Artificial Intelligence and Human Rights: A Business Ethical Assessment

Alexander KRIEBITZ* and Christoph LÜTGE**

Abstract

Artificial intelligence (AI) has evolved as a disruptive technology, impacting a wide range of human rights-related issues ranging from discrimination to supply chain due diligence. Given the increasing human rights obligations of companies and the intensifying discourse on AI and human rights, we shed light on the responsibilities of corporate actors in terms of human rights standards in the context of developing and using AI. What implications do human rights obligations have for companies developing and using AI? In our article, we discuss firstly whether AI inherently conflicts with human rights and human autonomy. Next, we discuss how AI might be linked to the beneficence criterion of AI ethics and how AI might be applied in human rights-related areas. Finally, we elaborate on individual aspects of what it means to conform to human rights, addressing AI-specific problem areas.

Keywords: artificial intelligence, corporate ethics, data privacy, digitization, human rights

I. Introduction

Artificial intelligence (AI) is increasingly conquering our reality and shaping how societies and their institutions are maintained, organized and controlled, ranging from face recognition tools to autonomous vehicles, search engines, translation tools and programs predicting price developments in stock markets. When compared with conventional technologies, AI excels, in terms of interpreting and reacting to data, which (for AI purposes) is documented, generated and stored in electronic devices; the data begin communicating with each other and generating what we term 'big data'. In this sense, we can aptly describe AI as a constellation of different processes and technologies,[1] leading

* Research Associate at the Chair of Business Ethics, Technical University of Munich, Munich, Germany. Alexander Kriebitz declares no conflict of interest.

** Chair Holder and Full Professor at the Chair of Business Ethics, Technical University of Munich, Munich, Germany. As Professor at Technical University of Munich and as director of the Institute of Ethics in Artificial Intelligence, Christoph Lütge received project funding from Facebook Inc., Fujitsu K.K. and Huawei Technologies Co. Ltd related to research on artificial intelligence and ethics.

[1] David Kaye, 'Mandate of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression', Open Letter to Office of the High Commissioner for Human Rights (1 June 2017), https://www.ohchr.org/Documents/Issues/Opinion/Legislation/OL-DEU-1-2017.pdf (accessed 27 November 2019). Josh Cowls et al, 'Designing AI for Social Good: Seven Essential Factors' (15 May 2019), https://dx.doi.org/10.2139/ssrn.3388669 (accessed 27 November 2019).

Business and Human Rights Journal, 5 (2020), pp. 84–104
doi:10.1017/bhj.2019.28

to an incremental substitution of human actions by automated data processing. Although AI clearly offers major advantages for humankind, in the sense of more accurate diagnostic tools, enhanced measures to combat crime and curb terrorism, critics still point to the risks that may accompany this technological revolution. The Open Letter on AI of 2015, signed by major scientists and businesspeople, has sparked an intensive debate on how to regulate AI and how to avoid potential pitfalls attributed to the mismanagement of this technology.[2] In this context, Stephen Hawking referred to AI as potentially the worst event in human history,[3] capable of spelling the end of humankind, while other prophecies about the technology related to AI sound as ominous as the warnings offered in Orwell's *Nineteen Eighty-Four* or Huxley's *Brave New World.*

The uncertainties accompanying this period of technological change call for intensive debate on how to steer the further development of AI, as well as its ethics and governance; these debates pose new questions concerning the design of ethical frameworks[4] and legislation all across the globe. Although AI certainly contributes to the realization of various social and environmental goals, such as the United Nations (UN) Social Development Goals, there remains a risk of conflict between the normative foundations of our civilization and factual use of AI. Therefore, legislators and ethicists worldwide have begun to develop legal norms and standards to tackle potential cases of misuse of AI and to regulate the matter. These include, for instance, the Montreal Declaration for Responsible AI, the Asilomar AI Principles, the AI4People's principles for AI ethics,[5] the two High-Level Expert Groups on AI's reports, on ethics[6] as well as governance of AI,[7] the House of Lords Artificial Intelligence Committee, the GDPR and the German Ethics Code for Automated and Connected Driving, which entail important aspects of ethical issues related to AI. In addition to the UN Report on Artificial Intelligence and its implications for human rights, a legal or ethical codification tailored to the application of AI in the context of human rights has yet to be articulated. Human rights, however, play an essential role in the context of AI governance, as they are regarded as fundamental norms of Western civilization and play an increasing role, generally, in international law.[8] Apart from the ethical duties of States and international organizations such as the UN in safeguarding and protecting human rights, the focus of human rights has been gravitating towards to the enforcement of human rights by companies, most prominently after the formulation of the

[2] Matthey Sparkes, 'Top Scientists Call for Caution over Artificial Intelligence', *The Telegraph* (13 January 2015), https://www.telegraph.co.uk/technology/news/11342200/Top-scientists-call-for-caution-over-artificial-intelligence.html (accessed 27 November 2019).

[3] 'Stephen Hawking Warns Artificial Intelligence Could End Mankind', *BBC* (2 December 2014), https://www.bbc.com/news/technology-30290540 (accessed 27 November 2019).

[4] Compare, e.g., Luciano Floridi et al, 'AI4People – An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations' (2018) 28 *Minds & Machines* 4, 689 and Stuart Russell et al, 'Research Priorities for Robust and Beneficial Artificial Intelligence' (2015) 36 *Artificial Intelligence Magazine* 4, 105.

[5] Floridi et al, note 4.

[6] High Level Expert Group on Artificial Intelligence, 'Ethics Guidelines for Trustworthy AI' (8 April 2019), https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419 (accessed 27 November 2019).

[7] Floridi et al, note 4, as well as Josh Cowls and Luciano Floridi, 'Prolegomena to a White Paper on an Ethical Framework for a Good AI Society' (2019), https://dx.doi.org/10.2139/ssrn.3198732 (accessed 27 November 2019).

[8] Bruno Simma and Dirk Pulkowski, 'Of Planets and the Universe: Self-Contained Regimes' (2006) 17 *European Journal of International Law* 3, 483.

https://doi.org/10.1017/bhj.2019.28 Published online by Cambridge University Press

Ruggie principles.[9] Given the increasing obligations and duties of companies as stewards of human rights,[10] this article sheds light on the responsibilities of corporate actors to the enforcement and realization of human rights standards in the context of AI. What implications do human rights obligations have for the way companies are developing and using AI?

In a broader sense, our article aims to connect the discourse on AI ethics to the discourse on the human rights obligations of corporate decision makers. In our view, both discourses do not represent competing or exclusive views, but rather complement and enrich each other as they integrate the larger domains of business ethics and technology ethics.

