from the AI revolution threatens the principles of sovereignty and equality. States with large data sets or technological capabilities are at an automatic advantage when dealing with tech companies, and some states may not have the technology at all. Again, it is for this reason that smaller countries must follow Denmark's example, and acknowledge the changing landscape of diplomatic practice by generating dialogue with the private sector producers of AI technology. This is an opportunity for those states with less technological prowess to become knowledgable about the industry.

[87] Jess Pilegaard, 'Virtually Virtual? The New Frontiers of Diplomacy' (2016) 12 The Hague Journal of Diplomacy 316, 335.

5.2 AI Governance

Through the cooperation of the public and private sectors, a set of international norms can be established to supplement existing law. As AI impacts so many areas of society, the codes that govern it must consider issues of ethics, morality and politics, rather than being purely technical.[88] AI threatens to erode the state-based legal system as we know it. With tech companies operating at the same level as nation states, it is no longer correct to conclude that nation states exclusively should make international law and therefore this interaction is crucial. The dynamic and evolutionary nature of AI technology means that the establishment of strict legal definitions is futile and therefore, it must be governed by soft law. Large tech companies have already proposed regulatory frameworks for this purpose, such as the Digital Geneva Convention[89] by Microsoft, or the Tech Accord.[90] It is unlikely that there will be one streamlined code of conduct that can apply to all actors, and therefore a network of regulation, underpinned by international norms is the most appropriate form of governance for AI.

6. Conclusion

It is undeniable that the AI revolution will have a considerable impact on diplomatic practice, as it will on virtually every aspect of society. The stage in which diplomacy operates has changed substantially and will continue to do so, as AI becomes a topic on every state's agenda. Issues of security, economics and politics that emerge with technological development mean that governments simply cannot ignore AI any longer. AI's effect on diplomatic practice is two-dimensional: the employment of technology within every day diplomatic tasks and the broader evolution of diplomatic actors. Use of technology within

---

[88] Thomas Burri, 'International Law and Artificial Intelligence' (2017) 60 German Yearbook of International Law 91.

[89] Observer Research Foundation, 'Why we urgently need a Digital Geneva Convention' (*Microsoft,* 29 December 2017) <https://www.microsoft.com/en-us/cybersecurity/blog-hub/why-we-urgently-need-digital-geneva-convention> accessed 17 January 2021.

[90] 'Cybersecurity Tech Accord,' (*Tech Accord*) <https://cybertechaccord.org/accord/> accessed 19 January 2021.

diplomatic practice has both positive and negative consequences, and if it is to be beneficial in the future, diplomats must be educated on how it can be utilised appropriately. The VCDR is flexible in that it can encapsulate new diplomatic functions and be interpreted in a way that protects new technology, as demonstrated in relation to cyber operations. However, the transformation in the diplomatic landscape is radical enough that the existing legal framework is no longer sufficient. Initiatives like TechPlomacy and the Global Partnership on AI must be adopted universally if the world is to be prepared for a new AI world order. Discussions between governments and experts from the tech field can facilitate informed use of technology within diplomatic practice and furthermore can work towards establishing legal norms that reflect the interests of both the public and private sector. Ultimately, AI will not render the current law obsolete in that there will always be the need for traditional diplomacy, and thus, regulation of this. New methods of diplomacy, aided by AI, can be integrated into this conventional framework. However, the wider ramifications of AI necessitate reform of the VCDR, or new regulation to encompass non-state actors, and supplementary soft law to shape the future impact of AI on international diplomatic law.

Volume 5, Number 1, Spring 2021

Why Does Artificial Intelligence Challenge Democracy?

A Critical Analysis of the Nature of the Challenges Posed by AI-Enabled Manipulation

Caroline Serbanescu*

The Cambridge Analytica scandal has shown how Artificial Intelligence ('AI') applications can be used to influence electoral decisions. The apparent discomfort which such AI applications have in doing so triggered perhaps best illustrates the systemic nature of the disruption posed by AI applications to legal institutions in general, and to democracy in particular.

In an attempt to find the root causes of such systemic legal disruption, this paper investigates why AI applications challenge democracy pursuant to the problem-finding approach. It argues, akin to a hypothesis, that the reason why AI applications disrupt democracy lies in the new forms of manipulation which they enable and which this paper calls "AI-enabled manipulation". This paper thus presents a critical analysis of the nature of the challenges posed by AI-enabled manipulation to democracy by showing that such new forms of manipulation disrupt three of the main principles or assumptions on which democracy relies. These three democratic assumptions are citizens' autonomy, the principle of equal participation and the public forum. Throughout the paper, suggestions for regulatory responses to AI-enabled manipulation are also made. Furthermore, this paper exposes potential new problems and new regulatory concerns which its analysis generates, thereby opening to potential further research. Lastly, this paper suggests a shift of regulatory focus in the face of AI-enabled manipulation.

* LL.M. student in IP and ICT Law at KU Leuven (Belgium) [caroline.serbanescu@gmail.com]

Introduction

The Cambridge Analytica scandal has shown how Artificial Intelligence ('AI') applications such as fake news, fake accounts and algorithmic profiling can be used to influence electoral decisions. The apparent discomfort which such AI applications have in doing so triggered perhaps best illustrates the systemic nature of the disruption posed by AI applications to legal institutions in general, and to democracy in particular.

In an attempt to find the root causes of such systemic legal disruption, this paper investigates why AI applications challenge one of the foundations of many legal systems: democracy. My thesis, akin to a hypothesis, is that AI applications disrupt democracy due to the new forms of manipulation that they enable, which I call "AI-enabled manipulation". Because AI-enabled manipulation disrupts three of the main presumptions or principles on which democracy relies, these applications pose structural challenges to democracy. The approach I am taking is thus internal to individuals, focussing on AI applications' ability to shape individuals' decision-making processes, and not external to them, as I do not discuss AI applications' potential to make decisions about them.[1]

I understand democracy in this paper in its etymological sense, i.e. the rule by the people. More precisely, I focus on liberal democracy, a form of government based on citizen representation and respect for individual freedoms and choices, due to its wide application.[2] Furthermore, this paper studies why AI applications challenge citizens' representation in parliaments, and thus the legislative branch of government. It does therefore not examine legal disruption resulting from AI applications to the executive and judicial branches of government and to constitutional review. Accordingly and because of my

[1] Daniel Susser, 'Invisible Influence: Artificial Intelligence and the Ethics of Adaptive Choice Architectures' (2019) <https://dl.acm.org/doi/pdf/10.1145/3306618.3314286> accessed 18 May 2020, 403.

