In Search of *uber*-Governance: Artificial Intelligence between Normative Challenges and Opportunities

Dr Aurora Voiculescu[1]

Abstract

In the past years, we have been confronted within the governance spheres with fast-paced developments aimed at addressing artificial intelligence (AI) technologies. Such movements have been prompted by the realisation of a great discrepancy between developments in AI and robotics technologies on the one hand, and the inherent, much slower pace of change in the normative spheres associated with these technologies, on the other hand. Spheres such as the ethical, the regulatory, the policy and the governance structures and processes have been revealed to be unprepared for handling the challenges brought by the advancements of the AI field.

The developments in AI and robotics raise a number of social issues that have acquired particular importance. The issue of what and whose ideas are reflected in the public space and are shaping it has long-stopped being a domestic only issue. Economic and communication processes, enhanced by modern technology, makes the way in which AI is conceptualised, designed and disseminated become of global importance. Appropriate governance mechanisms and processes are therefore needed for maximising AI-supported development opportunities while minimising the risks of harm associated with AI.

In this context, this paper is focusing on issues of inclusivity and multi-stakeholder participation as challenging, yet fundamental governance issues in shaping AI. Stemming from the AI's great potential to empower as well as from its just as great or even greater potential to disrupt and harm, the paper is looking into the capacity of existing structures to generate relevant and effective governance mechanisms and processes. These would be expected to address the challenges raised in the AI field according to more inclusive and deliberative processes.

In the past years, a number of AI governance initiatives have emerged, acknowledging at different levels that technology, and AI more than any other, stops being neutral as soon as it acquires a sense of social action, whether this is in the lab or in the real world. As numerous examples have shown, through the use of AI, biases can be amplified, disinformation spread, risks heightened yet left un-addressed. Given the fast pace of the AI technological developments and the need to foster such developments with the least possible regulatory impediments and red tape, the governance ethos seems to be in overdrive and, at the same time, at a loss. Overcoming a misguided infatuation with the exclusive focus on ethics and voluntary initiatives (a stage echoing largely the well-known motions of the debates present in the fields of business and human rights and in corporate social responsibility), a number of governance-focused initiatives feature markedly in the current debate. These initiatives stem largely from the usual governance structures, with the UN, OECD, EU as well as the private sector as key actors.

In this sense, the UN Roadmap for Digital Cooperation came to build on a previous initiative that looked at digital interdependence and highlighted the need for building an inclusive digital economy and society, focusing on developing human and institutional capacity; on protecting human rights and human agency; on promoting digital trust and stability; and on fostering global digital cooperation. The Roadmap emphasised global connectivity as key to meaningful participation and the need to define digital public goods fostered by an open-source ethos that encourages collaboration and experimentation. Key parameters such as race and gender are also put forward as important in promoting a digitally inclusive economy and society. The Roadmap also puts forth the need for building human and institutional capacity that is able to both take advantage and support the rise of the new technologies, through a digital capacity-building that is needs-based rather than supply-driven.

Putting forward governance structures apt to provide greater coherence and coordination is, therefore, essential for answering such expectations. The framework also acknowledges the fact that new governance structures and processes are required in order to address the threat to human rights. Digital technologies are seen as creating opportunities for new means to advocate, defend and exercise human rights, but they can also bring about new and heightened dangers that can lead to the suppression, limitation and even the violation of rights. This calls for specific attention being paid to issues such as data protection and privacy, digital identity, use of surveillance technologies, overall online interactions and content governance.

Other initiatives, such as the OECD Principles on Artificial Intelligence and the OECD.AI Policy Observatory, the European Commission's White Paper on Artificial Intelligence, the Global Partnership on Artificial Intelligence (GPAI), and the International Congress for the Governance of Artificial Intelligence (ICGAI) have identified similar issues, highlighting the fact that much more is needed in order to produce inclusive, responsive and effective global cooperation structures that can meet the challenges posed by the digital world and the AI technologies. Most voices in this debate, however, continue to suffer from inherent biases such as gender, race or geography. Therefore, bringing sociological insights to bear onto how discriminatory structures are reproduced becomes essential. Existing structures such as the UN, OECD, the EU can play here

[1] Law & Theory Lab, University of Westminster. Email: a.voiculescu@westminster.ac.uk

an important role in bringing governments, the private sector, the civil society, the academia and the technology community to the same table for working together. However, in the process, they must themselves demonstrate capacity for self-reflection.

In this context, a number of challenges remain to be taken forward and addressed, such as the enhancing of representation and inclusiveness in global discussions. Who the stakeholders around the table are and whose voices are being heard will come to constitute very important aspects of the new type of governance that is expected in the AI domain. Another type of challenge relates to the fragmentation and overall lack of coordination among the various AI normative initiatives, which impedes stakeholders' access to the existing governance groupings. A common, inclusive and democratic platform of debate and governance is yet to take shape. Last but not least, the development of global oversight and governance depend on key public sector stakeholders such as the governments and civil society sector, benefiting from additional capacity and expertise to engage on AI matters.

Each of the various initiatives so far has its own distinct role to play in the search for AI governance structures and processes that are robust and inclusive as well as agile. However, achieving this type of *uber*-Governance that displays flexibility and adaptability, while retaining a regard for inclusivity, due process and democratic deliberation, will clearly have to draw on the resources and expertise of a plurality of actors and on the different registers of engagement according to which these actors intervene in the AI governance processes.

Future robots as objects of governance: Analysing EU policy narratives

Jesse de Pagter[1*] and **Michael Filzmoser**[1]

Abstract. This extended abstract describes the analysis of European Union policy narratives on the future of robotics. It focuses on the way in which the imaginaries of the future concerning robots are rendering them into objects of technological governance. Narrative analysis, as a useful method for this endeavor, is presented and the implementation of narrative analysis with state-of-the-art annotation software is described.

1 INTRODUCTION

Visions of emerging technologies such as robotics is often promising a better future while those technologies are framed as the drivers behind new industrial revolutions. At the same time, however, those technologies come with many narratives about their far-reaching effects on our societies, democracies and economies. The promising expectations, as well as the fear for their disturbing societal effects, render robots (in combination with AI) into important drivers behind the imagination of the public, policymakers and corporations. Focusing on such imaginations, this project studies the governance of robotics in the context of European Union (EU) technology governance. Thereby it aims to develop an understanding of the EU's efforts regarding the development of policies that regulate and control the design, implementation and application of future robots [5].

2 CENTRAL CONCEPTS

Technological governance can be considered as one of the driving forces behind the development of new social, economic and political connections in the EU. Governance of (new) technology is as such a sociopolitical mechanism that gets deployed in order to further the socioeconomic integration and political unification of the EU [2]. For the analysis of the EU's technological governance (i) sociotechnical imaginaries and (ii) governance objects are central concepts.

2.1 Sociotechnical Imaginaries

First of all, the concept of sociotechnical imaginaries is used in order to analyse the EU's efforts to develop its technological governance of future robots. Sociotechnical imaginaries can be defined as *"collectively imagined forms of social life and social order reflected in the design and fulfillment of nation-specific scientific and/or technological projects"* [7, pp. 120]. Future-oriented abstractions that guide this governance process are an important object of study, as they have a generative character: those abstractions determine the activities, investments and interests of governments, rendering them into strong instruments of legitimation [8]. Within the context of policy-making, expectations regarding emerging technology can as such become enactments of a desired future [4]. The constitutive and performative interpretation of expectations themselves emphasizes the pressing conditions of new technologies. The EU's sociotechnical imaginaries, describing and prescribing the future of the EU and its robotic agenda, are therefore analyzed.

2.2 Governance Objects

A further important concept for developing an understanding of robots in EU policy documents is the notion of governance objects. This concept has gained recent interest because of its capabilities of tracing hybrid, co-produced entities that emerge from complex interactions of expert knowledge, political interventions and mundane practices [1]. The process of their emergence and stabilization is interesting in the sense that new networks of cooperation are developed between elements that were disconnected beforehand [3]. The use of this object-oriented perspective enables to understand technological networks in a way that allows for high levels of complexity and contingency. Furthermore, this approach can explain how and why a particular version of the object of governance is emerging. Such an analysis aims to provide new insights into the dynamic processes and (path-dependent) characteristics of technological governance. In that sense, the policy narratives on robots are understood as constitutive and "real" instances of agency in the constitution of an object of governance.

3 NARRATIVE ANALYSIS

Narrative analysis of EU policy documents is an adequate method to analyse the development of robotics as an object of governance through the study of its imaginaries of the future. The narrative focus of this analysis entails that policy documents from the EU are understood to be containing a collection of narratives.