II. What are Human Rights, and Why do they Matter?

Addressing AI from a human rights perspective requires a short description of the concept of human rights. In Western thought, human rights are regarded as the supreme norm of law and they form the basis for most legal systems. According to the majority of experts on international law,[11] human rights are not merely an enumeration of individual rights, but rather form a self-contained regime. The integral pillar of this regime is an anthropology based on the self-determination and autonomy of the human being.[12]

According to this understanding, human rights oblige the state and other social organizations to observe certain principles and procedures when dealing with subordinates; these principles encompass, for example, strict adherence to the rule of law principle and the right to a fair trial. At the same time, the philosophy of human rights views freedom as the basic condition of human beings, concluding that restrictions to this freedom must serve the common good, and not the will of a monarch or tyrant. This concept largely corresponds to the notion espoused by Isaiah Berlin, who defined freedom as 'the absence of obstacles to possible choices and activities',[13] and who contributed to the understanding of human rights as 'claim rights' limiting the power of the state.[14]

Under these circumstances, interventions in the autonomy of the individual are only legitimate if they are based on the consent of the individual concerned, or if the liberty of one individual conflicts with the interest of others. The transfer of property or the implementation of a medical treatment – an intervention in the inviolable integrity of

[9] United Nations Guiding Principles on Business and Human Rights: Implementing the United Nations 'Protect, Respect and Remedy' Framework, https://www.ohchr.org/documents/publications/Guidingprinciples Businesshr_eN.pdf (accessed 27 November 2019)

[10] Compare Nicola Jägers, *Corporate Human Rights Obligations* (Cambridge: Intersentia, 2002); Michael A Santoro, *Profits and Principles: Global Capitalism and Human Rights in China* (Ithaca: Cornell University Press, 2000); Florian Wettstein, *Multinational Corporations and Global Justice* (Bibliovault OAI Repository, the University of Chicago Press, 2009); John Gerard Ruggie, *Just Business: Multinational Corporations and Human Rights* (New York: W.W. Norton & Company, 2013); Peter Muchlinsky, 'Implementing the New UN Corporate Human Rights Framework. Implications for Corporate Law, Governance, and Regulation' (2002) 22 *Business Ethics Quarterly* 1, 145–177.

[11] Simma and Pulkowski, note 8. Stephen Gardbaum, 'Human Rights as International Constitutional Rights' (2008) 19 *European Journal of International Law* 4, 749.

[12] Immanuel Kant, *Groundwork of the Metaphysics of Morals* (Cambridge: Cambridge University Press, 1998).

[13] Isaiah Berlin, 'Two Concepts of Liberty', in *Four Essays on Liberty* (Oxford: Oxford University Press, 1969), 118.

[14] Wesley Hohfeld, 'Fundamental Legal Conceptions as Applied in Judicial Reasoning' (1917) 26 *Yale Law Journal* 8, 710.

https://doi.org/10.1017/bhj.2019.28 Published online by Cambridge University Press

the body – are only lawful if they enjoy the explicit consent of the individual, some necessary exemptions, such as emergencies, notwithstanding. The main exception, which allows constraining freedom of one person, requires that it service the prevention of harm to third parties. According to the harm principle, which forms the basis of human rights as claim rights that explicitly bind institutions and other third parties, 'the only purpose for which power can be rightfully exercised over any member of a civilised community, against his will, is to prevent harm to others.'[15] As a result, incursions of the state, limiting the freedom of individuals, face substantial restrictions and are only legitimate in cases of norm collision. In this way, traffic regulation, insofar as it constitutes a limitation to individual freedom, serves to minimize traffic accidents, ultimately deriving from the task of the state to protect human lives.

However, in accordance with the principle of proportionality, the interference of the state must be proportionate to the damage averted. This notion derives from the high privilege accorded to the idea of equality before the law, which can be traced back to the Aristotelian idea of 'corrective justice',[16] implying that the harm initiated by a regulation or an act of the state must be proportionate to the harm avoided. An important aspect of this term is the Weberian concept of *Augenmaß*[17] – a sense of proportion or 'common sense' – which should guide political actions and legislators.[18] Imprisoning a child for stealing an apple or withdrawing a driver's license after he or she has exceeded the speed limit by 5 km/h would be examples of disproportionate interference in the liberties of individuals.

However, there are cases in which the principle of proportionality is not applicable, as it is not possible to derogate specific types of human rights. According to the majority of human rights experts, the very nature of human rights prohibits particular actions, such as torture, slavery, rape or extremely humiliating behaviour, even if they serve other fundamental rights.[19]

From this perspective, we derive some basic implications of human rights for the regulation of AI:

(a) The rights of an individual can be transferred only by his or her consent (principle of consent).

(b) The only justification for the use of power against the will of a person is the prevention of harm (harm principle).

(c) The use of force must be proportionate to the threat (principle of proportionality).

---

[15] John S Mill, *On Liberty* (London: Longman, Roberts & Green, 1859).

[16] Gerhard M Ambrosi, 'Aristotle's Geometrical Model of Distributive Justice' (2007), https://www.uni-trier.de/fileadmin/fb4/prof/VWL/EWP/Publikationen/Ambrosi/Aristotle-4.pdf (accessed 27 November 2019); Louis P Pojman, *Ethical Theory: Classical and Contemporary Readings* (Belmont, CA: Wadsworth Publishing, 1996).

[17] Max Weber, 'Politik als Beruf', in *Geistige Arbeit als Beruf. Vier Vorträge vor dem Freistudentischen Bund* (Munich, Germany: Duncker & Humblot, 1919).

[18] Lorenzo Zucca, *Constitutional Dilemmas: Conflicts of Fundamental Legal Rights in Europe and the USA* (Oxford, UK: Oxford University Press, 2007); Tor-Inge Harbo, 'The Function of the Proportionality Principle in EU Law' (2010) 16 *European Law Journal* 2, 158; Paul Craig and Grainne de Bruca, *EU Law – Text, Cases and Materials* (Oxford, UK: Oxford University Press, 2015).

[19] Committee against Torture, 'Concluding Observations on the Fourth Periodic Report of the United Kingdom', UN Doc. CAT/C/CR/33/3 (25 November 2004), para 4(a)(i).

Human rights, however, transcend a purely defensive character to influence the objectives of social organizations, as in the form of 'state objectives' or imperatives.[20] The state, as the highest normative authority, and international organizations thereby confront the task of guaranteeing human rights on a material and economic basis. The Covenant on Social, Economic and Cultural Rights, for example, contains the principle of progressive realization urging states to enact 'policies and techniques to achieve steady economic, social and cultural development […]' (Article 6.2).[21] Finally, the notion of human rights and the idea that human beings are born free and equal requires the *anchoring of participation* in the political process. This notion is largely equivalent to the Lockean principle of government as underpinned by the 'consent by the governed' mentioned in the American Declaration of Independence.