[2] Amartya Sen, 'Democracy as a Universal Value' (1999) 10(3) Journal of Democracy 3, 4–5; Russell Hardin, *Liberalism, Constitutionalism, and Democracy* (OUP 1999), 6; Anders Westholm, José Ramón Montero, Jan W van Deth, 'Introduction: Citizenship, Involvement, and Democracy in Europe' in Jan W van Deth, José Ramón Montero, Anders Westholm (eds), *Citizenship and Involvement in European Democracies: A Comparative Analysis* (Routledge 2007) 1, 4; William A Galston, 'The Populist Challenge to Liberal Democracy' (2018) 29(2) Journal of Democracy 5, 9–10.

etymological understanding of democracy, I decided to focus on citizens' participation in democratic processes by way of voting for their representatives. Despite that there are other ways for citizens to participate in democracy, my view is that studying the challenges posed by AI applications to elections is the most illustrative of the systemic nature of such challenges to democracy.

Furthermore, this paper takes a problem-finding approach, in contrast to the problem-solving approach generally characterising legal thinking. Pursuant to this approach, this paper seeks to find questions – not definitive solutions – arising from an ill-defined problem, AI-enabled disruption of democracy, for which there are "no known methods of solution" and if solutions are found, there are no criteria for assessing their correctness. Therefore, this paper seeks to open up to new questions in relation to AI challenges to democracy in order to help drafting regulatory responses without gaps and without generally threatening regulatory efforts.[3]

In this paper, I defend my thesis that AI-enabled manipulation is the root cause of or the thread behind AI challenges to representative democracy as it impairs three main principles or assumptions underlying democracy. These assumptions are citizens' autonomy (Section 1); the principle of equal participation in democratic processes (Section 2); and the public forum (Section 3). Section 1 also discusses AI-enabled manipulation as new forms of manipulation whereas Section 4 proceeds with a suggestion for a shift of regulatory focus.

[3] Patricia Kennedy Arlin, 'Wisdom: The Art of Problem-Finding' in Robert J Sternberg (ed), *Wisdom: Its nature, origins, and development* (CUP 1990) 230, 231, 235, 239; Hin-Yan Liu, 'From the Autonomy Framework towards Networks and Systems Approaches for 'Autonomous' Weapons Systems' (2019) 10 Journal of International Humanitarian Legal Studies 89, 89, 90, 92, 93.

1. Threat to Autonomy as Enabling Condition of Democracy

In this Section, I first discuss the new forms of manipulation enabled by AI and how such manipulation threatens citizens' autonomy (1.1). I then respond to some possible objections to my argumentation (1.2).

1.1 Unfolding the Threat

I develop the following thesis in this part: AI applications threaten democracy as they may be used to manipulate citizens' (i.e. both voters and representatives) decision-making processes, thereby threatening their autonomy as necessary precondition for participating in democratic processes. Manipulation is here understood as "imposing a hidden influence on someone's decision making".[4] In my view, AI-enabled manipulation operates as an influence external to one's cognitive processes whose hiddenness is such that the influence is unconsciously internalised in individuals' decision-making processes.

More precisely, I believe that AI applications have the potential to manipulate citizens as they are able to shape in a personalised, dynamic and concealed manner the "choice architecture" of citizens, i.e. both the set of available choices and the way they are formulated. The choice architecture thus represents the context or environment in which individuals make decisions.[5]

AI applications are able to shape our choice architectures *in a hidden manner* since these technologies have become transparent to us literally, i.e. we experience the world *through* them. Such transparency has been made possible by our increasing use of technologies in our daily activities, with the result that technologies increasingly and pervasively mediate our experiences and perceptions of the world. In other words, because we increasingly use AI applications daily, we focus on the activities facilitated by these technologies

[4] Susser (n 1) 405; Daniel Susser, Beate Roessler, Helen Nissenbaum, 'Online Manipulation: Hidden Influences in a Digital World' (2019) 4 Georgetown Law Technology Review 1, 26.

[5] Susser (n 1) 404; Susser, Roessler, Nissenbaum (n 4) 39.

instead of focussing on the technologies themselves which influence the context in which we make decisions.[6]

The ubiquity of technology mediation gave rise to Big Data, consisting in (the collection of) vast amounts of data, including personal data[7], and in the analysis thereof at high speed so as to make valuable interferences. Accordingly, Big Data subsequently led to the development of AI-enabled manipulation as the latter relies on the former for its operation. Indeed, the more that is known about each individual, the easier it is to shape her choice architecture so as to steer her decision making in the desired decision.[8]

Manipulation is in itself not a new phenomenon. It involves the exploitation of the manipulee's cognitive or affective weaknesses and vulnerabilities in order to steer her decisions towards the manipulator's ends. It does so without the manipulee's conscious awareness or in a way thwarting her capacity to become consciously aware thereof by undermining usually reliable assumptions. The exploited vulnerabilities can result from individual contingencies (e.g. habits, personality, personal history, etc.). Valuable vulnerabilities for manipulation purposes can also arise from trends emerging from the demographic groups to which each individual belongs, thereby potentially disclosing weaknesses that

---

[6] Yoni Van Den Eede, 'In Between Us: On the Transparency and Opacity of Technological Mediation' (2011) 16 Foundations of Science 139, 141, 144; Susser (n 1) 404–05; Susser, Roessler, Nissenbaum (n 4) 33–34, 38.

[7] "Personal data" is understood in this paper within the meaning of the General Data Protection Regulation: "any information relating to an identified or identifiable natural person ('data subject')" (Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L119/1, article 4 (1)).

[8] Zeynep Tufekci, 'Engineering the Public: Big Data, Surveillance and Computational Politics' (2014) <https://pdfs.semanticscholar.org/04e2/f184505a6b67c611bc57c05864385c024418.pdf?_ga=2.34130535.792266196.1589824767-2048156165.1589824767> accessed 18 May 2020, 20; Karen Yeung, 'Algorithmic Regulation: A Critical Interrogation' (2018) 12 Regulation & Governance 505, 514.

individuals themselves do not see.[9] Accordingly, identifying the group vulnerabilities of individuals allows to refine manipulation.