Narrative analysis has for a long time been a method of research in the social sciences [6]. Recently, the narrative as the locus of analysis has gained attention in disciplines that have traditionally been more quantitative. Robert Shiller for instance, has argued for what he calls "narrative economics". He argues that human brains are strongly based on narratives and have the ability to produce social norms that *"govern our activities, including our economic actions"* [9, pp. 37]. With the aid of narrative analysis the formation and character of specific narratives can be studied. In this way it is specified how the expectation's material and rhetorical manifestations are facilitating the emergence of sociopolitical actions and agencies. An important aspect of the narrative approach is therefore that narratives

[1] TU Wien, Institute of Management Science, Austria
*corresponding author: jesse.de.pagter@tuwien.ac.at

are seen as a way to disseminate certain views: Narratives can serve as frames of reference and can as such be shared easily among different people or groups.

The tool that is used for the narrative analysis in this project is the web-based annotation tool Tagtog[2]. Tagtog provides annotation assistance on the basis of natural language processing. By annotating EU documents, the learning algorithm of the tool is in fact continuously improved which makes annotation of future policy documents easier.

4 EXPECTED RESULTS

Aim of this study is to describe how future imaginaries of robots can render them as objects of EU governance. The ongoing narrative analysis has already exposed how the many expectations regarding the different societal issues of future robots are in fact constituting a fertile ground for EU regulation. Furthermore the documents show how intentions to create robots that respect human values are extremely difficult to implement in actual regulation. Next to its aims of critical engagement this project tries to help in smoothing those processes of policy-making regarding technology that is not yet existent.

REFERENCES

[1] Bentley B. Allan, ‘From subjects to objects: Knowledge in International Relations theory’, *European Journal of International Relations*, **24**(4), 841–864, (December 2018).

[2] Andrew Barry, *Political Machines Governing a Technological Society*, Athlone, London, 2001.

[3] Daan Boezeman and Henk Jan Kooij, ‘Heated debates: The transformation of urban warming into an object of governance in the Netherlands’, in *Evolutionary Governance Theory*, 185–203, Springer, (2015).

[4] Mads Borup, Nik Brown, Kornelia Konrad, and Harro Van Lente, ‘The sociology of expectations in science and technology’, *Technology Analysis & Strategic Management*, **18**(3-4), 285–298, (July 2006).

[5] Dominik B. O. Bösl and Martina Bode, ‘Roboethics and Robotic Governance – A Literature Review and Research Agenda’, in *ROBOT 2017: Third Iberian Robotics Conference*, eds., Anibal Ollero, Alberto Sanfeliu, Luis Montano, Nuno Lau, and Carlos Cardeira, volume 693, 140–146, Springer International Publishing, Cham, (2018).

[6] Barbara Czarniawska, *Narratives in Social Science Research*, Introducing Qualitative Methods, Sage Publications, London ; Thousand Oaks, Calif, 2004.

[7] Sheila Jasanoff and Sang-Hyun Kim, ‘Containing the Atom: Sociotechnical Imaginaries and Nuclear Power in the United States and South Korea’, *Minerva*, **47**(2), 119–146, (June 2009).

[8] Sheila Jasanoff and Sang-Hyun Kim, *Dreamscapes of Modernity: Sociotechnical Imaginaries and the Fabrication of Power*, University of Chicago Press, 2015.

[9] Robert J. Shiller, *Narrative Economics: How Stories Go Viral and Drive Major Economic Events*, Princeton University Press, Princeton, 2019.

[2] https://www.tagtog.net/

Epistemic Exploitation and Contributory Injustice in Law and AI Research: Towards a Paradigm Change in Legal Research

Margarita Amaxopoulou

Abstract

Shaping the development of AI and its capacity to augment human efforts to achieve collective long-term goals (e.g. tackling climate change, sustainable development) is gradually becoming a focal interest of many legal researchers. Scholars have been exploring the potential of using AI to regulate social behaviour,[i] as well as the proliferation of AI in society and the corresponding challenges for the rule of law and fundamental legal principles.[ii] While there is a fair amount of variety of opinions within the legal research community as to *what* should be prioritised in the law and AI agenda, there is less attention being paid to *how* these priorities should be set, i.e. the epistemological and methodological prerequisites of conducting legal research into AI.[iii] In other words, we have not addressed the question: how do our perspectives, disciplinary training and worldviews as legal researchers influence, and potentially limit or bias, our capacity to produce research that will meaningfully impact the future of AI and, through that, the future of humanity?

This article seeks to flesh out the risks of epistemic bias in law and AI research, particularly looking at two prominent types of such bias: *epistemic exploitation*, i.e. requiring members of one group to educate the members of another group without proper consideration of this burden,[iv] and *contributory injustice,* i.e. using conceptual resources that hamper the ability of others to contribute to the knowledge community.[v] Findings pointing to both types of bias have come up in the analysis of originally collected qualitative data from semi-structured interviews with legal and computer science researchers working in the field of AI in the UK, US and EU, conducted within my doctorate research. This research has been exploring the role of computer science expertise as a mode of authority and governance that shapes the rise and fall of legal orders across borders and contributes to structural changes within the State.[vi]

The interaction between legal and computer science experts is crucial, considering that it is often the main communicative avenue through which not only AI development can conform to legal principles by-design,[vii] but also legal research can comprehend what questions should be prioritised *today* for meaningful AI governance *in the future*. Nonetheless, my findings suggest that this communication is hampered by the tendency of both sides to privilege their own conceptual resources and epistemological premises. Both sides often expect the other side to go at unrealistic lengths to educate them about common challenges in a language that they will be able to understand. The result is an impasse to produce meaningful exchanges and a minimal, tokenistic inclusion of one side's contributions to the other side's research work. This might pose insurmountable barriers to legal researchers' efforts to shape AI development in the long run.

In setting the priorities for longtermist legal research that can meaningfully influence the future development of AI, it is urgent that these epistemological challenges and limitations are tackled and a paradigm change for legal research is embraced. Within my summer research fellowship, I propose, first, to conduct a survey of attitudes of legal researchers to interrogate the forces that give rise to potential sources of epistemic bias, and, second, to make concrete recommendations for ameliorating communicative possibilities between legal and computer science AI researchers.

References

[i] Roger Brownsword, *Law, Technology and Society: Reimagining the Regulatory Environment Technology and Society* (Routledge 2019) 24-27; Karen Yeung and Martin Lodge, 'Algorithmic Regulation: An Introduction' in Karen Yeung and Martin Lodge (eds) *Algorithmic Regulation* (OUP 2019) 4-6.

[ii] Adrian Zuckerman, 'Artificial intelligence - implications for the legal profession, adversarial process and rule of law' (2020) 136 Law quarterly review 427, 453; Monica Zalnieriute, Lyria Bennett Moses and George Williams, 'The Rule of Law and Automation of Government Decision-Making' (2019) 82 Modern Law Review 425, 427-428; Mireille Hildebrandt, 'Algorithmic Regulation and the Rule of Law' (2018) 376 (2128) Philosophical Transactions of the Royal Society A: Mathematical, Physical & Engineering Sciences 1, 2-3.

[iii] Christoph Winter et al, 'Legal Priorities Research: A Research Agenda' Legal Priorities Project 5-7.

[iv] Giale Pohlhaus, ‘Varieties of Epistemic Injustice 1’ in Ian James Kidd, José Medina, Gaile Pohlhaus (eds) *The Routledge handbook of epistemic injustice* (Routledge 2017) 13.

[v] Kristie Dotson, ‘Cautionary tale: on limiting epistemic oppression’ (2012) 33 (1) Frontiers: A Journal of Women Studies 24, 31-32.

[vi] Gregory Shaffer, ‘The Dimensions and Determinants of State Change’ in Gregory Shaffer (ed) *Transnational Legal Ordering and State Change* (CUP 2013) 23-24.

[vii] Mireille Hildebrandt, ‘Saved by Design? The case of Legal Protection by Design’ (2017) 11(3) NanoEthics 307, 308-309.

Algorithmic Accountability on Social Media Platforms: Digital Identity Portability and Self-Regulation.

Anna Hovsepyan

Abstract. There is no "one-size-fits-all" solution to the problems that social media platforms bring. Nevertheless, algorithm-based systems' ubiquity has brought increasing attention to the lack of accountability and whether to regulate or not to regulate as those platforms allow greater levels of personalisation and precision by using the potential to synthesize large amounts of data. As a result, these great new powers might jeopardize stability and social justice.