Due to the increasing role of enterprises within international law in the wake of globalization, John Ruggie writes that enterprises became part of the human rights discourse.[22] The UN Global Compact, the ILO Declaration on Fundamental Principles and Rights at Work, the UK Modern Slavery and Human Trafficking Act, as well as the UN Guiding Principles on Business and Human Rights (UNGPs) have largely influenced this development. The UNGPs, according to Ruggie, provide that companies should 'identify and access any actual or potential adverse human rights impact with which they may be involved either through their own activities or as a result of business relationships'. Hence, attention is being drawn to the responsibilities of companies in regard to the above-mentioned 'claim rights' in an international context.[23]

III. What Distinguishes AI from Other Technologies?

The discussion of the relationship between human rights and AI requires an examination of the properties and peculiarities of AI. What makes AI different from other technologies, such as traditional vehicles, smartphones or nuclear power plants? Why do we need an ethics tailored to AI at all? The definition closest to our understanding of AI is given by the Merriam-Webster Dictionary, which defines artificial intelligence as the capability of a machine to imitate intelligent human behaviour. Strictly speaking, the word 'intelligent' does not refer to the machine, but rather to the fact that if the task of the AI solution would have been solved by a human being, the mode of accomplishing

[20] Patricia C Kuszler, 'Global Health and the Human Rights Imperative' (2007) 2 *Asian Journal of WTO & International Health Law and Policy* 1, 99.

[21] United Nations Human Rights Office of the High Commissioner, 'International Covenant on Economic, Social and Cultural Rights', https://www.ohchr.org/en/professionalinterest/pages/cescr.aspx (accessed 27 November 2019).

[22] John Gerard Ruggie, *Just Business: Multinational Corporations and Human Rights* (New York: W.W. Norton & Company, 2013); JA Zerk, *Multinationals and Corporate Social Responsibility. Limitations and Opportunities in International Law* (Cambridge: Cambridge University Press, 2006), p. 19.

[23] Human Rights Council, 'UN Guiding Principles on Business and Human Rights', https://www.business-humanrights.org/sites/default/files/reports-and-materials/Ruggie-report-7-Apr-2008.pdf (accessed 27 November 2019).

https://doi.org/10.1017/bhj.2019.28 Published online by Cambridge University Press

the task would have been called intelligent.[24] The term comparison to human intelligence refers therefore, in the first instance, to the output of an action and not to the input or to the process of the decision-making processes in machines.[25] The Open Letter on AI[26] here refers to the statistical and economic notions of intelligence. This understanding has an important implication for human rights, as AI – not being an ontological entity – cannot be regarded as an independent actor or potential perpetrator of human rights violations, at least not yet. Instead, human rights compliance that relates to AI solutions remains in the domain of human responsibility and works to bind nation states, companies or non-governmental organizations (NGOs), using these technologies.

Based on this conceptualization of AI, human rights violations may originate in different impulses and inclinations. While imitating intelligent behaviour, AI combines large amounts of data with fast, iterative processing and intelligent algorithms, allowing the software to learn automatically from patterns or features in the data. In this sense, AI is working to obviate the need of collecting and interpreting data, by using neural networks, and may therefore develop conclusions unforeseen by humans, as they are excluded from the definition of the objectives and outputs of AI.[27] At this point, the special feature of AI is that some of its processes run automatically, and in these, humans cannot intervene directly; they also cannot be foreseen *ex-ante*, resulting in unintended consequences.

These characteristics pose questions more general to its conformability with human rights, namely, whether AI, insofar as it consists of the transfer of human agency to a machine, represents an inherent conflict with the idea of moral self-determination. If this does not pose an inherent conflict, questions remain about the ways that certain characteristics and aspects of AI may affect human rights. The following questions shed light on some of these issues:

- Are there scenarios where AI has a positive impact on human rights?
- Are there scenarios where the data input of AI violates human rights?
- Are there scenarios where the output of AI violates human rights?
- Does the usage of AI in specific domains violate human rights, most notably, participation rights?
- Is it possible to use AI to violate or constrain human rights?

[24] David Kaye, 'Mandate of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression', Open Letter to Office of the High Commissioner for Human Rights (1 June 2017), https://www.ohchr.org/Documents/Issues/Opinion/Legislation/OL-DEU-1-2017.pdf (accessed 27 November 2019) and Cowls et al, note 1 and John McCarthy, 'Programs with Common Sense', in *Proceedings of the Teddington Conference on the Mechanization of Thought Processes* (City: Publisher, 1959), 75.

[25] John McCarthy, 'Programs with Common Sense', in *Proceedings of the Teddington Conference on the Mechanization of Thought Processes* (City: Publisher, 1959), 75.

[26] Stuart Russell et al, 'Research Priorities for Robust and Beneficial Artificial Intelligence', *AI Magazine* (2015), https://futureoflife.org/ai-open-letter/ (accessed 27 November 2019).

[27] David Kaye, 'Mandate of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression', Open Letter to Office of the High Commissioner for Human Rights (1 June 2017), https://www.ohchr.org/Documents/Issues/Opinion/Legislation/OL-DEU-1-2017.pdf (accessed 27 November 2019).

https://doi.org/10.1017/bhj.2019.28 Published online by Cambridge University Press

IV. Artificial Intelligence and Human Rights – an Inherent Conflict?

In the first part of our analysis, we address the question of whether there is an inherent conflict between the usage of AI and human rights. This requires a clarification of the term 'inherent' used here. In general, we can distinguish between acts that inherently conflict with human rights and acts that represent a contingent conflict with human rights.[28] In our view, an act is inherently at odds with human rights if it constitutes a human rights violation, regardless of the circumstances. A paradigmatic example is slavery, which controverts the nature of human rights and self-determination, regardless of the exact circumstances and causes (cf. Article 4, International Covenant on Civil and Political Rights). The construction of a road may be one example to illustrate the difference between inherent and contingent limits. Usually, there is nothing wrong with constructing roads. If, however, the government constructs the road using forced labour or by the expropriation of native tribes, this represents a clear violation of human rights. Cases of this kind constitute human rights violations for contingent reasons, as their ethical assessment depends not on the action *per se*, but rather on its circumstances.