In my view, the unprecedented threat of AI applications to democracy stems from the new forms of manipulation which they enable.

There is a twofold qualitative change brought about by AI-enabled manipulation in contrast to "analogue" manipulation (i.e. not involving AI). On the one hand, AI-enabled manipulation allows for *tailored* influences over decision-making processes due to technology's pervasiveness in – or even surveillance of – daily lives. Such pervasiveness thus allows to more easily identify and subsequently exploit individuals' weaknesses arising from both individual and group contingencies. On the other hand, AI-enabled manipulation is *dynamic* or *adaptive* in the sense that it can adapt and refine itself quickly in the light of individuals' conducts notably on the internet, which can reveal changes of preferences.[10]

Furthermore, AI-enabled manipulation involves a quantitative change due to the unparalleled reach of these new forms of manipulation. Combined with the possibility to personalise and dynamically change each individual's choice architecture, the wide reach of AI-enabled manipulation can enable its exploiter to more effectively reach her goal. Indeed, if a large number of voters manipulated by AI applications votes in the direction reflecting the manipulator's interests, such interests would likely be achieved as they would represent "the will of the people".[11]

Overall, the rise of AI applications has thus led to more effective forms of manipulation.

—

In my view, AI-enabled manipulation exacerbates the known threat of manipulation to citizens' autonomy, which thereby also constitutes a threat to democracy (*infra* 5). Autonomy is here understood as the capacity to "rule oneself" or to act independently on the basis of one's own reasons that one

[9] Karen Yeung, 'Hypernudge: Big Data as a Mode of Regulation by Design' (2017) 20 Information, Communication & Society 118, 122; Susser, Roessler, Nissenbaum (n 4) 3, 26, 32.

[10] Susser (n 1) 404; Susser, Roessler, Nissenbaum (n 4) 3.

[11] Susser, Roessler, Nissenbaum (n 4) 4, 29.

recognizes and endorses.[12] In fact, I believe that the presence of manipulation always implies a threat to one's autonomy as imposing a hidden influence on one's decision making leaves uncertain as to this person's (remaining) autonomy. Consequently, just as AI-enabled manipulation is the thread of my argumentation, so is the resulting threat to autonomy.

AI-enabled manipulation exacerbates the undermining of citizens' autonomy due to its quantitative and qualitative improvements (*supra*) as well as its hiddenness. More precisely, such manipulation challenges the following two features of autonomy generally recognized by autonomy theorists.

On the one hand, autonomous persons have the "cognitive, psychological, social, and emotional *competencies* [or capacities] to deliberate, to form intentions and to act on the basis of that process".[13]

On the other hand, "autonomous persons can (at least in principle) critically reflect on their values, desires, and goals, and act for their own reasons, i.e. endorse them *authentically* as their own". The latter condition amounts in fact to individuals' capacity for self-authorship over their actions.[14]

AI-enabled manipulation threatens both autonomy features. On the one hand, such manipulation impairs autonomous individuals' capacity to *competently* deliberate as it designs the features of individuals' choice architecture in a concealed, tailored and adaptive manner so as to influence them without their conscious awareness. On the other, such influence steers citizens to act, such as to vote for a candidate, for reasons they cannot understand, as they are not their own, and therefore cannot *authentically* endorse as their own.[15]

As a counter argument to the above, it may be contended that citizens do not always make decisions on the basis of reasons. Indeed, when citizens do not know what to decide, even for important decisions such as voting decisions, they may

---

[12] Andrew Sneddon, *Autonomy* (Bloomsbury UK 2013) 3; Susser (n 1) 406–07.

[13] John Philip Christman, *The Politics of Persons: Individual Autonomy and Socio-Historical Selves* (CUP 2009) 154–55; Yeung, 'Hypernudge …' (n 9) 124; Susser, Roessler, Nissenbaum (n 4) 36.

[14] Sneddon (n 12) 7; Yeung, 'Hypernudge …' (n 9) 124; Susser, Roessler, Nissenbaum (n 4) 18, 36.

[15] Cass R Sunstein, *The Ethics of Influence: Government in the Age of Behavioral Science* (CUP 2016) 83; Susser (n 1) 406-07; Susser, Roessler, Nissenbaum (n 4) 38.

"pick" a decision simply because they "felt like it".[16] However, in that case, I believe that AI-enabled manipulation may still represent a threat to citizens' autonomy and hence to democracy. Indeed, AI-enabled manipulation may perceive this lack of rational deliberation as a cognitive vulnerability which it can exploit in order to drive manipulees to cast the desired vote. This is because AI-enabled manipulation has the unparalleled potential to inculcate the manipulees reasons – the most appealing to them as made possible by targeted manipulation – to make the decision preferred by the manipulator.

By undermining citizens' autonomy, I believe that AI-enabled manipulation endangers the foundations of liberal representative democracies.

The core idea of liberal representative democracies is that "political power derives its authority from the autonomous consent of the governed".[17] Accordingly, democracy presupposes that the governed are politically autonomous, which is the case when they are sufficiently able to participate in democratic processes in a way that reflects their capacity to self-rule.[18]

However, with AI-enabled manipulation, the manipulated citizens may have given a consent to political power which does not reflect their capacity to self-rule as they may not understand and hence endorse the reasons backing their consent. Therefore, they may not have given a consent reflecting their own will but rather that of the manipulators. In that case, votes do not represent the will of the people, which conflicts with the core of representative democracy as it requires Parliaments to express the will of the people and the public interest. Ultimately citizens' acknowledgment of political institutions, especially the Parliament, as their own may be undermined.[19] Moreover, the (input) legitimacy of the representatives, which is "based on the assumption that political choices are legitimate if and because they reflect the will of the people", may be

[16] Sunstein (n 15) 66.
[17] Sneddon (n 12) 4.
[18] Westholm, Montero, van Deth (n 2) 6–7; Sneddon (n 12) 3–4, 7.
[19] Lawrence Pratchett, 'The Core Principles of European Democracy' in *Reflections on the Future of Democracy in Europe* (Council of Europe Publishing 2005) 31, 33; Susser (n 1) 406.

impaired.[20] The same legitimacy problem occurs if representatives are manipulated when making decisions as in that case, their decisions do also not reflect the will of the people but that of the manipulator (*infra*).