For that reason, the paper argues that it has become necessary to change the legislation and to strengthen the existing enforcement mechanisms not to allow corporations to become too big to fail and, as a result, too big to jail as the moral hazard has already resulted in a high recidivism rate in financial crimes. But most importantly, regulation should trigger self-regulation first, which can be done by securing competition on the market and by greater responsibility on holding individuals accountable for their misconduct. Finally, the Right to Digital Identity Portability needs to be seen as one of the pillars that will give greater security to consumers and provide an impetus to the democratisation process on social media platforms.

INTRODUCTION

One of the biggest challenges within the tech industry has become the overwhelming power in the hands of a few giant corporations, often referred to as the Big Tech. This power has led to a decrease in competition on the market and the impossibility for small and medium businesses to properly function. But it also has put consumers in a very vulnerable position. What kind of vulnerability we will see further in the paper, but what is also important to understand while talking about possible solutions is that there is no one-size-fits-all solution, and despite those companies often being discussed under the same umbrella term, it is a big mistake that needs to be changed. The threats that Facebook presents and the threats that Amazon, Apple or Google possess are very much different. For that reason, the paper will discuss how social media providers (SMPs) can be regulated.

The discussion about regulating the Internet has been out there for decades. It can be summed up to "to regulate, or not to regulate",[1] but with the rise of new technologies and in particular machine learning and this new era of surveillance capitalism, micro-targeting and behavioural modification, the interest and the need to regulate it has become the biggest priority.[2] The idealistic view of the Internet as something that will self-regulate itself and will liberate and democratise societies around the world by providing a free public sphere with unlimited access to knowledge and information used to look like dreams come true,[3] but of course, with great new powers comes great new responsibilities. There is a legitimate concern about the possible harm, whether intentional or unintentional, caused by those systems, as the first public deployments of SMPs have seen ample evidence of the technologies being disruptive and destructive at the same time. In their works, the Net Delusion by Morozov and Surveillance Capitalism by Shoshana Zuboff, we can see how exactly SMPs can threaten democracies worldwide. Including the ones that have had a democratic heritage for centuries.

For that reason, the ever-increasing application of algorithms to decision making has prompted demands for algorithmic accountability, which has become an important but complex notion that has created legal challenges. The article begins by arguing that tech companies have become the new too big to fail, which automatically makes them 'too big to jail'[4]. It discusses the issues through the lenses of two historical cases, the Breakup of the Bell System as well as the failure of holding financial institutions accountable and what we can learn from there that resembles the same problems that SMPs present. Further, it introduces the idea of Digital Identity Portability and how it differs from the Right to Data Portability and what the history of Porting Authorisation Code can tell us.

Tech Companies are too Big to Fail.

To understand more about regulation, we need to embrace the Hegelian spirit of history, but in this case of legal history. To understand how to regulate the Big Tech, and in particular social media providers, we need to take a look at two cases (i) the breakup of the Bell System and (ii) the financial institutions that have become too big to fail and as a result too big to jail.

The list of mergers and acquisitions by Facebook goes on and on. The most notable ones have been the acquisition of Instagram in 2012 for $1 billion and WhatsApp for $19 billion in 2014. There is also a less famous acquisition made in 2019 where Facebook agreed to acquire brain-computing start-up

[1] Frederick Mostert, Free speech and internet regulation, *Journal of Intellectual Property Law & Practice*, Volume 14, Issue 8, August 2019, Pages 607–612, https://doi.org/10.1093/jiplp/jpz074

[2] Shoshana Zuboff. (2019) *The Age Of Surveillance Capitalism*. Profile Books.

[3] Evgeny Morozov. (2012). *The Net Delusion* (PublicAffairs).

[4] Garrett, B. (2016). *Too big to jail: How Prosecutors Compromise with Prosecutors*. Harvard University Press.

CTRL-labs that is “a start-up working on ways for people to control computers using their minds”.[5] Since all those M&A, there has been a legitimate concern over antitrust laws and whether the company has been trying to stifle competition. Now it is possible that as some former Facebook employees have put it that the purpose was not to “squash a would-be competitor”[6] but rather an attempt to survive competition from Twitter. Nevertheless, what we have now seen is Facebook with more than 2.7 billion monthly active users, Instagram with 1 billion active users, and WhatsApp with over 2 billion active users across 180 countries.

This seems to have way greater power than the well-known Bell System. The Facebook antitrust case resembles the concerns that were raised over the course of the 20th century over AT&T Corporation that gained so much power that the whole market was incapable of competing because of the monopoly. But things have changed after the U.S. Department of Justice (the Antitrust Division) brought a lawsuit against AT&T on the bases of violations of the Sherman Act which resulted in the breakup of the Bell System in 1982. This brought competition back to the market, and with greater competition comes a greater quality of products and greater security for consumers.

And this brings the attention to another story. Some companies have become too big to fail and as a result too big to jail. Even though the idea was coined by Garett in the context of financial crimes,[7] it seems that one of the issues with *big tech* is very similar to what is happening within financial institutions. In other words, when the global financial crisis of 2007-2008 shocked financial institutions, but as a result of those individuals responsible for it not usually being held liable, it increased a global debate on how to keep those in senior corporate positions accountable for their actions. For example, in the aftermath of the same financial crises, in Iceland, 26 financiers were held liable, while in the US and the UK no individual was.**[8]** Thus, the idea developed that the largest institutions are too big to fail. This then also raises questions if the whole institution of punishment, whether from the moralists’ perspective or from a utilitarian perspective plays any role at all. As from what we will see, neither individuals are being held accountable, nor a deterrence effect plays a role for companies that are too big to fail.

Case No 1: When the *LIBOR* scandal arose (stands for London Inter-bank Offered Rate), the *HSBC* bank already had a Deferred Prosecution Agreement (DPA) which was the outcome of a money laundering case that ended in 2012. *HSBC* was fined a record $1.92 billion by the US authorities and was also sued by the families of the US citizens that were murdered by drug gangs in Mexico claiming the bank let cartels launder billions of dollars to operate their businesses from jail.[9] One of the requirements of the deferred prosecution agreement (DPA) was for the bank to comply with its internal policy and regulations.[10] However, in a few years, *HSBC* was involved in a new scandal known as the *LIBOR* scandal. This time it paid $100 million to end a *LIBOR* rigging lawsuit in the U.S.[11] In aftermaths, the UK Serious Fraud Office closed the investigation, and only one person (Tom Hayes) was convicted for money laundering.[12]

Case No 2: Another example is one of the most important cases in this field, the legal action against British Petroleum (BP). The company paid the largest fine in the history of the U.S. It started in 2005 in Texas when an explosion killed fifteen workers, and hundreds more were injured. The prosecution agreed a $50 million fine in a settlement.[13] However, it did not prevent another oil platform explosion a few years later, which killed eleven workers and produced a massive oil spill in the Gulf of Mexico.[14] As a result, BP then paid $20 billion in compensation, yet no individual was held accountable.[15]

There is a vast amount of criticism against criminal justice agencies for not being able to impose fair, proportionate punishment, in many cases such as these, where large global companies and their directors have not been held liable to a high enough level to which many victims and the public regard as proportionate.[16] Big Tech has no less power than any of the financial institutions. Those companies know everything about us: “Facebook defines who we are, Amazon what we want, and Google defines what we think”,[17] and we know nothing about them. In addition, they are the owners and developers of the most advanced AI systems in the world. Some of them are as

[5] Salvador Rodriguez, 'As Calls Grow To Split Up Facebook, Employees Who Were There For The Instagram Acquisition Explain Why The Deal Happened' *CNBC*(2019) <https://www.cnbc.com/2019/09/24/facebook-bought-instagram-because-it-was-scared-of-twitter-and-google.html>.

[6] Salvador Rodriguez. (2019). 'As Calls Grow To Split Up Facebook, Employees Who Were There For The Instagram Acquisition Explain Why The Deal Happened' <https://www.cnbc.com/2019/09/24/facebook-bought-instagram-because-it-was-scared-of-twitter-and-google.html>

[7] Garrett, B. (2016).

[8] Pontell, Henry & Geis, Gilbert. (2007). Black Mist and White Collars: Economic Crime in the United States and Japan. Asian Journal of Criminology. 2. 111-126. 10.1007/s11417-007-9032-1.

[9] Pontell, Henry & Geis, Gilbert. (2007).

[10] Pontell, Henry & Geis, Gilbert. (2007).

[11] Justice.gov. (2018). *HSBC Holdings Plc Agrees to Pay More Than $100 Million to Resolve Fraud Charges*. Office of Public Affairs.[online] Available at: https://www.justice.gov/opa/pr/hsbc-holdings-plc-agrees-pay-more-100-million-resolve-fraud-charges

[12] The Serious Fraud Office, *'SFO Concludes Investigation Into LIBOR Manipulation'* (2019).