Does the usage of AI constitute an inherent violation of human rights, thereby justifying a universal prohibition? The reason that we address this rather theoretical question owes to the general rejection of the usage of AI by some ethicists and religious entities, positing the existence of an insurmountable conflict between moral self-determination and the use of AI. A frequent line of argument is that the use of AI represents a conflict with human autonomy, because even weighty decisions may be taken over by AI, thereby standing in direct conflict with the very meaning of human rights and leading to alienation.[29] In the statement of the Southern Baptist 'Convention Artificial Intelligence: An Evangelical Statement of Principles', the church positioned itself against the assignment of AI to a level of human identity, worth, dignity, or moral agency.[30]

To analyse whether this claim is valid, we refer to one of the more controversial examples, namely, the use of AI in weighing life-or-death decisions. In autonomous driving, as in other applications, actions such as steering the vehicle or slowing down, originally performed by humans, are increasingly managed by mechanical and automatic processes. Moreover, software solutions may intervene, in the case of an accident, to minimize the number of casualties. In the event of an unavoidable crash, if the car must choose between hitting one of two individuals crossing a street, the autonomous car becomes the deciding entity, on which target to hit. Beyond general concerning accountability and responsibility considerations, the permissibility of automation in

---

[28] Compare Jason Brannon and Peter M Jaworski, *Markets without Limits: Moral Virtues and Commercial Interests* (Abingdon, UK: Routledge, 2015) and Julian F Mueller, 'The Ethics of Commercial Human Smuggling' (2018) *European Journal of Political Theory* 1–19.

[29] IAP Wogu, 'Artificial Intelligence, Alienation and Ontological Problems of Other Minds: A Critical Investigation into the Future of Man and Machines', Conference or Workshop Item (2018), http://eprints.covenantuniversity.edu.ng/id/eprint/11499 (accessed 27 November 2019).

[30] The Ethics & Religious Liberty Commission of the Southern Baptist Convention, *Artificial Intelligence: An Evangelical Statement of Principles* (2019), https://erlc.com/resource-library/statements/artificial-intelligence-an-evangelical-statement-of-principles (accessed 27 November 2019).

https://doi.org/10.1017/bhj.2019.28 Published online by Cambridge University Press

such instances matters for human rights human rights perspective,[31] as it relates to the question of whether AI is able to make decisions involving the life and death of people. This, in turn, connects to the larger debate on the relationship between human dignity and automatization. From a strict deontological position, weighing up lives cannot be legitimate based on the assumption that it conflicts with the idea of human dignity, in the sense that human beings should not be objectified; whereas utilitarian considerations would urge the programmer to choose the alternative with the lesser casualties.

From our point of view, however, one practical and one theoretical argument speak against the claim that AI usage in such cases constitutes an inherent breach of human rights. In practice, the decision of a given driver does not usually constitute a well-crafted thought process, but rather an unconscious or panicked reaction. It is therefore questionable whether the act really entails a moment of human agency. After reaching a certain level of development, sensory and mechanical processes may be superior to human reactions, as they could move much faster than any human brain – even in crisis. From a theoretical point of view, the Rawlsian concept of a veil of ignorance[32] offers a way out of the dilemma, by demonstrating that the principle of consent can harmonize with the maximization of the right to life. If we imagine the concrete scenario of an unavoidable crash that involves the death of the driver or the death of a group of people[33] and we further assume that all individuals have a desire to save their own life first, then it follows that the involved parties would be unable to reach a unanimous consensus. Due to the high value of life, it is also questionable whether an individual would be willing to give up his or her life for the sake of others, empirical findings notwithstanding.[34] Obviously, the prospect of reaching an agreement in this concrete situation is doomed to failure, without sacrificing the principle of unanimous consent. Although this is possible, according to some largely utilitarian theories, the application of a decision based on the – hypothetical – will of the majority remains contentious.[35]

The only way out of this gridlock situation is to shift the focus from the individual case to a general rule. Given the premise that anonymous individuals do not know their position in advance and must agree *ex-ante* on a procedure[36] on how to deal with the use of AI in unavoidable situations, they are likely to agree on abstract and impartial principles. That individuals are unaware of their exact role in the scenario establishes a setting that is conducive to impartiality, ensuring 'that no one is advantaged or disadvantaged in the choice of principles by the outcome of natural chance or the contingency of social circumstances.'[37] Based on this, reasonable individuals may propose a regulation, according to which a car shall be programmed to minimize the number of casualties in

[31] This links up to the Asilomar principle 'Non-subversion and Human Control'.

[32] John Rawls, *A Theory of Justice* (Cambridge: Belknap Press of Harvard University Press, 1971).

[33] Wendell Wallach and Colin Allen, *Moral Machines: Teaching Robots Right From Wrong* (Oxford: Oxford University Press, 2008).

[34] Ezio Di Nucci, 'Self-sacrifice and the trolley problem' (2013) 26 *Philosophical Psychology* 5, 662.

[35] Noah J Goodall, 'Machine Ethics and Automated Vehicles', in Gereon Meyer and Sven Beiker (eds.), *Road Vehicle Automation* (Springer, 2014), 93 and Edmond Awad et al, 'The Moral Machine Experiment' (2018) 563 *Nature*, 59.

[36] John Rawls, *A Theory of Justice* (Cambridge: Belknap Press of Harvard University Press, 1971), note 29.

[37] Ibid.

https://doi.org/10.1017/bhj.2019.28 Published online by Cambridge University Press

unavoidable dilemma situations, as the overall probability of being hit or killed in an autonomous car decreases with a diminishing number of casualties. The rule must therefore imply that the larger group of people is always saved, after weighing up decisions, regardless of their status as drivers, pedestrians or other personal features. This conforms to the principle of human dignity, which asserts that all individuals involved should have equal chances to life. Lifting the question of programming unavoidable accidents to the level of social participation seems to be more meaningful, because the formulation of the rule has an impact on every individual participating in road traffic. The exact choice of the mechanisms and whether to apply randomization in cases when the car must decide between two groups of the same size is ultimately a question of social and democratic consent. This implies that companies are unable to formulate own interpretations of human rights frameworks for such instances and need to refer to the overarching legal frameworks or the rulings of constitutional courts. Moreover, the programming must distinguish between situations involving anonymous people and situations in which people know each other to avoid extreme harm (compare Rule 3). Therefore, the ultimate decision, perhaps in the form of a parliamentary decision and finally in its implementation by companies, must fulfil the following criteria:

- Rule 1: The likelihood of being killed in an accident must decrease for all persons.
- Rule 2: All individuals must have equal chances to survive.
- Rule 3: Severe hardships for individuals may overrule Rule 2, if it is based on the explicit consent of the parties affected (for example, the grandmother consents that the car hits her instead of her grandchild.)

The principle of 'meta-autonomy', which refers to the decision to delegate specific decisions to machines,[38] is not a very new concept, as we already voluntarily delegate freedom to collective organizations and standards in the sense of a 'consent of the governed'. Parallels can be drawn between the aforementioned autonomous driving example and the notion of delegating personal liberties to social organizations, a last will and testament, or considerations entailing collateral damage in armed conflicts. In this sense, the example also yields a conclusion related to the general relationship between human agency and AI. The transfer of the decision to the autonomous vehicle in the concrete accident situation does not represent a conflict with the human agency, as long as it is a recourse to general principles legitimated by democratic consent and is based on rational and abstract rules, which have been established in an open discourse.