In practice, AI-enabled manipulation means that AI applications can, for instance, influence in a tailored, dynamic and hidden way the electoral information each individual receives on social media or when making internet searches. In doing so, AI applications shape each individual's choice architecture in a unique way in the sense that no voter receives the same electoral information (*infra*).[21] As a result, citizens may be led to vote for a candidate for whom they would not necessarily have voted in the absence of influence or to abstain from voting while they would have perhaps voted failing the interference (*infra*). Whether individuals decide to vote or abstain from voting while being influenced, the chosen decision may not always be the most beneficial to each citizens' interests.

Accordingly, besides what was stated above, AI-enabled manipulation further undermines individuals' autonomy as autonomous citizens usually act in their own interest for the sake of enhancing their own welfare. If citizens are not able to make voting choices reflecting their own welfare, their interests and will cannot be served by their elected representatives, as they are not made aware thereof. As a result, democracy is further undermined as elected representatives cannot express the will of the "people" in their legislative activities.[22]

The same threat to representative democracy arises in relation to elected representatives' manipulation. AI applications may indeed shape representatives' decision-making processes so as to induce them to make legislative choices

---

[20] Magdalena Godowska, 'Democratic Dilemmas and the Regulation of Lobbying - the European Transparency Initiative and the Register for Lobbyists' (2011)14 Yearbook of Polish European Studies 181, 183.

[21] Jonathan Zittrain, 'Engineering an election' (2014) 127 Harvard Law Review 335, 336, 340; Tufekci (n 8) 26; Vyacheslav Polonski, 'How Artificial Intelligence Silently Took Over Democracy' (2017) World Economic Forum <www.weforum.org/agenda/2017/08/artificial-intelligence-can-save-democracy-unless-it-destroys-it-first/> accessed 19 May 2020; Karl M Manheim, Lyric Kaplan, 'Artificial Intelligence: Risks to Privacy and Democracy' (2019) 21 Yale Journal of Law and Technology 106, 147–50.

[22] Susser (n 1) 406.

beneficial to the AI exploiter's interests. Therefore, the ultimate decision made by representatives may again reflect the manipulator's interests, and not citizens' interests.

The new forms of manipulation to which AI applications give rise thus threaten the way citizens participate in decision-making processes and the way public authority is exercised.[23]

___

An objection to this argument could be that manipulation rarely *totally* deprives its target of autonomy because a manipulator never fully controls her. Indeed, in contrast to coercion, which is another way of influencing decision-making processes, manipulation does not imply the entire displacement of the target as the decision maker by way of compulsion. Manipulation involves in fact a more subtle insinuation into the target's decision-making processes as it impairs her capacity for conscious decision making. In other words, the manipulee makes a decision while not *fully* understanding why she took this decision or whether it served her own or someone else's interests.[24] In my view, such argument implies that if manipulated citizens cast a vote, this vote can still be deemed to reflect citizens' own will as it is "saved" by citizens' remaining autonomy. This remaining autonomy could indeed have steered the manipulees to cast a vote different than that preferred by the manipulator or could have deemed the manipulator's preferred decision to be in these manipulees' best interest, without being aware of the manipulation. Accordingly, the threats to democracy would disappear.

I submit the following counter arguments which qualify such objection.

On the one hand, as already stated (*supra* 5), it is common for people to lack a full understanding of the reasons for their choices, even without being manipulated, as many of them are based on unconscious processing (so-called "System 1"). Indeed, System 1 of processing information is intuitive and prone

---

[23] Snežana Samardžic-Markovic, 'AI and Democracy' (High-level Conference on 'Governing the Game Changer – Impacts of Artificial Intelligence Development on Human Rights, Democracy and the Rule of Law', Helsinki, 26–27 February 2019) <https://rm.coe.int/-artificial-intelligence-and-democracy-introductory-speech-by-snezana-/1680933353> accessed 19 May 2020.

[24] Susser, Roessler, Nissenbaum (n 4) 17.

to bias. I have already argued before that such lack of rational deliberation can be exploited by manipulators. Nevertheless, it may be that manipulators do not perceive such lack of rational deliberation and that citizens actually act or vote intuitively. This would imply that regulatory responses to AI-enabled manipulation should distinguish between situations where our actions are determined by our own impulses and situations where they are determined by factors intentionally framed by others to influence our decision making.[25] This also begs the question, deserving further research, as to what legal value should be awarded to intuitive votes given that they are hardly identifiable but that, from a democratic viewpoint, they do not seem desirable. From the latter viewpoint, intuitive votes do indeed not ensure that voters truly express their will, which should become apparent after a rational deliberation involving a consideration of all candidates' positions.

On the other hand, I believe that distinguishing between full and partial undermining of autonomy leads to the following difficulties and hence does not rule out concerns about democracy.

Firstly, it is difficult, if not impossible, to quantify the extent to which manipulation impairs individuals' autonomy. Therefore, it seems more appropriate, from a regulatory perspective, to shift the focus from the extent of autonomy impairment to the presence of AI-enabled manipulation, which is thereby deemed reprehensive as such.[26] Accordingly, any interference with one's decision-making processes is problematic, regardless of the extent of autonomy impairment, precisely because it creates uncertainties as to the extent of individuals' remaining autonomy.

Secondly, I believe that the view that the manipulee does not fully understand the reasons for her decisions or whether it served her interests presupposes that the manipulee suspects afterwards to have been subject to some kind of undue influence. I believe that AI-enabled manipulation may be so effective that it is difficult, if not impossible, for manipulees to even suspect such influence afterwards *by themselves*. In other words, AI-enabled manipulation may give the

[25] Sunstein (n 15) 89-90; Mark Egan, *An Analysis of Richard H. Thaler and Cass R. Sunstein's Nudge: Improving Decisions about Health, Wealth and Happiness* (Macat Library 2017) 35.

