The Impact of Artificial Intelligence on Human Rights, Democracy and the Rule of Law

draft d.d. March 20, 2020

---

by Catelijne Muller, LLM.

ALLAI.

Table of Contents

I. Introductory remarks

AI, as a general purpose technology, has an impact on the entire fabric of society, In 2017, the European Economic and Social Committee, in what is widely considered the 'inception report' on the broader societal impact of AI, identified the most important societal impact domains including: safety; ethics; laws and regulation; democracy; transparency; privacy; work; education and (in)equality.[1] This means that AI has an impact on our human rights, democracy and the rule of law, the core elements upon which our European societies are built.

In 2019, the AI High Level Expert Group on AI presented Ethics Guidelines for Trustworthy AI.[2] These guidelines define trustworthy AI as being lawful, ethical and socio-technically robust. For the ethical element of trustworthy AI, the guidelines explicitly take fundamental rights as a basis for AI ethics.[3] While these guidelines do contain elements that are derived directly from existing (human) rights, they are not yet legally binding by themselves. Recently, the call for stronger (existing or new) legally binding instruments for AI has become louder. The European Commission announced potential elements of a legislative framework in its Whitepaper on AI[4] and stresses the importance of AI being in line with EU fundamental rights and the laws that ensure those rights.

This paper outlines the impact of AI on human rights, democracy and rule of law. It identifies those human rights, as set out by the European Convention on Human Rights ("ECHR"), its Protocols and the European Social Charter ("ESC"), that are currently most impacted or likely to be impacted by AI (Chapter II). In Chapters III and IV, it aims to provide a number of possible strategies that could be implemented simultaneously, if necessary. Chapter III looks at addressing the impact within the existing framework of human rights, democracy and the rule of law. Chapter IV looks at strategies, should the existing framework fail to adequately protect us. As technology and society are evolving quickly this paper cannot be exhaustive but prioritises the most relevant impacts to the extent that they can be identified today.

---

[1] EESC opinion on AI & Society (INT/806, 2017)

[2] Ethics Guidelines for Trustworthy AI, High Level Expert Group on AI to the European Commission, 2019

[3] Subsequently, the guidelines describe 7 requirements for trustworthy (i.e. lawful, ethical and robust) AI:

- Human agency and oversight
- Technical robustness and safety
- Privacy and data governance
- Transparency
- Diversity, non-discrimination and fairness
- Social and environmental well-being
- Accountability

[4] Whitepaper on artificial intelligence, COM(2020) 65 final.

II. Defining AI

AI has a myriad of applications that have already been introduced into society: biometric (including facial) recognition, object recognition, risk and success prediction, algorithmic decision making or support, automatic translation, recommender systems, and so on. These applications have found their way into sectors such as law enforcement, justice, human resource management, financial services, transport, healthcare, public services, etc.

AI remains an essentially contested concept, as there is no universally accepted definition. Nevertheless, definitions can broadly be clustered in two camps: rationalist and human-centric definitions. The most prominent rationalist definition, defines AI as "an agent created by humans that decides and performs actions based on its perception".[5] The best-known example of a human-centric definition is the Turing test, which is passed by a computer, as soon as it performs a task that would otherwise require human (conversational) intelligence. The High Level Expert Group on AI has provided a definition of AI in 2019[6]:

> *Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions.*

Often, AI is described as a collection of technologies that combine data, algorithms and computing power. While this is correct for the most widely used AI-systems at present, this is only a very limited description of what AI is. AI is a container term for many computer applications, some of which combine data and algorithms, but other, non-data-driven AI approaches, also exist, e.g. expert systems, knowledge reasoning and representation, reactive planning, argumentation and others.

Most AI systems that have been penetrating our societies lately, are indeed examples of data-driven AI, with particular impact on human rights, democracy and rule of law.

[5] Russell, S. J., Norvig, P., & Davis, E. (2010). Artificial intelligence: A modern approach (3rd ed). Prentice Hall.
[6] EU High Level Expert Group on AI. A definition of AI, main capabilities and scientific disciplines, 2019.

In the following the most relevant terms are defined:

- Narrow AI: AI systems that can perform only very specific 'narrow' tasks;
- Big (historical) Data: AI systems that need a lot of (historical) data to perform well. The quality, volume and content of the data influence the operation of these AI system, and often lead to replication and amplification of errors, gaps and biases in the data;
- Correlation: many AI systems only look for relations in data, which do not establish (or "see") a causal relationship between a 'case' and a decision, but merely makes a prediction based on shared characteristics with other 'cases';
- Black boxes: many AI systems often are so-called black boxes, within which (decision) processes take place that cannot be fully explained in human terms;
- Common sense: AI systems do not have common sense, meaning that while a system might be able to recognize a cat or a cancer cell, it has no conception of the idea of what a cat or a cancer cell is. It merely provides a label to a specific pattern. It also cannot use the information about a cat or a cancer cell to identify a dog or a headache.