V. Beneficence and Human Rights

In this section, we examine the direction in which the development and research of AI by companies should go from a human rights-based perspective. We intend to link the existing debate on the beneficence principle in AI regulations to the socio-economic

[38] Floridi et al, note 4.

https://doi.org/10.1017/bhj.2019.28 Published online by Cambridge University Press

implications of human rights. Although human rights are classically defined as 'claim rights' *vis-à-vis* the state, we intend to shed light on the broader implications of human rights on the realization of human rights standards, as well as their implications on socio-economic factors that might be linked to human rights.[39] The idea that using scientific technologies should serve the greater good stems from the beneficence criterion of AI, which is present throughout different regulatory frameworks and ethical guidelines. In this regard, the Beijing criteria on Artificial Intelligence are quite representative for the opinion that 'AI should be designed and developed to promote the progress of society and human civilization […]'.[40]

The idea of beneficence that economic and scientific development should contribute to normative goals fits into the framework of the United Nations Sustainable Development Goals, which encompass a wide range of social and environmental goals such as 'eradicating poverty in all its forms and dimensions'. Using the SDGs as a proxy for human rights might be somewhat problematic, as there is an ongoing debate currently about the exact relation between the two frameworks.[41] Nevertheless, we argue that some of these goals correspond to the socio-economic rights enumerated in the International Covenant on Economic, Social and Cultural Rights such as the 'right to an adequate standard of living', the 'right to education' or the 'right to health'.[42] The 2030 Agenda for Sustainable Development has pointed out that the United Nations SDGs 'seek to realize the human rights of all' and referred to the socio-economic rights of the United Nations Declaration on Human Rights.[43] In the following section, we illustrate therefore how companies might link their AI strategies to the realization of these socially important issues and how to integrate the realization of human rights enshrined in international treatises in corporate approaches related to AI.

A. Sustainable Development Goal 1: No Poverty

According to a recent publication, AI enables technologies that trigger economic growth and raise the productivity of the economy.[44] In this regard, AI appears to have direct positive consequences on poverty and global prosperity, which are, according to the

[39] Compare John Rawls, 'A Theory of Justice' (Cambridge, MA, USA: Belknap Press of Harvard University Press, 1971); The World Bank, 'Human Rights and Economics: Tensions and Positive Relationships', http://siteresources.worldbank.org/PROJECTS/Resources/40940-1331068268558/Report_Development_Fragility_Human_Rights.pdf (accessed 27 November 2019).

[40] Beijing Academy of Artificial Intelligence (BAAI), 'Beijing AI Principles' (28 May 2019), https://www.baai.ac.cn/blog/beijing-ai-principles (accessed 27 November 2019).

[41] The debate is also linked to the general role of human rights and corresponding economic obligations. Some posit that human rights unfold economic obligations (e.g., Human Rights and the SDG – Pursuing Synergies, https://www.universal-rights.org/wp-content/uploads/2017/12/RAPPORT_2017_HUMAN-RIGHTS-SDGS-PURSUING-SYNERGIES_03_12_2017_digital_use-2.pdf (accessed 12 November 2017); others are claiming that human rights do not entail an economic dimension or to a lesser extent (Robert Nozick: 'Anarchy, State, and Utopia').

[42] Office of the United Nations High Commissioner for Human Rights, The Right to Health, https://www.ohchr.org/Documents/Publications/Factsheet31.pdf (accessed 15 November 2019).

[43] A/RES/70/1.

[44] Ricardo Vinuesa et al, 'The Role of Artificial Intelligence in Achieving the Sustainable Development Goals', *arXiv* (2019), https://arxiv.org/ftp/arxiv/papers/1905/1905.00501.pdf (accessed 12 September 2019); Daron Acemoglu and Pascual Restrepo, 'Artificial Intelligence, Automation and Work', NBER Working Paper No. 24196 (2018).

https://doi.org/10.1017/bhj.2019.28 Published online by Cambridge University Press

United Nations, the world's greatest challenges against humanity.[45] Besides the general implications of AI on economic growth, many of the positive implications of AI are more direct. The use of drones in agriculture, for example, helps farmers work, produce and maintain their farms and livestock efficiently. The Stanford Poverty & Technology Lab has been doing intensive research on finding solutions for poor farmers, in agriculture, as well as in every other topic lifting humans up out of poverty. The Lab also uses AI and imagery to predict poverty, and these predictions have been borne out as 81–99% accurate.[46]

B. Sustainable Development Goal 5: Gender Equality

The use of AI might also be conducive to achieve gender equality and the empowerment of all women and girls.[47] In Pakistan, the AI chatbot 'RAAJI' talks to women about female reproductive health, hygiene and safety. Education and gender equality are not only a human right and a target of the SDGs, but also the main tool for enhancing other human rights-relevant topics, such as equality, personal freedom or human dignity. The chatbot enterprise is partnered with UNESCO to create content shown in the rural areas of Pakistan.

Although the two depicted cases only represent just a few cases for beneficent AI, they underscore that companies can contribute to the greater good by implementing AI solutions. This links up to the normative goal of beneficence in the sense of contributing to the economic, social and environmental goals. In the long-term perspective, the beneficence criterion deriving from AI governance might influence the general understanding of corporate human rights responsibilities in the sense that companies are increasingly committed to ethical improvements. This might be comparable to the implications of States, which have to raise the development level of the respective country. So far, the UN Guidelines on Business and Human Rights regard the responsibilities of companies mainly from their dimension of claim rights.[48] Hence, the embedding of the beneficence criterion in AI ethics might contribute to an interpretation of human rights as normative objectives for corporate decision-making.

VI. Contingent Violations of Human Rights by Artificial Intelligence

In the following section, we discuss cases in which the use of AI may conflict with human rights. The cases depicted here are representative and do not represent an exhaustive list.

---

[45] 'Task of Eradicating Poverty Must be Met "With a Sense of Urgency", says Deputy UN Chief', *UN News Centre* (8 May 2017), https://www.un.org/sustainabledevelopment/blog/2017/05/task-of-eradicating-poverty-must-be-met-with-a-sense-of-urgency-says-deputy-un-chief/ (accessed 2 September 2019); 2030 Agenda for Sustainable Development, https://www.un.org/ga/search/view_doc.asp?symbol=A/RES/70/1&Lang=E (accessed 12 November 2019); The World Bank, note 41.