[26] In that sense, see Susser, Roessler, Nissenbaum (n 4) 41.

impression to the target that she votes autonomously. It does so by subtly influencing the target's decision making in a such a personalised and dynamic way that the target may unknowingly believe the manipulator's interests to be her interests and that she fully understands the reasons for her votes.[27]

In that sense, I believe that manipulees cannot become aware of AI-enabled manipulation by themselves but through the help of external actors. Thus far, only the manipulator (or individuals from the group of persons behind the manipulation) or persons having been in touch with the manipulator (without participating in the manipulation) can disclose AI-enabled manipulation.[28] However, the questions remain on how legislation could be drafted so as to incentivise such persons to disclose AI-enabled manipulative practices and on what other steps or actors States could take or involve to facilitate such disclosure, especially in the electoral context. Such questions fall out of the scope of this paper but warrant further research which I strongly encourage.

The Cambridge Analytica scandal is a good illustration of my thought. Until the disclosure of the hidden AI-enabled manipulation to which millions of Facebook users have been subject, I do not think that many users questioned their voting decisions. Instead, several actors, including researchers who had been contacted by Cambridge Analytica but did not participate therein, have disclosed the practices of this firm to newspapers. According to one of these researchers, it was strong ethical opposition that led him to disclose the AI-enabled manipulation.[29] Nevertheless, in line with my above suggestion, further

---

[27] In that sense, see Manheim, Kaplan (n 21) 109, 150.

[28] I however do not exclude the possibility that in future more and more journalists will be able to reveal AI-enabled manipulation, notably through the command of AI applications (although such tools currently have inherent limitations (*infra*)).

[29] See statement of one of the researchers having contributed to the disclosure of the scandal: Michal Kosinski, 'Statement on Cambridge Analytica' (2018) <https://drive.google.com/file/d/1zRaTAx0mpRC0m7-3wQRaDPYTOGMdvNBt/edit> accessed 19 May 2020. See one of the first press articles disclosing the scandal: Harry Davies, 'Ted Cruz using firm that harvested data on millions of unwitting Facebook users' *The Guardian* (11 December 2015) <https://www.theguardian.com/us-news/2015/dec/11/senator-ted-cruz-president-campaign-facebook-user-data> accessed 19 May 2020.

research could focus on whether other, legally enforceable and less subjective, incentives to disclose AI-enabled manipulation could be sustained by law.

In the light of the above, I conclude that by undermining citizens' autonomy, AI-enabled manipulation has the potential to transform citizens into passive actors (or "puppets").[30] Consequently, citizens are no longer democratic agents as undermining their autonomy implies that they can no longer participate in democratic processes so as to express their will. As a result, democracy as the rule by the people is undermined. Another related consequence in my view is the erosion of the notion of the law, as the incarnation of the will of the people, since the latter's existence is undermined.[31] This begs the question, deserving further research, as to whether there are other reasons why AI applications disrupt our understanding of the law.

1.2 Possible Objections

Besides the previous objections stated above, I here respond to further potential and more general objections to my above argumentation.

Firstly, some may contend that it is difficult, if not impossible, to assess if citizens are being manipulated. I agree with this objection and I have already explained above why I believe that regulatory efforts should focus on the presence of AI-enabled manipulation as such. A further suggestion building thereupon is for regulatory efforts to take a general instead of an individual approach in the face of AI-enabled manipulation. Hence, instead of scrutinizing individual cases, focussing on and examining the manipulation strategy at a general level could be more effective and less time-consuming as it could allow to reveal individual instances of manipulation.[32] This would notably require from regulators to qualify a given attempt of influencing citizens as AI-enabled manipulation. Such qualification exercise however requires the existence of a

---

[30] Susser, Roessler, Nissenbaum (n 4) 17; Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (PublicAffairs 2019) 21 ff.

[31] In that sense, see Anthony J Casey, Anthony Niblett, 'The Death of Rules and Standards' (2017) 92 Indiana Law Journal 1401.

[32] Sunstein (n 15) 86; Susser, Roessler, Nissenbaum (n 4) 41.

legal definition of AI-enabled manipulation which could include its constitutive elements (e.g. dynamic, tailored, hidden, large scale, etc.).

Secondly, one may question whether AI-enabled manipulation really threatens democracy when a manipulee would have casted the same vote without manipulation. In my opinion, AI-enabled manipulation remains problematic for democracy in that case. Indeed, I believe that it is only a coincidence if the manipulee's autonomous vote equates the vote she casts when manipulated. Hence, it is not the outcome of the manipulation that counts but rather the interference into one's decision making process as it leaves uncertain as to the manipulee's autonomy (*supra* 7).[33] Since AI-enabled manipulation can change the set of available electoral options and their understanding, it may shape the reasons why manipulees vote for a certain candidate, hence affecting the authenticity component of autonomy. In other words, the reasons why manipulees cast the same vote without and under manipulation may differ and the manipulees may thus not authentically endorse the reasons for their manipulated vote. Therefore, concerns for democracy do not disappear.

Thirdly, one may argue that as choices in general, including votes, are always influenced and conditioned by social, cultural, economic and political contexts, it is difficult to distinguish these contextual influences from the manipulator's influences.[34] Nevertheless, such contextual influences can be exploited by AI-enabled manipulation as group vulnerabilities of each citizen, thereby allowing to better tailor the manipulation (*supra* 4). As a result, there is perhaps no need to distinguish between these contextual influences and AI-enabled manipulation as the latter may incorporate the former in its process of shaping individuals' decisions.

Fourthly, another objection could be that it is difficult to distinguish the effect, on individuals' decision making, of AI-enabled manipulation from that of analogue manipulation. I agree with that objection but perhaps in certain cases, the exploiters of AI applications are the same persons who seek to manipulate voters via analogue means. By combining different forms of manipulation, manipulation is indeed rendered more effective. Therefore, in

[33] Susser, Roessler, Nissenbaum (n 4) 42.
[34] Yeung, 'Hypernudge …' (n 9) 129; Susser, Roessler, Nissenbaum (n 4) 42–43.

these cases, there is perhaps no need to distinguish between analogue and AI-enabled manipulation. Furthermore, such difficulty could perhaps be overcome if the focus is again placed on the presence of analogue and AI-enabled manipulation, which would allow to distinguish them. If that is the case, regulatory responses should in my view take account of the fact that AI-enabled manipulation is more effective in achieving the desired outcome than analogue manipulation notably due to its ability to constantly personalise manipulation (*supra* 4). Therefore, manipulators using AI applications to achieve their ends should bear a larger or special legal responsibility for the more effective threat to democracy they pose than manipulators using analogue means.[35]

2. Threat to Equal Participation

Apparent from the previous Section is that AI-enabled manipulation disrupts liberal representative democracy as it has the potential to eliminate any form of democratic participation by citizens. Such elimination would in fact be concealed as citizens would still physically vote without having the cognitive capacities or autonomy to do so. Another – less radical – challenge of AI-enabled manipulation to democracy is the potential to disrupt the democratic principle of equal participation. According to this principle, each citizen has the same ability to express her will to representatives.[36]

In my opinion, both the elimination of participation and unequal participation as a result of AI-enabled manipulation can occur at the same time. As the elimination of participation cannot currently reach all members of an electorate (as it notably depends on *all members* accessing the internet, which is not the case of e.g. some elderly), if some members are not cognitively able to

[35] In that sense, see Susser (n 1) 407.