All these characteristics can make current AI brittle, unstable and unpredictable, but also popular and widely applied.

Most importantly, AI systems are more than just the sum of their software components. AI systems also comprise the socio-technical system around it. When considering governance, the focus should not just be on the technology, but also on the social structures around it: the organisations, people and institutions that create, develop, deploy, use, and control it, and the people that are affected by it, such as citizens in their relation to governments, consumers, workers or even entire society.

Defining AI for regulatory purposes

A complicating factor is that legal definitions differ from pure scientific definitions whereas they should meet a number of requirements[7] (such as inclusiveness, preciseness, comprehensiveness, practicability, permanence), some of which are legally binding and some are considered good regulatory practice[8].

[7] A Legal Definition of AI Jonas Schuett Goethe University Frankfurt September 4, 2019 *(Legal definitions must be: (i) inclusive: the goals of regulation must not over- or under-include. (Julia Black. Rules and Regulators. Oxford University Press, 1997. [32] Robert Baldwin, Martin Cave, and Martin Lodge. Understanding Regulation: Theory, Strategy, and Practice. Oxford University Press, 2nd edition, 2012.); (ii) Precise: it should be clear which case falls under the definition and which does not; (iii) Comprehensive: the definition should be understandable by those who are regulated; (iv) Practicable: legal professionals should be able to easily determine whether a case falls under the definition; (v) Permanent: the need for continued legal updating should be avoided.*

[8] Inclusiveness can be derived from the principle of proportionality in EU law (art. 5(4) of the Treaty on European Union. The criteria precision and comprehensiveness are based on the principle of legal certainty in EU law. The criteria practicability and permanent are considered good legislative practice.

III. Impact of AI on Human Rights, Democracy and the Rule of Law

Taking an 'AI lifecycle approach' is important, in order to consider not only the development stage of AI, but also the deployment and use stages. Another element to keep in mind is that most AI-applications currently being used could enshrine, exacerbate and amplify the impact on human rights, democracy and the rule of law at scale, affecting larger parts of society and more people at the same time.

Four "Families of Human Rights" under the ECHR, its Protocols ESC are impacted by AI:

1. Respect for Human Value
2. Freedom of the Individual
3. Equality, Non-Discrimination and Solidarity
4. Social and Economic Rights

Moreover, AI has ample impact on:

5. Democracy
6. The Rule of Law

It is important to note that many AI-systems or uses can impact various human rights, democracy and the rule of law at the same time, or adversely affect one person's human rights while positively affecting another's.

1. AI & Respect for Human Value

Respect for human value is reflected by the ECHR in various rights, such as the right to liberty and security (art. 5), the right to a fair trial (art. 6), the right to no punishment without law (art. 7) and the right to a private life and physical and mental integrity (art. 8). AI can impact these rights in the following ways.

<u>Liberty and Security, Fair Trial, No Punishment without Law (art. 5, 6, 7 ECHR)</u>

The fact that AI can perpetuate or amplify existing biases, is particularly pertinent when used in law enforcement and the judiciary. In situations where physical freedom or personal security is at stake, such as with predictive policing, recidivism risk determination and sentencing, **the right to liberty and security combined with the right to a fair trial are vulnerable.** When an AI-system is used for recidivism prediction or sentencing it can have biased outcomes. When it is a black box, it becomes impossible for legal professionals, such as judges, lawyers and district attorneys to understand the reasoning behind the outcomes of the system and thus complicate the motivation and appeal of the judgement.

Less obvious is the impact of AI on **the right to reasonable suspicion and prohibition of arbitrary arrest**. AI-applications used for predictive policing merely seek correlations based on shared characteristics with other 'cases'. Suspicion in these instances is not based on actual suspicion of a crime or misdemeanor by the particular suspect, but merely on shared characteristics of the suspect with others (such as address, income, nationality, debts, employment, behaviour, behaviour of friends and family members and so on). Moreover, the actual characteristics used in the AI-system and the 'weights' given to those characteristics are often unknown.

If applied responsibly however, certain types or uses of AI can however also improve security, for example AI applications that can 'age' missing people in pictures to improve chances of finding them or AI-driven object recognition that can scan luggage at an airport for suspected contents.

Private and Family Life, Physical, Psychological and Moral Integrity (art. 8 ECHR)

Many AI-systems and uses have a broad and deep impact on the right to privacy. Privacy discussions around AI currently tend to focus primarily on data privacy and the indiscriminate processing of personal (an non-personal) data. It should however be noted that, **while data privacy is indeed an important element, the impact of AI on our privacy goes well beyond our data**. Art. 8 of the ECHR encompasses the protection of a wide range of elements of our private lives, that can be grouped into three broad categories namely: (i) a person's (general) privacy, (ii) a person's physical, psychological or moral integrity and (iii) a person's identity and autonomy.[9] Different applications and uses of AI can have an impact on these categories, and have received little attention to date.