[13] The United State Department of Justice (2007). *British Petroleum to Pay More Than $370 Million in Environmental Crimes, Fraud Cases*. https://www.justice.gov/archive/opa/pr/2007/October/07_ag_850.html.

[14] Justice.gov. (2015). *U.S. and Five Gulf States Reach Historic Settlement with BP to Resolve Civil Lawsuit Over Deepwater Horizon Oil Spill*. [online] Available at: https://www.justice.gov/opa/pr/us-and-five-gulf-states-reach-historic-settlement-bp-resolve-civil-lawsuit-over-deepwater

[15] Macadams, R. (2014). How Should Prosecutors Punish Corporate Criminals? Review of Too Big To Jail: How Prosecutors Compromise with Corporations, by Brandon L. Garrett. *Harvard University Press*. [online] Available at: https://newramblerreview.com/book-reviews/law/how-should-prosecutors-punish-corporate-criminals

[16] Macadams, R. (2014).

[17] Dyson, G. (2013). *Turing's cathedral*. London: Penguin Books.

large as the population of more than 1/3 of our planet, which means that to some extent, they owe a duty to more than 2 billion users, such as in case of Facebook.[18]

But the *Cambridge Analytica* scandal which resulted in the largest corporate fine in US history ($5 billion US dollars), to be paid to the United States Treasury, fined by US Federal Trade Commission for massive and repetitive company's privacy violations has shown, that it will take less than a month for Facebook to recover this money through income.[19] The history of corporate wrongdoings and in addition, weak regulation indicates the reasons for recidivism being at a high level. Tech companies are the new 'too big to fail', which automatically makes them 'too big to jail',[20] and results in lack of accountability.

A possible solution would be to break up tech giants and draw limits on how far the power of corporations can go. In other words, the companies should be designed as "platform utilities"[21] where all users will have to be treated fairly and equally. In addition, there needs to be a limitation on the type of services the same platform can provide.[22] But most importantly there either should be more substantial fines on corporations or there should be greater responsibility for individuals within those corporations that are involved in the development and decision-making process that results in designing potentially harmful autonomous systems. As the history of corporate wrongdoings has illustrated how well protected big corporations are.

Digital Identity Portability and Self-Regulation

Nevertheless, by only 'breaking up' the problems will not be solved. As Morozov puts it in his work "there is much to admire about Google, Twitter and Facebook, but as they began to play an increasingly important role in mediating foreign policy, "admiration" is not a particularly helpful attitude for any policymaker"[23] including privacy concerns, micro-targeting by SMPs and a threat of a new wave of authoritarian governments that both *are* and *will* exploit those platforms for their interests. There is a greater need for self-regulation for three reasons: first, the pace of innovation is extremely high that no government regulation can ever keep up the same pace. Secondly, despite all the concerns over the Internet, it is, arguably, but still an independent platform that provides greater access to information and communication. It is important to be careful not to overregulate it. Finally, as Joanna J Bryson puts it, "absolutely every complex entity – cells, people, corporations, governments, religions, whatever - of course they all self-regulate, or they wouldn't persist!"[24] and indeed, "the good governments and good governance make that self-regulation easier".[25]

The idea of self-regulation on SMPs has failed, and that was a logical outcome. In other words, when you have a market where there is no competition, it is hard to expect that there will be any proper and adequate self-regulation. Self-regulation will result only when companies are interested in providing the best service to consumers. Still, it doesn't work in situations where consumers have nowhere to go and when consumer vulnerability is what can best describe the current status of business to consumer.

When in January 2021, WhatsApp introduced its new rather poorly explained privacy policy that left the users without an option to opt-out, it took less than a week to witness the most significant digital migration from WhatsApp to Telegram and Signal. According to the UK parliament's home affairs committee, over the first three weeks of January, Signal gained 7.5 million users globally, and Telegram gained 25 million users. As a result, within ten days, WhatsApp fell from the eights most downloaded app in the UK to the 23rd. The following outcome was logical and expected, the new privacy policies were postponed. But while WhatsApp was losing its users on the 28th of January 2021, Telegram came up with a great way of advertising itself: "Starting today, everyone can bring their chat history – including videos and documents – to Telegram from apps like WhatsApp, Line and KakaoTalk".[26]

Why was this such an important tactic to undertake?
Our digital identity, which includes our accounts on Facebook, Instagram, WhatsApp, Twitter and many other platforms, have become, to some extent, our extensions. Not in the way that Chalmers and Clark would have argued, which is not an objective of a legal paper. But there is greater attachment to our online accounts due to them being representations of our identity, personality, being for some an equivalent of a diary with memories of their loved once or friends. In some cases, users have been on those platforms for almost a decade.

Here is where the Right to Digital Identity Portability or, in other words, an updated version of the Right to Data Portability will become an important step towards breaking up the monopoly and bringing competition back to the market. The right to data portability was one of the most important novelties of the GDPR concerning data protection regulation, but at the same time, it ended up being one of the most underestimated ones. The GDPR does not give a broad definition to it as it might appear at first:[27]

[18] TechCrunch, 'Facebook Now Has 2 Billion Monthly Users… And Responsibility' (2017) <https://techcrunch.com/2017/06/27/facebook-2-billion-users/>
[19] Vaidhyanathan, S. (2019). Billion-dollar fines can't stop Google and Facebook. That's peanuts for them. *The Guardian*. [online] Available at: https://www.theguardian.com/commentisfree/2019/jul/26/google-facebook-regulation-ftc-settlement
[20] Garrett, B. (2016).
[21] Elizabeth Warren. (2019). 'It's Time To Break Up Amazon, Google And Facebook' *Medium* <https://medium.com/@teamwarren/heres-how-we-can-break-up-big-tech-9ad9e0da324c>.
[22] Elizabeth Warren. (2019).
[23] Evgeny Morozov. (2012).
[24] Joanna J. Bryson. (2021). 'Minimising Regulation'. Adventures in NI. <https://joanna-bryson.blogspot.com/2021/02/minimising-regulation.html>.
[25] Joanna J. Bryson. (2021).
[26] Telegram. (2021). 'Moving Chat History From Other Apps' <https://telegram.org/blog/move-history>.
[27] Rosemary Jay. (2017). *Guide To The General Data Protection Regulation* (1st edn, Sweet & Maxwell). p. 239

"the right to receive the personal data concerning him or her, which he or she has provided to controller"[28] is not a good enough definition. We know that personal social media accounts (SMAs) possess greater information and data sets than that. As Rosemary Jay states that "data which are created as a consequence of a relationship between a data subject and a controller are not in scope of this right as they are not provided by the data subject to the receiving controller".[29] The Right to Digital Identity Portability – and that is the right to take the full information (the account itself) and not only what was shared by the user and transfer it to another SMPs is what can change the competition and give consumers the flexibility and the right to choose.

To make it clear, another interesting development within Communication services was the introduction of the General Conditions of Entitlement on 25 July 2003 by Ofcom.[30] It stated that regulated providers must provide a Porting Authorisation Code (PAC) to their mobile switching customers on request. They are not allowed to refuse a custom to take the phone number they have been using and move to another Communications Provider. In addition, Ofcom has "launched a pre-enforcement programme to monitor compliance by Mobile Network Operators (MNOs) and Mobile Virtual Network Operators (MVNOs) with the requirements of General Condition 18.1 (GC18.1)."[31] One of the reasons very much resembles what we are having with SMAs. There was a great attachment by a customer to the phone number he or she were using, which was seen as something that was stifling the competition and forcing a customer to stay with the provider even when terms and conditions were unfair. This led to higher competition on the market, greater options for consumers, the ability to choose a fair contract with a mobile provider and self-regulation that would attract more customers.

CONCLUSION

There is no "one-size-fits-all" solution to the problems SMPs bring. However, time is of the essence, and it is necessary to find solutions to mitigate its risks. For that reason, rather than asking whether technology is for good or for bad, we should ask ourselves whether we are implementing it in a way that is inclusive, fair, transparent and if it allows everyone to be included in restructuring society to secure the rights and freedoms guaranteed under the Declaration of Human Rights.

For that reason, the paper argues that the regulation should be strengthened to exercise tighter control over SMPs. It has become necessary to change the legislation not to allow corporations to become too big to fail and, as a result, too big to jail as the moral hazard has already resulted in a high recidivism rate in financial crimes. But most importantly, regulation should trigger self-regulation first, which can be done by securing competition on the market and by greater responsibility on holding individuals accountable for their misconduct. Finally, the Right to Digital Identity Portability needs to be seen as one of the pillars that will give greater security to consumers and provide an impetus to the democratisation process on SMPs.