[36] Sydney Verba, 'Political Equality. What is It? Why Do We Want It?' (Review Paper for Russell Sage Foundation 2001) <https://www.russellsage.org/sites/all/files/u4/Verba.pdf> accessed 19 May 2020, 2; Westholm, Montero, van Deth (n 2) 3; Jan Teorell, Paul Sum, Mette Tobiasen, 'Participation and Political Equality: an Assessment of Large-Scale Democracy' in Jan W van Deth, José Ramón Montero, Anders Westholm (eds), *Citizenship and Involvement in European Democracies: A Comparative Analysis* (Routledge 2007) 384, 385.

participate anymore in democratic processes, it is likely that equality in participation is impaired. Indeed, other members not reached by AI-enabled manipulation retain their cognitive capacities to participate in such processes.

There are in my view at least two ways in which AI-enabled manipulation may threaten equality in participation.

On the one hand, as mentioned, AI-enabled manipulation relies on large collection of personal data found notably but not exclusively on the internet. However, not all individuals leave the same amount of personal data on the internet as such amount notably depends on membership to (several) social networks and the number of internet searches initiated by each individual. Therefore, I believe that there are different degrees of manipulation to which AI applications may lead depending on the data that the algorithm finds on each individual. As a result of these different degrees of manipulation, there are different extents to which the manipulee's autonomy are impaired so that some individuals may be more vulnerable to AI-enabled manipulation than others. This means that AI-enabled manipulation's effect on some individuals' decision making may not be such as to hinder their capacity to express a will or vote that they endorse (*supra*). Accordingly, equality in participation is undermined due to these different degrees of autonomy impairments. In that relation, such different degrees of AI-enabled manipulation may also reflect discrimination against protected classes, thereby further leading to unequal participation. Since there is more data produced on such classes than on non-protected classes, the former are more vulnerable to AI-enabled manipulation, hence less likely to express autonomous choices.[37] However, I am aware that it may be impossible to quantify the interference of AI-enabled manipulation on citizens' autonomy and there will therefore always be an uncertainty in that regard. To overcome

[37] For instance, in the US, many communities of colour are more thoroughly surveilled than white communities (Alvaro M Bedoya, 'The Color of Surveillance: What an infamous abuse of power teaches us about the modern spy era' *Slate* (2016) <https://slate.com/technology/2016/01/what-the-fbis-surveillance-of-martin-luther-king-says-about-modern-spying.html> accessed 19 May 2020; Susser, Roessler, Nissenbaum (n 4) 40–41). Tufekci (n 8) 3; Anja Bechmann, 'Data as Humans: Representation, Accountability, and Equality in Big Data' in Rikke Frank Jørgensen (ed), *Human Rights in the Age of Platforms* (MIT Press 2019) 73, 77–78.

such difficulty, I reiterate that regulatory efforts should focus on the presence of manipulation, which should be deemed reprehensive as such.

On the other hand, in countries where voting is not compulsory, AI-enabled manipulation may be used to deter voters from certain (minority or vulnerable) groups to vote, i.e. suppress their votes. Such practice creates an inequality in the possibility of diverse voter groups to express their will in the form of votes as groups subject to AI-enabled manipulation are deprived from such possibility.[38] Accordingly, equality in participation is disrupted.

Attempts from a foreign government to suppress votes of certain (protected) groups using AI applications have actually already occurred in the 2016 US presidential elections.[39]

A problem for democracy arising from AI applications disrupting equality in participation is the undermining of the legitimacy of parliamentarians. Indeed, if not all voters had an equal chance to vote for their representatives, the latter cannot be said to represent and act in accordance with the "will of the people". A further problem is that such disruption is likely to lead to the "tyranny of the majority", whereby the majority, possibly representing the interests of the manipulator (*supra* 6), uses its political power to serve its own interests at the expense of the rights of others and of the public good.[40]

---

[38] Elaine Kamarck, 'Malevolent Soft Power, AI, and the Threat to Democracy' (Brookings Report 2018) <www.brookings.edu/research/malevolent-soft-power-ai-and-the-threat-to-democracy/> accessed 19 May 2020.

[39] The Russian government has been found to have interfered in the 2016 US Presidential election. More precisely, Russia, supportive of republican candidates, used AI applications (e.g. fake accounts, fake news, etc.) to suppress Afro-Americans' votes, which have historically leaned towards democratic candidates (Alec Tyson, Shiva Maniam, 'Behind Trump's victory: Divisions by race, gender, education' (Pew Research Center 2016) <www.pewresearch.org/fact-tank/2016/11/09/behind-trumps-victory-divisions-by-race-gender-education/> accessed 19 May 2020; Kamarck (n 38); S Mueller, 'Report on the Investigation into Russian Interference in the 2016 Presidential Election' (US Department of Justice 2019) <www.justice.gov/storage/report.pdf> accessed 19 May 2020, 14, 25).

[40] *Osmanoğlu and Kocabaş v Switzerland* App no 29086/12 (ECtHR, 10 January 2017) para 84; Ronald J Terchek, Thomas C Conte, *Theories of Democracy: a Reader* (Ringgold Inc 2002) 5; Verba (n 36) 3-4; David Held, *Models of Democracy* (Polity Press 2006) 72; Van Den Eede (n 6) 153.