AI-driven (mass) surveillance, for example with **facial recognition**, involves the capture, storage and processing processing of personal (biometric) data (our faces)[10], but it **also affects our 'general' privacy, identity and autonomy** in such a way that it creates a situation where we are (constantly) being watched, followed and identified. As a **psychological 'chilling' effect**, people might feel inclined to adapt their behaviour to a certain norm, which shifts the balance of power between the state or private organisation using facial recognition and the individual.[11] In legal doctrine and precedent the chilling effect of surveillance can constitute a violation of

[9] Guidance to art. 8 ECHR, Council of Europe.
[10] The jurisprudence of the European Court of Human Rights (ECtHR) makes clear that the capture, storage and processing of such information, even only briefly, impacts art. 8 ECHR.
[11] Examined Lives: Informational Privacy and the Subject as Object, Julie E. Cohen, 2000.

the private space, which is necessary for personal development and democratic deliberation.[12] Even if our faces are immediately deleted after capturing, the technology still intrudes our psychological integrity.

And while for facial recognition the impact on our 'general' right to privacy and our psychological integrity might be more obvious, one could argue that the indiscriminate **on- and offline tracking of all aspects of our lives** (through our online behaviour, our location data, our IoT data from smart watches, health trackers, smart speakers, thermostats, cars, etc.), could have the same impact on our right to privacy, including our psychological integrity.

Other forms of AI-driven biometric recognition have an even greater impact on our psychological integrity. **Recognition of micro-expressions, gate, (tone of) voice, heart rate, temperature, etc.** are currently being used to assess or even predict our behaviour, mental state and emotions.

It should be noted upfront that **no sound scientific evidence exists** corroborating that a person's inner emotions or mental state can be accurately 'read' from a person's face, gate, heart rate, tone of voice or temperature, let alone that future behaviour could be predicted by it. In a recent meta-study a group of scientists[13] concluded that AI-driven emotion recognition could, at the most, recognize how a person subjectively *interprets* a certain biometric feature of another person. An interpretation does not align with how that person actually feels, and AI is just labeling that interpretation which is highly dependent on context and culture. Far-fetched statements, that AI could for example determine whether someone will be successful in a job based on micro-expressions or tone of voice, are simply without scientific basis.

More importantly, the widespread use of these kinds of AI techniques, for example in recruitment, law enforcement, schools, impacts a person's physical, psychological or moral integrity and thus elements of that person's private life.

It should be noted that GDPR restricts the processing of biometric data only to some extent. Biometric data according to the GDPR is "personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of a natural person, **which allow or confirm the unique identification of that natural person**.

[12] The chilling effect describes the inhibition or discouragement of the legitimate exercise of a right. It has been shown that once people know that they are being surveilled they start to behave and develop differently.
Staben, J. (2016). Der Abschreckungseffekt auf die Grundrechtsausübung: Strukturen eines verfassungsrechtlichen Arguments. Mohr Siebeck.

[13] Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. Psychological Science in the Public Interest, 20(1), 1–68.

The last part of the sentence is crucial, because if biometric recognition is not aimed at identification (but for example at categorization, profiling or affect recognition), it might not fall under the GDPR-definition. In fact, recital 51 of the GDPR says that 'the processing of photographs [is considered] biometric data only when processed through a specific technical means allowing the unique identification or authentication of a natural person.'

Many biometric recognition technologies are not aimed at processing biometric data to uniquely identify a person, but merely to assess a person's behaviour (for example in class) or to categorize individuals (for example for the purpose of determining their insurance premium based on their statistical prevalence to health problems). These uses might not fall under the definition of biometric data (processing) under the GDPR.

Going back to data privacy, personal and non-personal data is not only being used to train AI systems, but also to profile and score people for various purposes such as predictive policing, insurance acceptance, social benefits allowance, performance prediction in hiring and firing processes. Moreover, massive amounts of 'data points' on how we go about our daily lives are used not only to send us targeted advertising, but also to push/influence/induce/nudge us towards certain information and thus influence our options, affecting our moral integrity.

2. AI & Freedom of the Individual

Freedom of the individual is reflected by the ECHR in various rights, such as freedom of expression (art. 10) and freedom of assembly and association (art. 11). AI can have a 'chilling' effect on these freedoms as well.

Freedom of Expression (art. 10 ECHR)

Art. 10 of the ECHR provides the right to freedom of expression and information, including the freedom to hold opinions, and to receive information and ideas. AI being used to profile, surveil, track and identify people and screen, define, sort and influence or nudge behaviour not only has a potential impact on the right to moral integrity as described above, it can also have a chilling effect on these particular freedoms.

Using facial recognition in public areas may interfere with a person's freedom of opinion and expression, simply because of the fact **that the protection of 'group anonymity' no longer exists, if everyone in the group could potentially be recognized**.