[28] General Data Protection Regulation. (2018) Art.20(1)
[29] Rosemary Jay. (2017). p.239
[30] 'General Conditions Of Entitlement' (*Ofcom*) <https://www.ofcom.org.uk/phones-telecoms-and-internet/information-for-industry/telecoms-competition-regulation/general-conditions-of-entitlement>.
[31] 'Own-Initiative Pre-Enforcement Programme Into Mobile Network Operators' And Mobile Virtual Network Operators' Compliance With General Condition 18.1 In Respect Of The Provision Of Porting Authorisation Codes' (*Ofcom*, 2009) <https://www.ofcom.org.uk/about-ofcom/latest/bulletins/competition-bulletins/all-closed-cases/cw_01018>.

Mitigating Discriminatory Impacts of Algorithmic Decision-making Systems in Hiring Processes: An Analysis Through the United Nations Guiding Principles on Business human Rights Framework

Ceyda Ilgen[1]

Abstract. Using algorithmic hiring decision-making systems by companies have great benefits on companies for managing a huge number of job applications, reducing cost, saving time and increasing efficiency. On the other hand, algorithmic decision-making systems can lead to discriminatory impacts bringing together deepening vulnerability, exclusion and marginalisation as well as infringing enjoyment of basic human rights and freedoms. Although companies play a central role in the development and use of algorithmic decision-making systems, so far only limited attention has been given to addressing these systems' discriminatory impacts on a corporate responsibility platform. Therefore, the intention of this paper is demonstrating how the United Nations Guiding Principles on Business and Human Rights (UNGPs), specifically emphasising the corporate responsibility to human rights, play a significant role in mitigating discriminatory impacts of algorithmic decision-making systems.

1 INTRODUCTION

In today's digitally driven world, AI applications are increasingly used by both public and private actors given their promises of savings in cost and time, minimizing risks and increasing efficiency [1][2]. Such applications use algorithms, unambiguous procedures relied on trained data to automatically solve complex problems or to make or support decisions about individuals that traditionally carried out by human intelligence [3][4]. Algorithmic decision-making systems, also called as automated decision-making systems, are analysing a vast amount of data to identify patterns that will be beneficial for reaching different decisions such as evaluating individuals' eligibility for benefits, credits or insurance, marketing and advertisements, predicting frauds, assisting in criminal sentencing or probation decisions [2]. Algorithmic decision-making systems are also becoming a significant part of hiring processes in companies to select employees which is not an easy process in itself [5]. Currently, many companies use algorithms to excel decision-making systems in hiring processes.

On the other hand, algorithms have the potentiality to learn and present existing social biases in the training data given by humans [5]. Thus algorithmic decision-making systems carry with the risks of perpetuating existing human biases and resulting in discriminative decisions [5][6]. If the data provided to train the algorithms are defective or are not represent a specific group of people, algorithmic decision-making systems can interference with the right to non-discrimination of individuals especially already marginalised or vulnerable groups, such as women or disabled people [7][6]. The right to non-discrimination means the protection of enjoyment rights and freedoms from unlawful interference on any ground such as gender, sex, race, religion, political opinion or disability. [8]. Interventions on the right to non-discrimination pave the way for the breaches on other human rights such as the right to privacy and freedom of expression and association [7].

Considering the importance of the right to non-discrimination in the exercising of basic rights and freedoms, it has become a significant need to analyse discriminatory impacts of using algorithmic decision-making systems by companies through a human rights-based perspective in order to be grounded on internationally accepted frameworks [7][9]. Under human rights lens, business and human rights (BHR) framework recognises the responsibility to respect human rights of companies from all sizes, sectors, locations, ownership and structure [10]. The United Nations Guiding Principles on Business and Human Rights (UNGPs), a milestone step in BHR, require companies to avoid violating human rights and to address their negative human rights impacts. To meet this responsibility, the UNGPs emphasise undertaking certain operational and ongoing processes, including human rights due diligence and impacts assessments that have the potentially to mitigate discriminatory impacts of algorithmic hiring systems.

This paper firstly reveals how algorithmic hiring decision-making systems are used by companies in the second section. Then, discriminatory impacts of algorithms on hiring decisions are analysed in the third section. This paper finally examines the role of the UNGPs framework in mitigating discriminatory impacts of algorithmic decision-making systems in the fourth section.

2 ALGORITHMIC DECISION-MAKING SYSTEMS IN HIRING PROCESSES

Hiring new employees is one of the most complex operations of companies and requires going through long processes including advertising job positions, creating a pool of candidates and determining who will be interviewed from among a huge number of applications

[1] Dept. of Law, Univ. of Westminster, W1W 7BY, UK. Email: {Ceyda, Ilgen}@my.westminster.ac.uk.

[11]. These stages are followed by the processes of conducting interviews and reaching the final decision about the most qualified candidate for the specific job position [12]. In these demanding processes, subjective views of recruiters could adversely impact on reaching unbiased decisions about selecting the right candidate [12].

AI-based applications, which have now become a significant part of almost all business industries, are transforming the traditional method of finding employees into automated systems driven by algorithms. There are several ways to automate hiring processes [2]. As the first starting phase of these processes, human resources departments of companies use job advertising platforms to post new job opportunities for finding an eligible candidate [5]. In accordance with the users' descriptions and previous preferences, algorithms already controls who could come across which job posting [13]. Social media platforms such as Linkedln, Instagram, CareerBuilder and Facebook are commonly used by companies to create algorithms for advertising job openings to suitable candidates [5][7][1]. Their algorithms analyse a huge amount of job applicants to detect the most qualified candidate by predicting candidates' potential success for the specific job position and then advise this candidate to the company [1][13].

To follow hiring process, applicant tracking systems are deploying by companies to gather information of candidates and separate them based on candidates' educational backgrounds, years of experience or other keywords to create a hiring database [5]. Pre-screening tools, such as intelligent CV screeners or virtual recruitment assistants, and pre-employment assessments, such as questionnaires and video interviews, are being used to collect information about candidates to identify most appropriate candidates based on their characteristics, skills and experiences [11]. Automated interview scheduling, candidate relationship management platforms, automated research on applicants, searching for red flags and name-matching technology are deployed to contribute the acceleration of hiring process in more accurate way. Such tools also help recruiters in identifying qualified candidates to be interviewed [5]. During interviews, candidates' language, tones of voice and emotions are evaluated through algorithms in facial recognition technologies and natural language processing in order to analyse their suitability for the job position [5].

Technology companies, including Google, Microsoft and IBM are providing algorithmic decision-making platforms to companies for being used in hiring processes [1]. Large companies, such as Unilever, Vodafone, Singapore Airlines and Intel are using time and cost-saving algorithmic hiring systems that scan applicants' word preferences, body languages and facial expressions during video job interviews [3]. Unilever uses the HireVue, a software from a US-based company, to automatically detect candidates' use of language, tone of speech and their smiles, head, eye or brows movements during video interviews, which is optional and based on the candidates consent [3]. HireVue software also evaluates candidates' skills in stress management and teamwork through natural language processing and facial expression processing [1].

Vodafone deploys Textio, a software to manage job advertisements, and Headstart, an algorithmic hiring system matching graduates to available positions based on analyses of the positions' needs regardless of which university the candidates graduated from. The latter system also enables recruiters to prioritise candidates from marginalised or vulnerable groups through certain optional features [4]. BDO, UK-based professional services company, started to conduct recorded video interviews with candidates to promote diversity in candidate pool and to include students who face challenges to attend a face-to-face interview [4].

Using algorithm-based hiring platforms accelerates the recruitment processes, increases productivity and thus enhancing the quality of hire [2]. Such platforms have important benefits on making or helping humans in reaching unbiased and fair hiring decisions by improving the accuracy of predictions about candidates' successes in the relevant job position. They can also contribute to the diversity and inclusion in the workplace through expanding the hiring pool of job applications.

On the other hand, algorithms could lead to discriminatory outcomes in employment decisions [1]. In the programming of automated decision-making systems, algorithms use input data to obtain information assumed beneficial to make decisions on the most qualified candidate [4][14]. If the data provided by designers to train algorithms are flawed, biased or unrepresentative, algorithmic systems could replicate biased patterns of humans especially for already marginalised or vulnerable groups of individuals [13][1]. The next section of this paper examines how algorithmic decision-making systems in hiring processes infringe the right to non-discrimination.