[46] Joseph Benninton-Castro, 'AI is a Game-Changer in the Fight against Hunger and Poverty. Here's Why', *NBC News* (12 June 2017), https://www.nbcnews.com/mach/tech/ai-game-changer-fight-against-hunger-poverty-here-s-why-ncna774696 (accessed 27 November 2019).

[47] Compare A/RES/70/1.

[48] The Ruggie Principles speak explicitly of 'protect, respect and remedy', which largely confirms the defence or claim right character of human rights.

https://doi.org/10.1017/bhj.2019.28 Published online by Cambridge University Press

Table 1. Typology of AI-Related Human Rights Violations

Situation	Exemplary cases
Situation I: the input of AI conflicts with human rights	• Use of data without or against the explicit will of customers • Disproportionate use of intimate and personal data of individuals by public institutions
Situation II: the output of AI leads to unintended human rights violations	• Unlawful discrimination in job applications based on ethnicity • Illicit discrimination of women in the public health system
Situation III: the use of AI in specific areas conflicts with human rights	• Infringement of the right to opinion, due to excessive use of algorithms in social media • Replacement of democratic decisions by AI decisions (robotocracy)
Situation IV: a human rights violator uses AI	• Use of AI to monitor the citizens criticizing the government • Use of AI to supress ethnic minorities and to track individuals

In contrast with the section above, we discuss here not what AI should do, but rather the should-nots of AI in terms of human rights. To facilitate an understanding of the individual aspects of what it means to conform to human rights, we have oriented ourselves to various situations, addressing AI-specific problem areas. Central aspects are the human rights conformity of the in- and output of AI solutions, the type of use and the intentions of the actor (Table 1).

A. Situation I: The Input of AI Conflicts with Human Rights

AI needs to process data in order to expand its capabilities and to perform certain tasks. Understanding this close linkage between AI and data is crucial for practical cases, as the data collection may conflict with individuals' right to privacy and with their data autonomy.[49] From a human rights perspective, the right to privacy could be regarded as an extension of human dignity, which has been confirmed by court rulings (cf. *Lawrence v Texas*) and legal texts such as the Universal Declaration of Human Rights and the European Convention on Human Rights (Article 8).[50]

Unlike human dignity, the ownership of data can usually be transferred to third parties by consent, which also applies to how the data are used.[51] The General Data Protection Regulation (GPDR) reads, 'where processing is based on the data subject's consent, the controller should be able to demonstrate that the data subject has given consent to the

[49] German Data Ethics Commission, 'Opinion of the Data Ethics Commission' (October 2019), https://www.bmjv.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_EN.pdf?__blob=publicationFile&v=2 (accessed 27 November 2019).

[50] Amitai Etzioni, 'Are New Technologies the Enemy of Privacy?' (2007) 20 *Knowledge and Policy* 115.

[51] Lawrence Lessig, 'Privacy as Property' (2002) 69 *Social Research* 1, 247.

https://doi.org/10.1017/bhj.2019.28 Published online by Cambridge University Press

processing operation'. This passage mainly reflects the increased use of data by AI, such as in highly automated and especially connected driving: interactions between vehicles would substantially enhance road safety and would – at least, in principle – probably meet the consent of all involved parties. While data protection in the case of connected driving serves to enhance vehicle safety, commercial interests might dominate data accumulation in other areas. This is not problematic *per se*, as long as the transaction is based on the consent of both parties. Nevertheless, the factual realization of the consent principle might hinge on the socio-economic environment, which needs to be addressed on the regulatory level or self-legislated sector-wide standards.[52]

Of major importance in this context is the principle of informed consent, which has been integrated in many frameworks dealing with AI.[53] The notion of informed consent implies that 'measures should be taken to ensure that stakeholders of AI systems are with sufficient informed-consent about the impact of the system on their rights and interests'.[54] At the same time, too much data at the wrong place might lead to a detrimental outcome, unintended by corporate decision makers. The proportion of Jews killed in the Netherlands during the Holocaust was relatively high because the city council of Amsterdam had a detailed population census, which statistically recorded the religion of the inhabitants. By accessing these data, the German Gestapo was enabled to transfer Jewish citizens to concentration camps.[55] In order to mitigate future human rights violations, businesses should uphold the principle of data minimization, which consists of not collecting more personal information than needed for a particular purpose.[56] This might be integrated with the principle of foresighted responsibility[57] meaning that companies need to take network effects and changing actor constellations into account. The due diligence of a company poses therefore not only questions regarding the origin of data, but also to their end destination and future use cases.

The cooperation with public authorities, in the form of providing data or in the form of receiving data, represents maybe the most important challenge for companies in the sphere of data input-related human rights violations. The main reason is that the relationship between data receiver and data provider is asymmetric and that the right to privacy might be derogated in norm conflicts.[58] The usage of data by the police or investigative units, for example, requires balancing the right to privacy with the public's

[52] Pagallo et al, 'On Good AI Governance: 14 Priority Actions, a S.M.A.R.T. Model of Governance, and a Regulatory Toolbox' (2019), https://www.eismd.eu/pdf/AI4PEOPLE%20On%20Good%20Ai%20 Governance%202019.pdf. (accessed 27 November 2019).

[53] Compare Future of Life Institute: Asilomar AI Principles (2017), https://futureoflife.org/ai-principles/ (accessed 27 November 2019).

[54] Beijing Academy of Artificial Intelligence (BAAI), 'Beijing AI Principles' (28 May 2019), https://www.baai.ac.cn/blog/beijing-ai-principles (accessed 12 November 2019).

[55] William Steltzer, 'Population Statistics, the Holocaust, and the Nuremberg Trials' (1998) 24 *Population and Development Review* 3, 511–552.

[56] Compare with General Data Protection Regulation Article 5.

[57] German Data Ethics Commission, 'Opinion of the Data Ethics Commission' (October 2019), https://www.bmjv.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_EN.pdf?__blob=publicationFile&v=2 (accessed 22 November 2019).

[58] Antonio T Reigada, 'The Principle of Proportionality and the Fundamental Right to Personal Data Protection: The Biometric Data Processing' (2012) 17 *Lex Electronica* 2.

https://doi.org/10.1017/bhj.2019.28 Published online by Cambridge University Press

general interest in investigating criminal or administrative offences. However, not all ends justify specific means, as the usage of such data inputs for 'minor' crimes such as drug consumption, tax evasion or undeclared work would render the notion of data privacy as a defence right *vis-à-vis* the state as obsolete. China's usage of surveillance technologies in Xinjiang province points to the misuse of data by state authorities.[59] The orientation towards the principle of proportionality and the strict application of necessity can therefore provide an important baseline for guaranteeing the lawful conduct of AI, leading us to the following broad rules for regulating and self-regulating the data input into AI:

- The data transfer by enterprises has to be in line with the involved parties' consent and needs to consider changing actor constellations.
- Data can only be used by AI solutions with the consent of the involved parties, to reduce harm to others (application of the harm principle).
- The invasiveness of the AI solution needs to be proportional to its aims. The access to data input needs to be the least invasive incursion possible.