This could lead to those individuals changing their behaviour for example by no longer partaking in peaceful demonstrations.[14]

The same goes for the situation where all our data is used for AI-enabled scoring, assessment and performance (e.g. to receive credit, a mortgage, a loan, a job, a promotion, etc.). People might become **more hesitant to openly express a certain opinion**, read certain books or newspapers online or watch certain online media.

With regards to the right to receive and impart information and ideas, AI used in media and news curation, bringing ever more 'personalized' online content and news to individuals, raises concerns. Search engines, video recommendation systems and news aggregators often are opaque, both where it comes to the data they use to select or prioritize the content, but also where it comes to the purpose of the specific selection or prioritization.[15] Many business models are based on online advertising revenue. In order to have people spend as much time on a platform or website as possible, they might be selecting and prioritizing content that will do only that: keep people on their platform, irrespective of whether this content is objective, factually true, diverse or even relevant.

Beyond commercial motives, political or other motives might lead to AI-systems being optimized to select or prioritize particular content in an effort to coerce and influence individuals towards certain points of view, for example during election processes.[16]

Moreover, AI is becoming more capable of producing media footage (video, audio, images) **resembling real people's appearance and/or voice** (also known as 'deep fakes'), enabling the deceptive practices for various purposes.

All this can give rise to filter bubbles and proliferation of fake news, disinformation and propaganda, and affects the capacity of individuals to form and develop opinions, receive and impart information and ideas and thus impact our freedom of expression.[17]

[14] Privacy Impact Assessment Report for the Utilization of Facial Recognition Technologies to Identify Subjects in the Field, 30 June 2011, p. 18.
[15] Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data & Society, 3(1), 2053951715622512.
[16] Cambridge Analytica, Netflix Documentary: The Great Hack
[17] UN Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, A/73/348

Freedom of Assembly and Association (art. 11 ECHR)

The internet and social media have shown to be helpful tools for people to exercise their right to peaceful assembly and association. At the same time however, the use of AI could also jeopardize these rights, when people or groups of people are automatically tracked and identified and perhaps even 'excluded' from demonstrations or protests.[18]

As already mentioned, **the use of facial recognition in public areas in particular might discourage people to attend demonstrations and join in peaceful assembly**, which is one of the most important elements of a democratic society. Examples of this were already seen in Hong Kong when protesters started wearing masks and using lasers to avoid being 'caught' by facial recognition cameras.

3. AI & Equality, Non-discrimination and Solidarity

Prohibition of Discrimination (art. 14 ECHR, Protocol 12)

One of the most reported impacts of AI on human rights is the impact on the prohibition of discrimination and the right to equal treatment. As noted earlier, in many cases, **AI has shown to perpetuate and amplify and possibly enshrine discriminatory or otherwise unacceptable biases**. Moreover, these data-driven systems obscure the existence of biases, marginalising the social control mechanisms that govern human behaviour.

As an example, Amazon's recruitment AI favoured men over women, because it was trained on profiles of successful Amazon employees, which happened to be men. The AI-system did not simply filter out women, it looked at characteristics of successful employees such as typical wordings and phrasing and filtered out CV's that did not show these characteristics.[19]

Going back to the workings of present day AI, where the systems merely look for correlations based on shared characteristics with other 'cases', all kinds of unacceptable biases can easily surface. The problem with these systems is that, even if they would excel at identifying patterns, e.g. typical phrases used by successful employees, the system has no understanding of the meaning of the phrases, let alone that it will be able to understand the meaning of success, or even grasp what an employee is. It will only be able to provide a label to a specific pattern.

[18] Algorithms and Human Rights, Study on the human rights dimensions of automated data processing techniques and possible regulatory implications, Council of Europe, 2018

[19] Dastin, J. (2018, October 10). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters.

Contrary to popular belief, not all biases are the result of low-quality data. **The design of any artifact is in itself an accumulation of biased choices**, ranging from the inputs considered to the goals set to optimize for; does the system optimize for pure efficiency, or does it take the effect on workers and the environment into account? Is the goal of the system to find as many potential fraudsters as possible, or does it avoid flagging innocent people? All these choices are in one way or another driven by the inherent biases of the person(s) making them. In short, suggesting that we can remove all biases in (or even with) AI is wishful thinking.[20]

4. AI & Social and Economic Rights

AI in and around the Workplace

AI can have major benefits when used for hazardous, heavy, exhausting, dirty, unpleasant, repetitive or boring work. AI systems are however also increasingly being used **to monitor and track workers, distribute work without human intervention and assess and predict worker potential and performance in hiring and firing situations**. These applications of AI could jeopardize the right to just conditions of work, safe and healthy working conditions, dignity at work as well as the right to organize (art. 2 and 3 ESC). If workers are constantly monitored by their employers, they might become more hesitant to organize (art. 5). AI-systems that assess and predict performance of workers could jeopardize the right to equal opportunities and equal treatment in matters of employment and occupation without discrimination on the grounds of sex (art. 20 ESC), especially when these systems enshrine biases within the data or of their creator.