3 DISCRIMINATORY IMPACTS OF ALGORITHMIC DECISION-MAKING SYSTEMS

Any unlawful discrimination based on any ground such as gender, sex, race, religion, political opinion or disability is assessed as an interference with the principle of non-discrimination [8] Discriminatory impacts of algorithmic hiring systems can be resulted from different sources including biases emerging from training data or societal or individual biases of humans [4][15]. In more detail, algorithms in hiring systems are trained through the data obtained from previous candidates' applications or recorded interviews to detect which characteristics of candidates can be suitable for work competence [16]. The algorithm finds connections between characteristics and outcomes, and then uses those connections to predict new candidates' potentiality to be successful [16]. The main problem in this system is that when the majority of previous applications belong to certain groups, such as male candidates, algorithms learn masculinity characteristics and identify such characteristics as suitable for the job competence. Thus, there will be a discriminatory impact for female candidates.

Algorithmic hiring decision-making systems allow showing advertisements to candidates who are estimated to be most relevant to the position, but can also exclude other candidates who do not fall within those variables from seeing the advertisement [5]. Algorithmic systems can restrict online advertisements' visibility to selected age groups and prevent older employees from seeing those advertisements [5]. These systems can eventually lead to discrimination especially for certain vulnerable or marginalised groups including females, elderly and disabled people [5]. Job descriptions can be designed 'text analysis software to flag gendered language' that has the potential to discourage highly qualified women from applying for the jobs [7]. For instance, Vodafone is using a data-driven tool to post jobs, Textio that reflects some words such as competitive which have disincentive impacts on female applicants to apply for the roles [1]. Facial recognition technologies that are used to assess facial expressions potentially find a black person's face more aggressive than a white person due to lack of diverse data and thus failure in learning how to analyse dark-skinned faces [16]. These technologies have also adverse impacts on people with disabilities as their disability could affect the use of facial expressions or body languages [16].

Job postings through social media platforms could also magnify pre-existing biases. These platforms show job advertisements based on their predictions accordingly previous hiring decisions [1]. If companies have selected males for high job positions more than females, new advertisements are potentially seen more often by male candidates [1]. For instance, it has been found that taxi firms' job postings were shown basically to male job seekers. Likewise, job advertisements in certain sectors such as science, engineering, technology and math, were mainly seen by male candidates [1].

What makes the discriminatory impacts even more complicated, the interference with the right to non-discrimination is merely occurred after algorithms create a decision [1]. For instance, Amazon's algorithmic hiring system is accused of leading to discriminatory decisions against female applicants. It has been reported that Amazon's algorithmic system on CV screening was not rating the candidates gender-naturally as the machine learning models were trained on the previous male-dominant applications and therefore the company decided to cease the use of the system [1].

The need for a vast amount of personal data algorithmic decision-making systems lead to complex black-box algorithms characterised by the opaque and non-transparency processes of what data are gathered, how they are deployed, who benefits from their deployment and what algorithms are utilised to reach decisions [17].

Discriminatory and biased impacts of algorithmic decision-making systems can lead to violations on the enjoyment of other basic human rights and freedoms. Algorithmic decision-making systems pose a serious risk to the right to privacy and data protection [4]. Personal data of job seekers that are collected to train algorithms in the decision-making system can be misused by companies [4]. Algorithmic hiring systems also threaten the freedom of expression or association of applicants and employees by leading to chilling effect especially due to using facial recognition tools during interviews [5]. They can abstain from expressing themselves or from associating with other groups or unions due to the fear of the negative effect on their employability [3]. Denials of employment resulted from the mistakes or errors in algorithmic decision-making systems are directly related to interference with the right to work by restricting access to work [7].

4 MITIGATING DISCRIMINATORY IMPATCS THROUGH THE UNGPs

The right to non-discrimination is protected as one of the basic rights of individuals under major international and regional human rights instruments, such as Universal Declaration of Human Rights (UDHR), International Covenant on Civil and Political Rights (ICCPR), International Covenant on Economic, Social and Cultural Rights (ICESCR) and European Convention on Human Rights (ECHR) [18]. As the right to non-discrimination constitutes an essential principle under human rights law, discriminatory impacts of algorithmic decision-making systems should be analysed through a human rights-based approach that provides clarity and are relied on universally accepted frameworks [9].

Companies play a central role in the development and use of algorithmic decision-making systems, however, so far only limited attention has been given to addressing these systems' discriminatory impacts on a corporate responsibility platform [9]. Under international human rights law, adverse human rights impacts of companies have led to the emergence of business and human rights (BHR) field. In 2011, the United Nations (UN) Human Rights Council has taken the most important step in BHR field by adopting the United Nations Guiding Principles on Business and Human Rights (UNGPs) [5], the first framework that providing a global standard for preventing and addressing the risk of adverse business-related human rights impacts [12]. Hence, addressing discriminatory impacts of automated hiring systems through focusing on corporations' human rights responsibilities under the UNGPs has the advantage of grounding the conversation in globally recognised standards, widely recognised human rights, and management toolkits [9].

Furthermore, in recent years, algorithm-based adverse human rights impacts have led to 'algorithmic accountability' debate [2]. The BHR field is one of the most convenient disciplines in this debate as accountability plays a central role in the UNGPs framework [14]. The UNGPs emphasised where business enterprises identify that they have caused or contributed to adverse impacts, they should provide for or cooperate in their remediation through legitimate processes. Therefore, assessing this debate in the UNGPs framework will be useful to increase the accountability of companies when they conduct human rights violations by using data-driven technologies [2].

The UNGPs are grounded on a three-pillar framework: a duty of states to protect human rights, the responsibility of corporations to respect human rights and providing access to judicial or non-judicial remedies for those harmed by business activity, by both states and corporations [10]. According to first pillar, states have to protect individuals against business-related human rights violations by preventing, investigating, punishing and redressing these violations with effective regulations and policies. The second pillar specifies the requirements of companies' responsibility to respect human rights. This responsibility requires companies to avoid causing or contributing to human rights infringements in their operations, and to address adverse such infringements when they exist. It also expects from companies to prevent or mitigate negative human rights impacts that are directly connected to their activities. The responsibility to respect human rights is universal and applies to all business enterprises regardless of their size, sector, operational context, ownership and structure. It has been stated that the human rights responsibilities of businesses involve 'not just to the services they provide and the product they sell, but also to their internal operations' [3]. It is a clear fact that recruitment and hiring practices of companies are a part of the internal operations and companies have a responsibility to respect human rights when carrying out these practices.

The Guiding Principle 15 demonstrates companies how to fulfil their human rights responsibilities through some operational policies and processes, including adopting policy commitments and enabling access to remedy. This principle also expects companies to conduct human rights due diligence process which aims to identify, prevent, mitigate and account for how they address their adverse impacts on human rights. This operational process has to involve evaluating actual and potential human rights impacts, taking steps according to findings, tracking responses and discussing how adverse impacts are addressed [9]. The process has to be also continuing as the threats on human rights could change in time depending on evolvements on companies' operations [10].

Human rights due diligence rests upon application of human rights impact assessments. The Guiding Principle 19 emphasised that businesses should effectively integrate their findings from these assessments and take needed steps, in order to prevent, mitigate and address their potential and actual adverse human rights impacts. Both in human rights due diligence and impact assessments processes, actual and potential threats on human rights should be identified and assessed firstly to take appropriate steps [10].

In the context of algorithmic hiring systems, companies should be aware of the risks of biased algorithms in different hiring phases and not merely rely on the information produced by algorithms [1]. Biased input data should be avoided and algorithms should be carefully audited in order to make the process unbiased and fair [19]. Even algorithms offered by an external provider, companies should take into consideration the discriminatory outcomes of automated decision-making systems and evaluate them as a potential human rights risks, especially for already underrepresented groups such as women, elderly and disabled people. In accordance with the UNGPs, necessary policies should be taken to prevent and mitigate those risks. Companies also should provide transparency and prevent black-box algorithms in using automated decision-making systems. To do this, they should take steps to understand how the algorithm collects and analyses the provided data. As it could be difficult to understand the discriminatory impacts of algorithmic systems before they reach a final decision, there should be an ongoing human rights due diligence process with relevant actors from developers to users in order to provide transparency and fairness in making decisions [9].

5 CONCLUSIONS

Algorithmic decision-making systems are increasingly used in the hiring processes of companies. While these systems have significant benefits on accelerating long employment processes and providing efficiency, they can lead to biased and incorrect results about job seekers. As they generally reflect human biases and prejudices due to their designing and training processes, algorithmic decision-making systems have the capacity to generate discriminatory outcomes especially for certain vulnerable groups, such as women, elderly and disabled people. The risks on discrimination arising from using algorithms in hiring practices can also lead to infringements on the exercise of the right to privacy and to work, freedom of expression and association.