3. Threat to Public Forum

Today, much political and civic speech takes place online, where AI applications such as fake news and Deepfakes flourish.[41] These applications constitute disinformation strategies as well as concrete examples of AI-enabled manipulation and are in my view the most illustrative of the nature of the challenges of AI applications to the public forum. These disinformation strategies consist in disseminating deceptive information and/or showing recipients the information that they are mostly likely to endorse.

Such disinformation strategies have always existed in the electoral context but AI allows them to more effectively deceive their recipients. This is especially the case of Deepfakes, which consist in altering audio and video messages so as to deceive their recipients' senses. In doing so, the latter are induced to believe that the information conveyed is accurate. Such AI applications can thus be used in order to discredit electoral candidates, so as to incite voters to cast their votes in favour of the manipulator's preferred candidate, or to discredit elected representatives. Other AI applications contributing to disinformation are the algorithms operating many online platforms and search engines which can determine the visibility of political content. Accordingly, groups which cannot afford to rely on such algorithms will be more and more hidden from public view, or there will be changes in their reach that are beyond their control.[42]

These disinformation strategies are problematic for democracy mainly because they disrupt the public space for deliberation. The unparalleled disruption of AI applications to the public forum stems in my view from the targeted disinformation deployed on each individual, whereby all citizens receive tailored and adaptive, hence necessarily different, electoral information. As a result, such strategies alter the set of information available in the public forum and required for citizens to vote autonomously, i.e. to cast enlightened votes, and to participate in public debates. Indeed, such disinformation strategies make it difficult, if not impossible, for citizens to distinguish official and accurate

[41] Tufekci (n 8) 25.

[42] Tufekci (n 8) 26; Kamarck (n 38); Manheim, Kaplan (n 21) 137, 142, 146-48; Birgit Schippers, 'Artificial Intelligence and Democratic Politics' (2020) 11(1) Political Insight 32 33.

information from fake news or to grasp *all* information which are part of the public forum. This implies that such disinformation strategies do not comply with democracy's requirement that, during election times, the different sides have an adequate opportunity to present their respective cases, and the electorates have the freedom to obtain news and to consider the views of *all* competing parties. Moreover, such practices may cause social polarisation within the public forum, with the impact on democracy being the formation of distinct groups that can no longer understand each other and hence making the reaching of political compromises almost impossible.[43]

Further AI applications altering the democratic public debate and amounting to AI-enabled manipulation are fake accounts or bots. These applications indeed instigate the belief that citizens are engaging in constructive dialogues with fellow citizens whereas in reality they do not discuss with "real" people and the genuine purpose of such dialogues is to manipulate citizens. Such dialogues may be used in order to change citizens' opinion on electoral candidates with the ultimate purpose being to steer citizens to vote for the candidate preferred by the bots' exploiters. Because fake accounts alter the public forum in the same manner as disinformation strategies, they are collectively referred to in this paper as "AI-enabled alteration" strategies.[44]

In the face of new disruptive technologies, regulatory responses could be to apply by analogy laws regulating allegedly similar phenomena. Hence, the question arises whether AI-enabled alteration strategies can be equalized with the (not necessarily condemnable) practice of propaganda. In my view, they cannot.

Propaganda refers to "an organised effort to spread a particular doctrine or belief" on a large scale.[45] Both propaganda and AI-enabled alteration can be deployed not only by political parties to promote their candidates but also by

---

[43] Sen (n 2) 9-10; Tufekci (n 8) 26; Sunstein (n 15) 45, 65; Dirk Helbing and others, 'Will Democracy Survive Big Data and Artificial Intelligence?: Essays on the Dark and Light Sides of the Digital Revolution' (2019) < www.researchgate.net/publication/327271384_Will_Democracy_Survive_Big_Data_and_Artificial_Intelligence_Essays_on_the_Dark_and_Light_Sides_of_the_Digital_Revolution> accessed 19 May 2020, 5; Manheim, Kaplan (n 21) 150; Schippers (n 42) 33.
[44] Polonski (n 21); Kamarck (n 38); Manheim, Kaplan (n 21) 151.
[45] Edward L Bernays, *Propaganda* (Horace Liveright 1928) 20.

other actors willing to interfere with elections such as foreign states (*supra* 11).[46] AI-enabled alteration raises however more profound concerns for democracy than propaganda, which would justify a distinct regulatory effort. A first set of concerns relates to the fact that contrary to propaganda, AI-enabled alteration is personalised and adaptive to each recipient. Another set of concerns regards the fact that, in contrast to propaganda, AI-enabled alteration and more generally AI-enabled manipulation rely on widespread surveillance of citizens for their operation. As a result of such widespread surveillance, firms developing AI-enabled manipulation possess unparalleled power of behavioural modification.[47]

In my view, such unparalleled power which these firms possess *risk* in the long term leading to the defeat of democracy over "technocracy". Indeed, these firms' expertise in surveillance technologies could allow them to manipulate citizens and their representatives to such an extent that they or their clients would *de facto* be running States.[48] Only a perception of democracy would remain as citizens would continue to vote whereas their voting decisions would not be autonomous and fully informed as a result of AI-enabled manipulation and alteration.

My reasoning however begs questions as to whether such a society would be achievable and sustainable notably given the conflicting interests of the various firms concerned and/or of their clients.

Furthermore, for the sake of clarity, my view is not that the displacement of democracy by technocracy *will* occur in future. My point seeks rather to open

---

[46] Karl E Ettinger, 'Foreign Propaganda in America' (1946) 10(3) The Public Opinion Quarterly 329, 329; Kamarck (n 38).

[47] Yeung, 'Hypernudge …' (n 9) 130. See also Zuboff, *The Age of Surveillance Capitalism* … (n 30); Shoshana Zuboff, '"We Make Them Dance": Surveillance Capitalism, the Rise of Instrumentarian Power, and the Threat to Human Rights' in Rikke Frank Jørgensen (ed), *Human Rights in the Age of Platforms* (MIT Press 2019) 3.

[48] Marc Hudson, 'Ending Technocracy with a Neologism? Avivocracy as a Conceptual Tool' (2018) 55 Technology in Society 136, 136-37; Yeung, 'Algorithmic Regulation …' (n 8) 518; Julie E Cohen, *Between Truth and Power: The Legal Constructions of Informational Capitalism* (OUP 2019) 3. In that sense, see J Benjamin Hurlbut, 'Laws of Containment: Control Without Limits in the New Biology' in Irus Braverman (ed), *Gene Editing, Law, and the Environment: Life Beyond the Human* (Taylor & Francis 2017) 77, 86–91.

up to a potential challenge to democracy to be taken into account by regulatory responses to AI-enabled manipulation and alteration.