There is a risk of loss of necessary skills when more and more work and decisions that were previously performed or taken by humans are taken over by AI-systems. This could not only lead to a less skilled workforce, it also raises the risk of systemic failure, where only a few humans are capable of working with AI-systems and reacting to events where these systems fail.

While it is unknown if, and if so how many jobs will be lost or gained as a result of AI, in the disruptive transformation period**,** a mismatch between vulnerable labour forces and required skills could lead to technological unemployment.[21]

[20] First Analysis of the EU Whitepaper on AI, Virginia Dignum, Catelijne Muller, Andreas Theodorou, 2020.
[21] Brynjolfsson, E., & McAfee, A. (2014). The second machine age: Work, progress, and prosperity in a time of brilliant technologies. W. W. Norton & Company.

5. AI & Democracy

AI can have (and likely already has) an adverse impact on democracy, in particular where it comes to: (i) social and political discourse, access to information and voter influence, (ii) inequality and segregation and (iii) systemic failure or disruption.

Social and political and social discourse, access to information and voter influence

Well-functioning democracies require a well-informed citizenry, an open social and political discourse and absence of opaque voter influence.

This requires a **well-informed citizenry**. In information societies citizens can only select to consume a small amount of all the available information. Search engines, social media feeds, recommender systems and many news sites employ AI to determine which content is created and shown to users (information personalization). If done well, this could help citizens to better navigate the flood of available information and improve their democratic competences, for instance by allowing them to access resources in other languages through translation tools.[22] However, if AI determines which information is shown and consumed, what issues are suppressed in the flood of online information and which are virally amplified, this also brings risks of bias and unequal representation of opinions and voices.

AI-driven information personalisation is enabled by the constant monitoring and profiling of every individual. Driven by commercial or political motives this technologically-enabled informational infrastructure of our societies could amplify hyper-partisan content one is likely to agree with and provide an unprecedented powerful tool for individualised influence.[23] As a consequence it may undermine **the shared understanding, mutual respect and social cohesion required for democracy to thrive**. If personal AI predictions become very powerful and effective, they may even threaten to undermine the human agency and autonomy required for meaningful decisions by voters.[24]

Thirdly, AI can undermine a **fair electoral process**. Political campaigns or foreign actors can use (and have been using) personalised advertisements to send different messages to distinct voter groups without public accountability in the agora.[25] However, it should be noted that it

[22] Schroeder, R. (2018). Social Theory after the Internet. UCL Press.
[23] Wu, T. (2016). The attention merchants: The epic scramble to get inside our heads (First edition); Zuboff, S. (2019). The age of surveillance capitalism: The fight for the future at the new frontier of power.
[24] Taddeo, M., & Floridi, L. (2018b). How AI can be a force for good. Science, 361(6404), 751–752.
[25] Bradshaw, S., & Howard, P. (2019). Social Media and Democracy in Crisis. Oxford University Press.

remains uncertain exactly how influential micro-targeted advertisement is.[26] AI can also be used to create and spread misinformation and deep fakes, in the form of text, pictures, audio or video. Since these are hard to identify by citizens, journalists or public institutions, misleading and manipulating the public becomes easier and the level of truthfulness and credibility of media and democratic discourse may deteriorate.

Inequality and segregation

AI is widely expected to improve the **productivity** of economies. However, these productivity gains are expected to be distributed unequally with most benefits accruing to the well-off. Similarly, data and design choices, combined with a lack of transparency of black box algorithms have shown to lead to a perpetuating unjust bias against already disadvantaged groups in society, such as women and ethnic minorities.[27] AI could lead to inequality and segregation and thus threaten the necessary level **of economic and social equality** required for a thriving democracy.

Systemic risks

AI decisions that previously only humans were able to make, create new challenges for the **security and resilience** of societal systems. In particular, if decisions that previously were made by many decentralised actors are replaced by few centralised AI-driven systems, the systemic risks increase, where only a failure of few centralised systems is enough to potentially create catastrophic results.

Financial markets illustrate how new systemic risks emerge when different AI agents interact at superhuman speeds, as the rise of financial flash crashes have demonstrated.[28] When critical energy infrastructures, transport systems and hospitals increasingly depend on automated decisions of artificial agents this introduces new vulnerabilities in the form of a single point of failure with widespread effects. Once efficient systems of critical infrastructure are introduced they are harder to replace or backed-up if they break down.[29]

[26] Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. Nature, 489(7415), 295–298.
[27] Eubanks, V. (2017). Automating inequality: How high-tech tools profile, police, and punish the poor (First Edition). St. Martin's Press; O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy (First edition).
[28] Wellman, M. P., & Rajan, U. (2017). Ethical Issues for Autonomous Trading Agents. Minds and Machines, 27(4), 609–624.
[29] Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., Éigeartaigh, S. Ó., Beard, S., Belfield, H., Farquhar, S., Amodei, D. (2018). The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation.