The perspective matters for companies, as they might interact with public entities in critical infrastructures such as surveillance, profiling or face recognition. The application iBorderCTRL, checking the credibility of flight passengers in European airports, might be an example of the cooperation between public entities and AI developing enterprises involving critical personal data. Due to their closeness to human rights sensitive topics, companies operating in such environments need to control for human rights violations and develop their own codes of conduct in terms of data use. For assessing the likelihood of conflicts with human rights, parallels could be drawn to other advanced interrogation tools such as DNA profiling,[60] the establishment of DNA databases[61] or the use of polygraphs by law enforcement, which are partly perceived as excessive and disproportionate breaches of privacy if used against or without consent or knowledge of the suspect.[62]

The application of AI in judicial systems is highly critical as well. A limit of interrogation tools would be the legal principle of *nemo tenetur se ipsum accusare*, according to which nobody can be forced to accuse himself or herself. The legal status of the so-called right to silence has been enshrined in the European Convention on Human Rights (Article 6) and as relevant in the criminal procedural law of many countries. According to some jurisdictions,[63] the usage of polygraphs would even qualify as an

---

[59] Uyghur Human Rights Project, 'China's Repression and Internment of Uyghurs: U.S. Policy Responses', *House Committee on Foreign Affairs: Subcommittee on Asia and the Pacific* (26 September 2018), https://docs.house.gov/meetings/FA/FA05/20180926/108718/HHRG-115-FA05-Wstate-TurkelN-20180926.pdf (accessed 27 November 2019).

[60] Tania Simoncelli and Helen Wallace, 'Expanding Databases, Declining Liberties' (2006) 19 *Genewatch: A Bulletin of the Committee for Responsible Genetics* 1, 3.

[61] Helen Wallace, 'The UK National DNA Database: Balancing Crime Detection, Human Rights and Privacy' (2006) 7 *EMBO Reports* 26; HM Wallace et al, 'Forensic DNA Databases – Ethical and Legal Standards: A Global Review' (2014) 4 *Egyptian Journal of Forensic Science* 3, 57.

[62] Ed Johnston, 'Brain Scanning and Lie Detectors: The Implications for Fundamental Defence Rights' (2016) 22 *European Journal of Current Legal Issues* 2.

[63] Federal Court of Justice of Germany, *Dispensation of Justice 1954 BGH*, 16.02.1954-1 StR 578/53.

https://doi.org/10.1017/bhj.2019.28 Published online by Cambridge University Press

absolute violation of human rights, indicating that using AI technologies to detect a person's trustworthiness,[64] by using very personal data, faces high barriers. The comparison between the effects of AI and more conventional techniques might be relevant here, as AI does not constitute an entire novelty here, in terms of invasiveness, and as similar comparisons between conventional and modern technologies have already been made in the context of regulating cyber war. In such situations, companies might not only refer to national legislation, but are bound by internationally recognized human rights, due to the criticality of the data input used.

B. Situation II: The Output of AI Leads to Unintended Human Rights Violations

In this section, we refer to human rights abuses originating from misalignment between the goals and the machine's implementation.[65] As with every technology, unintended consequences to a specific technology can have devastating results. The example of Microsoft's chat robot Tay exposes the potential consequences of design faults. Originally designed to mimic the language patterns of a 19-year-old American girl, the chat robot ended up praising Hitler and inciting hatred.[66] Moreover, examples of errors resulting from incorrect programming or training of AI which affect human rights range from security risks in autonomous vehicles to discrimination issues in job application software. Most of these cases might constitute not intentional violations of fundamental rights but rather technical faults or acts of negligence. In this sense, biases might constitute a new form of human rights violations, where the perpetrator has no interest in violating human rights. The impacts of these faults, biases and errors on human rights, however, should not be under-estimated. For instance, Angwin et al. found that AI solutions operated by the police were discriminating against black people, while Amazon withdrew an AI solution, which was biased against women in applications for technical jobs.[67] Obermeyer et al. illustrated that the US health care system relied on an algorithm to guide health decisions, which was affected by a bias, leading to a discrimination against black Americans.[68]

Biases resulting in illicit discrimination of individuals are the most representative examples of unintended human rights violations. The reasons behind discrimination accidentally produced by AI solutions differ, but a major source for faults and human rights violations of this kind is that AI is often unable to separate causation from correlation. Moreover, problems concerning discrimination by AI often relate to how the 'target variable' and the 'class labels' are calibrated, how the training data are labelled

---

64 Federal Court of Justice of Germany, *Dispensation of Justice 2003 BGH*, v. 24.06.2003-VI ZR 327/02.

65 Floridi et al, note 4.

66 Elle Hunt, 'Tay, Microsoft's AI Chatbot, Gets a Crash Course in Racism from Twitter', *The Guardian* (24 March 2016), https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter (accessed 27 November 2019).

67 Julia Angwin et al, 'Machine Bias', *ProPublica* (23 May 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (accessed 27 November 2019).

68 Ziad Obermeyer et al, 'Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations' 366 *Science*, 447–453.

https://doi.org/10.1017/bhj.2019.28 Published online by Cambridge University Press

and collected, and the feature selection and proxies. Different from other cases, where AI and human rights conflict, we have the advantage that companies are already falling under the radar of the law, as most countries forbid discrimination based on gender, religion and other factors. Legally, anti-discriminatory frameworks also consider cases of unintended discrimination, so that most laws would also apply to the use of AI. The use of AI where some forms of discrimination are not explicitly forbidden might require legal changes and voluntary codes of conduct tailored to the use of AI.[69] In fact, some codes and guidelines in the field of artificial intelligence, such as the Asilomar Principles, have already brought the aspect of non-discrimination to the fore. As artificial intelligence can have a great impact on life and property, processes that can prevent a bias and identify ethical problems early play an essential role. The mitigation of biases qualifies therefore to the risk 'of causing or contributing to gross human rights abuses as a legal compliance'.[70]

The implication for companies is that in the case of AI product quality, consumer affairs and human rights need to be seen from an interconnected point of view, implying that the pure likelihood that AI decisions are better than human decisions is therefore not enough for implementing human rights-compliant AI. From a regulatory perspective, solutions addressing the enforcement of hygiene standards need to be compatible with the prevailing incentive structure, as moral appeals are not always strong enough to deter companies from violating norms. The real problem behind the use of AI, therefore, may be rather the lack of quality standards or negligence on the producer's behalf. From the point of view of business ethics, the relevant precedent would be the Ford Pinto Case from the 1970s. In this particular case, Ford prioritized profit maximization over product safety and evaded additional investment in safety, while accepting a higher mortality rate.[71]

In addition to quality standards pertaining to human rights impacts and anti-discrimination frameworks, transparency plays another important role to prevent that the output leads to unintended human rights violations. The Council of Europe has therefore stressed the importance of 'transparent human rights due diligence processes that involve the identification of the human rights risks associated with their AI systems, and taking effective action to prevent and/or mitigate the harms posed by such systems'.[72] This rationale extends to the actions of private companies, which are dealing with services that are of material interest for the involved parties, for example, the health sector. Mechanisms for providing more transparency might include sustainability reports, which are already dealing with human rights impacts. For example, companies might report the criteria used in the algorithms of job application mechanisms and describe the processes used to reduce bias risks.