4. A Necessary Shift of Focus?

In the light of the above, perhaps the challenges posed by AI applications to representative democracy revolve around the use made of such technology. Indeed, it may be because humans use AI applications for illegitimate ends from a democratic perspective, i.e. manipulation, that democracy and its core principles are disrupted. This would imply that the regulatory target should shift from AI as a technology to human use of such technology.

An illustration of my argument is the fact that the threats posed by AI to democracy can possibly be solved *in future* by using AI applications themselves since several of them are already being used to detect and remove undesirable content online. However, as mentioned, such use of AI applications to remedy their threats to democracy will possibly only be achievable in future due to the current limitations of these technologies, such as intrinsic biases or the risks of unduly blocking desirable content.[49]

A likely objection to my proposal of targeting human use of AI could be that it is not because of human use of technologies that AI poses challenges to legal institutions in general, and democracy in particular, but because AI develops in an unforeseeable manner. Nevertheless, in my view, the focus should still be placed on human use of AI since humans may have intended the unforeseeable developments of AI. Indeed, it is likely that because AI applications have the capacity to generate unique solutions not considered by humans and potentially improving humans' lives, these technologies' exploiters have an incentive to develop AI systems which can generate unexpected solutions.[50]

---

[49] See Katarina Kertysova, 'Artificial Intelligence and Disinformation: How AI Changes the Way Disinformation is Produced, Disseminated, and Can Be Countered' (2018) 29 Security and Human Rights 55, 59–61.

[50] Matthew U Scherer, 'Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies' (2016) 2 Harvard Journal of Law & Technology 353, 364–66.

Regulatory efforts could accordingly place responsibility on AI applications' exploiters. But even in that case, such exploiters may not be concerned about their responsibility if the costs for breaching it would not override the potential profits thereby gained. This could be the case given notably the potential large profits made from surveillance of citizens.[51]

The core regulatory question that then arises is how to ensure that individuals use technologies for the common good, including enhancing democracy? As will be shown, current proposals to tackle this question seem limited and therefore this topic warrants further research.

Some proposals made in the electoral context relate to placing responsibility on the campaigns themselves, which should thus monitor AI-enabled manipulation strategies. However, campaigns may for instance have the incentive to block information which would tarnish the image of the campaigns' candidates whereas such information would actually be accurate. Therefore, I believe that perhaps a more neutral actor could be involved in such monitoring, and further research could focus on this topic. Avoiding to unduly block information should also be a central concern for regulatory responses to AI-enabled manipulation in order to avoid censorship which could lead to and perpetuate authoritarian regimes.[52]

Further proposals relate to educating citizens about AI-enabled manipulation and alteration by teaching them how to distinguish real from fake news, real from bot accounts and how to identify manipulation.[53] Nevertheless, perhaps educating citizens will not suffice to overcome AI-enabled manipulation's threats

---

[51] Zuboff, 'We Make Them Dance …' (n 47) 15.

[52] Kamarck (n 38). See also discussions regarding the EU Directive on Copyright in the Digital Single Market (Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC [2019] OJ L130/92): Martin Senftleben and others, 'The Recommendation on Measures to Safeguard Fundamental Rights and the Open Internet in the Framework of the EU Copyright Reform' (2018) 40 European Intellectual Property Review 149; Sophie Stalla-Bourdillon and others, 'A Brief Exegesis of the Proposed Copyright Directive' (2017) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2875296> accessed 19 May 2020.

[53] Kamarck (n 38).

to democracy as such manipulation affects individuals' choice architecture. Accordingly, AI-enabled manipulation may affect decision making at such an unconscious level of human thinking that one may still be manipulated even if one is aware of the manipulation to which one is subject.

Conclusion

This paper sought to critically analyse my thesis according to which the reason why AI applications disrupt democracy lies in the new forms of manipulation enabled by AI applications.

I showed that AI-enabled manipulation is especially disruptive due to its ability to tailor manipulation to each individual, to adapt to each individual's change of conduct or thinking, its transparency (or improved hiddenness) and its potential wide reach.

I then analysed my thesis that AI-enabled manipulation challenges democracy because it disrupts three of its main principles or assumptions: citizens' autonomy, equal participation in democratic processes such as elections and that the public forum disseminates all the information required to cast enlightened votes. I conducted this analysis in the light of potential objections which allowed to refine my thesis.

I moreover made throughout my argumentation several proposals for further research, notably on how to ensure that individuals use technologies to enhance democracy.

However, I am aware of the possible limitations of my argumentation.

Firstly, my argumentation seems more theoretical than concrete. This is notably due to the nature of the topic, democracy, which is in itself an ideal type. This was a conscious choice that I made as it allowed my paper to explore why AI applications pose systemic challenges to democracy, i.e. why they challenge the very assumptions of democracy. Perhaps future research could thus investigate why the concrete electoral laws that protect citizens' autonomy, equality in participation and the public forum such as laws regulating campaign advocacy are currently unable to respond to the challenges posed by AI-enabled manipulation.

Secondly, my argumentation presumes that all democracies are similar, and hence that democratic principles and presumptions operate similarly everywhere. This allowed my argumentation to be relevant to democratic systems in general, and not to specific democratic systems. However, I am aware that democracy is always dependent on context and develops differently in dissimilar countries.[54] Therefore, I believe that regulatory responses to AI-enabled manipulation should be tailored to each democracy or at least leave a margin of appreciation to each democracy in case of international responses.

Thirdly, I am aware that AI-enabled manipulation is only one of the many ways allowing to influence citizens' decision making, and will continue to develop as AI develops. I also acknowledge that AI applications disrupt other foundations of democracy, such as fundamental rights, than that explored in this paper. My thesis does thus not provide a definitive or complete answer as to why AI disrupts democracy.

Lastly, as AI-enabled manipulation is posing unprecedented threats to the foundation of our legal systems, democracy, it is up to us to seize such threats as opportunities for democracy to develop and strengthen its values and principles.[55]

[54] Pratchett (n 19) 32.
[55] Pratchett (n 19) 33.