A particular danger to **international security and peace** lies in seeing the development of AI as a competitive race. AI will not only lead to undesirable side effects, but also empower malicious actors ranging from cybercriminals to totalitarian states in their desire to control populations.

Digital power concentration

Many AI-applications are developed and deployed by only **a handful of large private actors**, sometimes referred to as the Big Five, GAFAM or even Frightful Five.[30] If too much political power is concentrated in a few private hands which prioritise shareholder-value over the common good, this can threaten the authority of democratic states.

6. AI & Rule of Law

Public institutions are held to a higher standard when it comes to their behaviour towards individuals and society, which is reflected in principles such as justification, proportionality and equality. AI can increase the efficiency of institutions, yet on the other it can also erode the procedural **legitimacy of and trust in democratic institutions and the authority of the law.**

Courts, law enforcement and public administrations could become more efficient, yet at the cost of being more opaque and less human agency, autonomy and oversight.[31]

Similarly, whereas previously courts were the only ones to determine what counts as illegal hate speech, today mostly private AI systems determine whether speech is taken down by social media platforms.[32] These AI systems de facto compete for authority with judges and the law and In general, AI can contribute to developing judicial systems that operate outside the boundaries and protections of the rule of law.

Automated online dispute resolutions provided by private companies are governed by the terms of service rather than the law that do not award consumers the same rights and procedural protections in public courts.[33]

---

[30] Google, Facebook, Microsoft, Apple and Amazon. See: Nemitz, P. (2018). Constitutional democracy and technology in the age of artificial intelligence. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376(2133), 20180089. Webb, A. (2019). The Big Nine: How the tech titans and their thinking machines could warp humanity.
[31] Danaher, J. (2016). The Threat of Algocracy: Reality, Resistance and Accommodation. Philosophy & Technology, 29(3), 245–268. Weizenbaum, J. (1976). Computer Power and Human Reason: From Judgment to Calculation. W. H. Freeman.
[32] Cohen, J. E. (2019). Between truth and power: The legal constructions of informational capitalism.
[33] Susskind, J. (2018). Future politics: Living together in a world transformed by tech.

The European Commission for the Efficiency of Justice already in 2018 outlined **5 principles for the use of AI in the judiciary** in the "European Ethical Charter on the use of AI in the judicial systems and their environment". The High Level Expert Group on AI has called for public bodies to be held to the **7 Requirements for Trustworthy AI** when developing, procuring or using AI. Similar principles and requirements should be imposed on law enforcement agencies.

However, AI can not only threaten the rule of law, it could also strengthen it.[34] If developed and used responsibly, it can empower agencies to identify corruption with the state.[35] Similarly, AI can either be used to detect and defend against cyberattacks.[36]

III. How to address the impact of AI on Human Rights, Democracy and Rule of Law?

The impact of AI on human rights, democracy and the rule of law has been receiving more attention lately, most prominently, in the recent Whitepaper on AI of the European Commission. How to **address** the impact, however, remains uncertain territory to date. This Chapter describes possible strategies that can be followed within the existing framework of human rights, democracy and the rule of law. These strategies are not exhaustive and should help move the discussion towards the next phase.

1. Putting human rights in an AI context

Many AI developers, deployers and users (public and private) seem to be unaware of the (potential) impacts of AI on Human Rights. As a first step, an iteration or (re)articulation exercise in which existing Human Rights of the ECHR are 'translated' to an AI context, is very useful and could be done by means of a Framework Convention.

2. Measures for compliance, accountability and redress

To properly address the impact of AI on existing human rights, democracy and the rule of law, certain existing compliance, accountability and redress mechanisms could be further developed and new mechanisms could be introduced. What is important however, is that the use of AI is often hidden or unknown, making it difficult or impossible to know whether there is

[34] Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M., & Nerini, F. F. (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. Nature Communications, 11(1), 1–10.
[35] West, J., & Bhattacharya, M. (2016). Intelligent financial fraud detection: A comprehensive review. Computers & Security, 57, 47–66. Hajek, P., & Henriques, R. (2017). Mining corporate annual reports for intelligent detection of financial statement fraud – A comparative study of machine learning methods. Knowledge-Based Systems, 128, 139–152.
[36] Taddeo, M., & Floridi, L. (2018a). Regulate artificial intelligence to avert cyber arms race. Nature, 556(7701), 296–298.

an impact on human rights, democracy and the rule of law in the first place. Measures for compliance, accountability and redress should thus start with the **obligation of transparency about the use of AI** systems, which can impact human rights, democracy or the rule of law. This includes an AI registry, which then specifies the risk class and required amount of transparency and accountability for a particular application.