The business and human rights law framework recognises companies' responsibility to respect human rights. This responsibility includes businesses' internal activities, including algorithmic hiring operations. According to the UNGPs, companies should implement human rights due diligence processes to address algorithm-related adverse human rights impacts. They also should apply human rights impacts assessments to prevent and mitigate human rights risks of algorithmic technologies by taking steps to increase transparency in using algorithmic decision-making systems. What is more, to increase the algorithmic accountability, the UNGPs should be taken into consideration as these principles specifically emphasise the accountability of companies.

In conclusion, this paper evaluated the discriminatory risks of algorithmic decision-making systems used by companies. This paper also revealed how the UNGPs provide a convenient framework to mitigate such risks of algorithmic decision-making systems through requiring adopting operational and ongoing policies and practices by companies.

REFERENCES:

[1] A. Köchling and M. C. Wehner, "Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development," *Business Research*, vol. 13, no. 3. pp. 795–848, 2020, doi: 10.1007/s40685-020-00134-w.

[2] L. McGregor, D. Murray, and V. Ng, "International human rights law as a framework for algorithmic accountability," *Int. Comp. Law Q.*, vol. 68, no. 2, pp. 309–343, 2019, doi: 10.1017/S0020589319000046.

[3] I. Ebert, T. Busch, and F. Wettstein, "Business and Human Rights in the Data Economy A Mapping and Research Study."

[4] C. Castelluccia and D. Le Métayer, *Understanding algorithmic decision-making*, no. March. 2019.

[5] F. Raso, H. Hilligoss, V. Krishnamurthy, C. Bavitz, and L. Kim, "Artificial Intelligence & Human Rights: Opportunities & Risks," *Berkman Klein Cent. Internet Soc. Res. Publ. Ser. No. 2018-6*, vol. 7641, p. 63, 2018, [Online]. Available: https://cyber.harvard.edu/publication/2018/artificial-intelligence-human-rights.

[6] M. Latonero, "Intelligence : UPHOLDING HUMAN RIGHTS & DIGNITY Governing Artificial Intelligence : UPHOLDING & DIGNITY."

[7] L. McGregor, V. Ng, and A. Shaheed, "The Universal Declaration of Human Rights at 70 - Putting Human Rights at the Heart of the Design, Development and Deployment of Artificial Intelligence," 2018, [Online]. Available: https://48ba3m4eh2bf2sksp43rq8kk-wpengine.netdna-ssl.com/wp-content/uploads/2018/12/UDHR70_AI.pdf.

[8] M. Duwell and D. Mieth, "The Charter of Fundamental Rights of the European Union," *Biomed. Ethics*, vol. 5, no. 2, p. 51, 2000, doi: 10.5040/9781509917440.ch-005.

[9] I. Ebert, T. Busch, and F. Wettstein, "Business and Human Rights in the Data Economy A Mapping and Research Study."

[10] S. Bijlmakers and S. Bijlmakers, "The UN Guiding Principles on business and human rights," *Corp. Soc. Responsib. Hum. Rights, Law*, pp. 45–63, 2018, doi: 10.4324/9781351171922-3.

[11] M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy, "Mitigating Bias in Algorithmic Employment Screening: Evaluating Claims and Practices," *arXiv*. arXiv, Jun. 21, 2019, doi: 10.2139/ssrn.3408010.

[12] A. Taher and K. Arifen, "Recruitment and selection process in human resource management," vol. 10, no. 2, pp. 45–49, 2000, doi: 10.47760/ijcsmc.2021.v10i02.006.

[13] C. Schumann, J. S. Foster, N. Mattei, and J. P. Dickerson, "We need fairness and explainability in algorithmic hiring blue sky ideas track," *Proc. Int. Jt. Conf. Auton. Agents Multiagent Syst. AAMAS*, vol. 2020-May, no. Aamas, pp. 1716–1720, 2020.

[14] T. J. Freeman, A. Mckain, and S. Hall, "The Legal Implications of Algorithmic Decision-Making," *The Nebraska Lawyer*, no. May/June, pp. 25–29, 2020.

[15] B. Lepri, N. Oliver, E. Letouzé, A. Pentland, and P. Vinck, "Fair, Transparent, and Accountable Algorithmic Decision-making Processes: The Premise, the Proposed Solutions, and the Open Challenges," *Philos. Technol.*, vol. 31, no. 4, pp. 611–627, 2018, doi: 10.1007/s13347-017-0279-x.

[16] A. Kelly-lyth, "Challenging Biased Hiring Algorithms," vol. 00, no. 0, pp. 1–30, 2021, doi: 10.1093/ojls/gqab006.

[17] N. Verlagsgesellschaft *et al.*, "Nomos Verlagsgesellschaft mbH Chapter Title : Introduction Book Title : Building-Blocks of a Data Protection Revolution Book Subtitle : The Uneasy Case for Blockchain Technology to Secure Privacy and Identity Book Author (s): Shraddha Kulhari Published ," *Indep. Rev.*, vol. 20, no. 1, pp. 4–8, 2018.

[18] F. J. Zuiderveen Borgesius, "Strengthening legal protection against discrimination by algorithms and artificial intelligence," *Int. J. Hum. Rights*, vol. 24, no. 10, pp. 1572–1593, 2020, doi: 10.1080/13642987.2020.1743976.

[19] European Union Agency for Fundamental Rights (FRA), "BigData: Discrimination in data-supported decision making," *FRA Focus*, p. 14, 2018, [Online]. Available: https://fra.europa.eu/sites/default/files/fra_uploads/fra-2018-focus-big-data_en.pdf.

Identifying the Ethical and Legal Risks of Autonomous Robotic Systems: A pilot study

Daniel Trusilo and Thomas Burri

EXTENDED ABSTRACT

Autonomous robotic systems have enormous potential to do good. A system using Artificial Intelligence (AI) to autonomously operate in a damaged building after an earthquake could conceivably locate trapped victims faster than traditional human search and rescue teams or assist in a Fukushima type nuclear disaster[1] while keeping human disaster response teams out of harm's way. However, there is a dark side to the capabilities that are needed to operate in disaster environments – these capabilities can be weaponized to cause harm.

Through the application of an evaluation tool we previously developed, we are able to empirically assess autonomous disaster relief and weapon systems.[2] In interviews we conducted during a May 2020 pilot study of this tool, one common and dangerous misconception stood out among the engineers of four autonomous systems: When asked what would stop such systems from being used as weapons in war the engineers stated that they believed that the law would deem such activities illegal. However, under the current discussion of the use of autonomous systems in war, "there is no existing legal impediment to their development, deployment or use" [1].

The actual question of the legality of new means or methods of warfare is addressed by International Humanitarian Law (IHL) under Article 36 of Additional Protocol I to the 1949 Geneva Conventions [2], but there are challenges to applying an Article 36 review to an autonomous system 3. It has also been argued that autonomous systems may impact the normative moral theory that IHL is based on, redefining fundamental notions such as proportionality and necessity 4 or the standards related to the use of force in international relations 5. As a result, the systems that are in development today are being designed without fully accounting for problematic properties.

Therefore, there is a need to identify and discuss the ethical and normative challenges presented by autonomous robotic systems and the capabilities they present – to inform a debate on properties that may meet current legal standards but are in and of themselves ethically problematic. This need leads to a series of questions:

- How do existing ethical and legal boundaries apply to autonomous robotic systems?
- In the absence of legal clarity, how can problematic properties of autonomous robotic systems be practically identified in order to inform and contribute to the establishment of norms?

This paper explores the results of a pilot study of our efforts to develop a method to practically identify properties of autonomous robotic systems that are legally or ethically problematic. Our objective is to develop a practical method of informing the research and design, procurement, and operationalization of actual systems while contributing to the debate about what is lawful and ethical through the identification of trends. To achieve this objective, our research applies a series of assessments to determine: 1) if a system qualifies as a robotic system, 2) if a system possesses a degree of autonomy and, 3) if a system is designed to cause harm. Once these initial assessments are complete, and it has been determined that our evaluation tool is applicable, we can then evaluate an autonomous robotic system according to 37 specific aspects or properties. This detailed evaluation results in a composite risk level of each aspect leading to the identification of potentially problematic properties.