Compliance could then start with what has recently been described as a new culture of **"Human Rights, Democracy and Rule of Law by design"**.[37] In such a culture, developers, deployers and uses of AI, from the outset would reflect on how the technology might affect human rights, democracy and the rule of law and adjust the technology or its use accordingly. This could be underpinned by a (legal) obligation to perform an **AI Human Rights, Democracy and Rule of Law Impact Assessment**.

Such a new culture would need to include the **obligation to account** for the appropriate structure to be put in place, but also for the outcomes of the AI Human Rights, Democracy and Rule of Law Impact Assessment as well as the design and governance decisions based thereon.

Redress in light of AI impact on human rights entails access to justice and effective remedy. As far as access to justice goes, it might be too soon to determine whether this is sufficiently guaranteed when it comes to AI and human rights impact. Only just recently have we seen the first couple of judgements by domestic courts on the (potential) impact of AI on one particular human right, the right to privacy of art. 8 ECHR.[38]

More importantly however, **access to justice is challenged when many AI-applications are developed and deployed by only a handful of large private actors**. These companies dominate both the development of AI as well as the (eco)systems AI operates in and on. While states are obliged to protect individuals and groups against breaches of human rights perpetrated by other actors, appreciation of non-state actors' influence on human rights has steadily grown.[39] As these large tech companies have now become operators that are capable of determining and perhaps altering our social and even democratic structures, the impact of their AI(-use) on human rights becomes more prevalent. In this respect, AI might serve as a good opportunity and think of a structure that would legally oblige private actors to comply

[37] Nemitz, P. (2018). Constitutional democracy and technology in the age of artificial intelligence.
[38] A Dutch court has considered a law that allowed public institutions to potentially use AI to predict fraud with social benefits in violation of the right to a private life of art. 8 ECHR, the UK High Court in Cardiff accepted that facial recognition affects art. 8 ECHR, as it enables the extraction of "intrinsically private" information, but it considered the use lawful for proportionality reasons. A French Court considered the use of facial recognition in schools in violation with art. 8 ECHR.
[39] Business and Human Rights, A Handbook for Legal Practitioners, Claire Methven O'Brien, Council of Europe

with human rights and to grant access to justice if they fail to do so.[40] The basic question is whether to a) accept the private power of AI companies and to make sure they use it responsibly, or to b) challenge it and try to reassert the power of the state.

When it comes to an effective remedy, AI is a topic where, as Sheldon also observed, remedies are 'not only about making the victim whole; they express opprobrium to the wrongdoer from the perspective of society as a whole' and thus 'affirm, reinforce, and reify the fundamental values of society'.[41] The European Court of Human Rights has stressed in its *Broniowski* judgment, that international law requires that **'individual and general redress (…) go hand in hand'**.[42]

To determine an effective remedy in case of a human rights violation as a result of AI, one thus needs to look at **both individual and general remedies**. Moreover, because AI has a myriad of applications, ranging from surveillance and identification, to profiling, nudging and decision making, remedies need to be tailored towards those different applications. Proper remedies should include cessation of unlawful conduct and guarantees of non-repetition, where states could for example be obliged to adopt and implement enforceable legislation to protect human rights from future AI impacts. The obligation to repair the injury or damage caused by the violation, either to an individual or to a community, should exist. For some AI applications just ensuring an effective remedy might not be sufficient to address the human rights impact of that application. More far-reaching measures, such as a ban or restrictive use might be necessary (see Chapter IV).

3. Protecting democratic structures and the rule of law

To prevent systemic failure or disruption due to centralisation of AI-driven decision making processes in vital structures, distributed decision making processes, rather than centralised should be implemented to prevent risk of catastrophic failure. **These processes should have proper structures of human oversight built in**.[43]

Human oversight helps ensure that an AI system does not undermine human autonomy or causes other adverse effects. Oversight may be achieved through governance mechanisms such as **human-in-the-loop (HITL), human-on-the-loop (HOTL), or human-in-command (HIC)**

[40] This means going beyond merely referring to the Recommendation CM/Rec(2016)3 on human rights and business of the Committee of Ministers of the Council of Europe (and the UN Guiding Principles on Business and Human Rights)
[41] Dinah Shelton, 'The Right to Reparations for Acts of Torture: What Right, What Remedies?', 17(2) *Torture* 96 (2007), at 96
[42] Broniowski v. Poland, ECHR
[43] Ethics Guidelines for Trustworthy AI, High Level Expert Group on AI, 2019

approach. HITL refers to the capability for human intervention in every decision cycle of the system, which in many cases is neither possible nor desirable. HOTL refers to the capability for human intervention during the design cycle of the system and monitoring the system's operation. HIC refers to the capability to oversee the overall activity of the AI system (including its broader economic, societal, legal and ethical impact) and the ability to decide when and how to use the system in any particular situation. This can include the decision not to use an AI system in a particular situation, to establish levels of human discretion during the use of the system, or to ensure the ability to override a decision made by a system.

To address the risk of inequality, governments need to actively halt the use of AI applications that increase inequality.

Preventing election influence or public manipulation through AI-driven personalised information is not an easy task. Regulations for online campaigning, either for the (social) media platforms or for political parties, could be considered. Obviously, this raises questions regarding the freedom of speech. Keeping humans in/on the loop and in command (see above) could help detect and eliminate undesirable voter influencing.

A crucial leverage in ensuring responsible use of AI in public services is public procurement. If the legally binding requirements for **public procurement** are updated to include criteria such as fairness, accountability and transparency in AI this can serve two purposes. On the one hand, it ensures that governments strictly only use systems that are compatible with the rule of law, but also creates economic incentives for the private sector to develop and use systems that comply with the principles of the rule of law. Furthermore, the use of AI in government should be subject to oversight mechanisms, including **court orders and ombudspersons** for complaints.

IV. What if current human rights, democracy and the rule of law fail to adequately protect us?

1. Question Zero

Due to the invasiveness of some AI-applications or uses, there might be situations in which our current framework of human rights, democracy and the rule of law fails to adequately or timely protect us and where we might need to pause for reflection and find the appropriate answer to what one could consider **"question zero": Do we want to allow this particular AI-system or**

use and if so, under what conditions? Answering this question should force us to look at the AI-system or use from all perspectives, which could result in several 'solutions':

- A particular AI-system or use is put under a moratorium, (temporarily or indefinitely) banned or put under restrictions ("Red Lines")
- New Human Rights are introduced as safeguards against the 'new' adverse impact of AI
- Existing Human Rights are adapted to allow for responsible development and use of AI
- A particular AI-system or use is made subject to a specific democratic oversight-mechanism
- Private owners of powerful AI-systems are obliged to align their AI development and governance structures with the interests of those affected by the system and society at large, which could include measures to involve relevant parties (such as workers, consumers, clients, citizens, policy makers)

First and foremost, 'AI impact' is to be considered **both at individual and at societal/collective level** whereas AI can impact both the individual as well as larger parts of our collective society. Secondly, **context, purpose, severity, scale and likelihood** of the impact is important to determine the appropriate and proportionate action. For AI applications that generate unacceptable risks or pose threats of harm or systemic failure that are substantial, a precautionary and principle-based regulatory approach should be adopted. For other AI applications a risk based approach could be more appropriate.

2. Red Lines

Red lines could be drawn for certain AI-systems or uses that are considered to be too impactful to be left uncontrolled or unregulated or to even be allowed. These AI-applications could give rise to the necessity of **a ban, moratorium and/or strong restrictions or conditions for exceptional and/or controlled use**:

- Indiscriminate use of facial recognition and other forms of biometric recognition either by state actors or by private actors
- AI-powered mass surveillance (using facial/biometric recognition but also other forms of AI-tracking and/or identification such as through location services, online behaviour, etc.)
- Personal, physical or mental tracking, assessment, profiling, scoring and nudging through biometric and behaviour recognition
- AI-enabled Social Scoring
- Covert AI systems and deep fakes
- Human-AI interfaces

Exceptional use of such technologies, such as for national security purposes or medical treatment or diagnosis, should be evidence based, necessary and proportionate and only be allowed in controlled environments and (if applicable) for limited periods of time.

3. Some adapted or new human rights

In addition to Red Lines-measures, the following adapted or new Human Rights could be considered (non-exhaustive):

- A right to human autonomy, agency and oversight over AI
- A right to transparency/explainability of AI outcomes, including the right to an explanation of how the AI functions, what logic it follows, and how its use affects the interests of the individual concerned, even if the AI-system does not process personal data, in which case there is already a right to such information under GDPR.[44]
- A separate right to physical, psychological and moral Integrity in light of AI-profiling, affect recognition
- A strengthened right to privacy to protect against AI-driven mass surveillance
- Adapting the right to data privacy to protect against indiscriminate, society-wide online tracking of individuals, using personal and non-personal data (which often serves as a proxy for personal identification)

Diverging from these rights in exceptional circumstances such as for security purposes should only be allowed under strict conditions and in a proportionate manner.

4. Future scenarios

Extrapolating into the future with a longer time horizon, certain critical long-term concerns can be hypothesized and are being researched, necessitating a risk-based approach in view of possible unknown unknowns and "black swans". While some consider that Artificial General Intelligence, Artificial Consciousness, Artificial Moral Agents, Super-intelligence can be examples of such long-term concerns (currently non-existent), many others believe these to be unrealistic. Nevertheless, close monitoring of these developments is necessary in order to determine whether ongoing adaptations to our human rights, democracy and rule of law systems are necessary.

[44] Nemitz, P. (2018). Constitutional democracy and technology in the age of artificial intelligence.