To present our research, we will first briefly discuss the way we define systems as well as the moral ethical and international humanitarian law foundations that our approach aims to supplement. We will then discuss the oft-cited historical example of a pre-emptive ban of new technology in warfare, namely the banning of blinding lasers. Next, we will discuss our applied ethics approach in greater detail through the real-world example of a quad-pedal autonomous robot, which was part of the pilot study of our evaluation tool. Lastly, we will discuss the future direction of our work and how our tool can support the crystallization of legal norms.

REFERENCES

[1] United Nations Institute for Disarmament Research, *The Weaponization of Increasingly Autonomous Technologies: Autonomous Weapon Systems and Cyber Operations*, www.unidir.org/publication/weaponization-increasingly-autonomous-technologies-autonomous-weapon-systems-and-cyber. (2017).

[2] ICRC (International Committe of the Red Cross). *A Guide to the Legal Review of New Weapons, Means and Mehtods of Warfare: Measures to Implement Article 36 of Additional Protocol 1 of 1977.* (2006).

[1] For a discussion of the use of robotic systems in hazardous environments see: "The Use of Robots to Respond to Nuclear Accidents: Applying the Lessons of the Past to the Fukushima Daiichi Nuclear Power Station" at: https://www.annualreviews.org/doi/pdf/10.1146/annurev-control-071420-100248

[2] The University of Zurich Digital Society Initiative White Paper titled "An Evaluation Schema for the Ethical Use of Autonomous Robotic Systems in Security Applications," by Markus Christen, Thomas Burri, Joseph Chapa, Raphael Salvi, Filippo Santoni de Sio, and John Sullins, in which the Schema was initially developed, can be found at: https://ssrn.com/abstract=3063617.

[3] Boulanin, V. and Verbruggen, M. *Article 36 Reviews: Dealing with the challenges posed by emerging technologies*. Report. Stockholm International Peace Research Institute, Stockholm, Sweeden, https://www.sipri.org/sites/default/files/2017-12/article_36_report_1712.pdf. (2017).

[4] Johansson, L. Ethical Aspects of Military Maritime and Aerial Autonomous Systems. *Journal of Military Ethics, 17*(2-3), 140-155. (2018).

[5] Huelss, H. Norms Are What Machines Make of Them: Autonomous Weapons Systems and the Normative Implications of Human-Machine Interactions. *International Political Sociology,* 14(2), 111-128. (2020).

Ethical Funding for Trustworthy AI: initial workshop outcomes

Gardner, A.[1], Leon Smith, A., Oldfield, M., A., Steventon, A., Coughlan, E.

1 INTRODUCTION

A number of ethical AI frameworks [1] have been published to guide developers in producing AI systems that help to mitigate the risks and harms that can occur. Systems that have been developed along ethical guidelines can be considered “Trustworthy AI”. However, despite the prevalence of these guidelines we continue to experience the deployment of AI systems that infringe on equality and human rights, demonstrating significant bias [2]. The Ethical Funding for Trustworthy AI (EFTAI) project was formed as a response to the increasing concerns regarding the development and deployment of Artificial Intelligence (AI) systems that result in bias, discrimination and infringements on human rights [3]. Specifically, we focus on how and why such AI systems have been funded and what controls are in place at this stage.

2 THE ROLE OF FUNDING AGENCIES

To date, the role of investors and funding bodies within the debate around AI ethics has been limited. How is it that systems that subsequently prove inadequate, discriminatory and harmful are being funded? What responsibilities do such bodies have in ensuring that “untrustworthy AI” is not funded or developed, and indeed can or should investors be held accountable? This is particularly relevant considering the seven principles of public life [4]. Addressing ethical issues in technology and innovation is not new, and funding agencies have significant experience in ensuring ethical compliance. For example, in developing policies and procedures related to diversity and inclusion [5]. Hence, it should be possible to implement similar safeguards for AI.

3 THE PROPOSALS

Following a review of a number of ethical AI frameworks and also of comparable ethical oversight strategies deployed by funding bodies a set of proposals was developed to enhance the application and governance of ethically aligned AI projects.

Proposal 1 outlines the suggestion that grant application forms contain a requirement for a Trustworthy AI Assessment. Here, applicants are required to outline the actions they plan to take to ensure they follow AI ethics guidelines. The proposal also offered an outline corresponding ‘guidance for applicants’.

Proposal 2 outlines the wider operational aspects funding agencies could employ to ensure the ethical oversight of funded projects and provide guidance to the funding agency itself. Central to Proposal 2 is the establishment of multi-disciplinary AI Ethics Boards, which may assist at different stages from grant screening, to advisory on policy.

4 STAKEHOLDER WORKSHOP

To assess the acceptability and feasibility of the proposals we conducted a workshop with key stakeholders. This included representatives from major funding organisations, AI Ethics experts and academics who apply for AI related projects. Prior to the workshop all participants were sent the full white paper for the EFTAI project, participant information sheet and consent form. The two-hour workshop was conducted online via MS Teams and commenced with position pieces by a funding organisation, an academic and an AI Ethicist to act as provocations for discussion. The position pieces do not constitute any part of the analysis of the workshop outcomes.

Participants were then divided by stakeholder group into 4 breakout groups of approximately 7 participants per group and one facilitator. This consisted of 2 funding groups and 2 academic/researcher groups. The breakout group had 45 minutes to discuss 3 questions:

Question 1 - What are your thoughts on the acceptability of these proposals as a solution to the identified problem?

Question 2 - Proposal One: Introduction of a Trustworthy AI Statement. What would effective implementation of this criteria look like to you?

Question 3 - Proposal Two: Management and Monitoring by Funding Bodies. What would effective governance of funded projects look like to you?

Feedback was via a Miro board [6] with participants anonymously adding their answers to each question following online discussions. Detailed Analysis of the feedback will be conducted by thematic analysis and sentiment analysis. An initial sentiment analysis, utilising an online sentiment analysis tool [7] was conducted to get an initial overview of reactions to the proposals.

5 INITIAL WORKSHOP FINDINGS

In terms of question 1 all groups indicated a positive sentiment that overall there is an identified problem and the proposals go in some way to address them. However, there was a significant split in opinion once the two proposals were considered separately. Question 2, regarding the inclusion of a trustworthy AI statement in grant applications had a strong positive outcome with the

[1] School of Computing and Maths, Keele University, Keele, Staffordshire, ST5 5BG. Emails: a.gardner@keele.ac.uk

funders (97.4% confidence) with a moderate positive outcome by researchers (at 64.9% and 73.3% confidence). With regards to question 3, looking at the governance mechanism such as AI ethics boards, funders were classified as neutral (56.7%) and researchers were strongly negative (89.9% and 94% respectively). One funder breakout group did not subdivide answers by question hence answers could not be included in the question specific analysis but combined analysis across all three questioned was considered negative (52.6% confidence).

5 CONCLUSIONS

The initial findings indicate that understanding of how to operationalise ethical AI guidelines is still in the early stages. In general, it appears that it is accepted there is a problem that funders have some role in addressing but there is little agreement on how that should be managed. Although there are difference in degree of approval it appears that both funders and researchers are aligned in their responses to the proposals. Further work using thematic analysis is due to be conducted to extract more detail from the workshop outputs and will be used to re-evaluate the proposals.

We acknowledge that the proposal for inclusion of a Trustworthy AI Statement on applications on its own is not enough to address the problems with untrustworthy AI. Solutions such as this exist within the context of a much wider set of actions, governance and regulations throughout the whole AI lifecycle. However, it is anticipated that a difference can be made through a small "nudge" in the application procedure, and by influencing the direction of flow of money. These proposals could have a wide and strong effect in terms of awareness, education of AI ethics and enabling the development of Trustworthy AI.

References

[1] HLEG,"Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment European Commissio," 2020. [Online]. Available: https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

[2] BBC, "Home Office drops 'racist' algorithm from visa decisions," 2020. [Online]. Available: https://www.bbc.co.uk/news/technology-53650758.

[3] Redden, J. *et al* "Datafied child welfare services: unpacking politics, economics and power.," *Policy Studies,* pp. 41:5, 507-526, 2020.

[4] House of Lords Liaison Committee, "AI in the UK: No Room for Complacency," 2020. [Online]. Available: https://publications.parliament.uk/pa/ld5801/ldselect/ldliaison/196/19602.htm.

[5] Directorate General for Research and Innovation, "Horizon 2020 Programme: How to complete your ethics self-assessment," 2019. [Online]. Available: https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/ethics/h2020_hi_ethics-self-assess_en.pdf.

[6] "Miro" [Online]. Available: https://miro.com/ .

[7] "Sentiment Analysis Online," [Online]. Available: .https://monkeylearn.com/sentiment-analysis-online/.