Transparency of the management approach to AI, explanations of AI solutions and their potential impacts on human rights as well as remedy mechanisms would play central

---

[69] Frederik Z Borgesius, 'Discrimination, Artificial Intelligence, and Algorithmic Decision-Making', Council of Europe: Strasbourg (2018), https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73 (accessed 27 November 2019).

[70] Compare UN Guiding Principles on Business and Human Rights, note 23.

[71] Compare Lee P Strobel, *Reckless Homicide? Ford's Pinto Trial* (South Bend and Books, 1980).

[72] Council of Europe, 'Unboxing Artificial Intelligence: 10 steps to protect Human Rights', https://rm.coe.int/unboxing-artificial-intelligence-10-steps-to-protect-human-rights-reco/1680946e64 p. 18 (accessed 27 November 2019).

https://doi.org/10.1017/bhj.2019.28 Published online by Cambridge University Press

roles to ensure that AI conforms to human rights. The gravity of the mechanisms enforcing compliance with these principles, in the forms of licensing and auditing, would depend on the AI solution's impact and materiality to human rights.

C. Situation III: The Use of AI in Specific Areas Conflicts with Human Rights

In the previous sections, we have discussed the impacts of AI on human rights from a 'claim rights' perspective. However, a holistic perspective on the issue requires the inclusion to rights that enable participation in the political and social decision-making process. Citizen political participation classically expresses itself in the right to vote and the right to express one's opinion. Both concepts might be linked up to the notion of 'democratic participation', which has been implemented as a guiding of AI regulation by the Montréal Declaration for Responsible Development of Artificial Intelligence,[73] and to the 'principle of autonomy' in the Asilomar AI Principles. In order to illustrate the importance of specific areas for human rights violations and particularly in the form of participation rights, we concentrate on the handling of companies of AI in terms of freedom of speech, as it belongs to the most contested area in AI ethics. Its relevance has been stressed by the UN Report 'Promotion and protection of the right to freedom of opinion and expression', which elaborated on the impact of AI to freedom of expression and opinion and stressed the internet's role as a platform for forming and articulating opinions.[74]

In general, AI's influence on the freedom of opinion and expression seems to have two implications for corporate decision makers. On the one hand, biases created by AI might impact individuals' self-determination and autonomy to form and develop personal opinions based on factual and varied information.[75] AI's impacts might be particularly strong, if we consider the important role that discourse theories attach to the presupposition that no relevant arguments are suppressed or excluded by the participants.[76] As a result, filtering by AI and could change the discourse's direction and suppress parts of the opinion spectrum. The Montreal declaration has therefore argued that safeguarding 'democracy against the manipulation of information for political ends' constitutes one of the major ethical challenges for decision makers, implying that companies need to 'prevent and mitigate' negative effects on the discourse by the use of risk assessments and quality standards.[77]

---

[73] Artificial Intelligence Cluster Steering Committee Quebec, 'Montréal Declaration for Responsible Development of Artificial Intelligence', https://ai.quebec/wp-content/uploads/sites/2/2018/12/News-release_Launch_Montreal_Declaration_AI-04_12_18.pdf (accessed 27 November 2019).

[74] David Kaye, 'Mandate of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression', Open Letter to Office of the High Commissioner for Human Rights (1 June 2017), https://www.ohchr.org/Documents/Issues/Opinion/Legislation/OL-DEU-1-2017.pdf (accessed 27 November 2019).

[75] Jack M Balkin, 'Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation', Yale Law School, Public Law Research Paper No. 615 (2018); Maja Brkan, 'Freedom of Expression and Artificial Intelligence: On Personalisation, Disinformation and (Lack of) Horizontal Effect of the Charter (17 March 2019), http://dx.doi.org/10.2139/ssrn.3354180 (accessed 27 November 2019).

[76] Juergen Habermas, *The Theory of Communicative Action* (Boston, MA, USA: Beacon Press, 1992).

[77] Compare UN Guiding Principles, note 23.

https://doi.org/10.1017/bhj.2019.28 Published online by Cambridge University Press

On the other hand, the use of AI to censor specific political comments remains highly problematic. The current debate centres on the question of whether AI should limit hate speech.[78] The first issue concerns the aspect of technical feasibility. AI does not yet seem to be able to distinguish between appropriate commentaries and hate speech. The lines between 'still allowed' expressions of opinion and hate comments might be blurring and situation dependent: AI solutions require an understanding of irony and of milieu- and culture-specific expressions, making the standardization of individual case decisions a very tedious task. The second issue is of a more theoretical nature. The transfer of the censoring of comments to companies might be questionable not only in terms of freedom of opinion but also when considering the rule of law principle, as formerly judicial decisions are replaced by algorithms used by private enterprises.[79] Under these circumstances, it is therefore indispensable for companies and especially for social networks to anchor remedy mechanisms enabling individuals to protect themselves in cases of injustice. Moreover, the UN Report on the 'Promotion and protection of the right to freedom of opinion and expression' has referred to the responsibility of companies to 'publish data on content removals, […] alongside case studies and education on commercial and political profiling'.[80] Due to the fact that the 'overuse and underuse'[81] of censoring comments in social media might be detrimental to human rights to the same extent, companies might be required to emphasize these procedural aspects like transparency to third parties, risk assessments and operational standards to a greater extent than in other cases of AI.

D. Situation IV: A Human Rights Violator Uses AI

In the final section, we deal with the most problematic use case, namely the use of AI to infringe upon human rights. This form of AI use typically called 'malicious AI' and AI regulation has dedicated strong attention to this aspect, which might be even regarded as the oldest principle of AI regulation.[82] The difference from the aforementioned cases lies in the intention of using AI as an instrument to violate human rights. In this sense, the human rights violation is not an externality or an act of negligence, but rather the goal of the AI solution.

Possible scenarios of this sort would include using algorithms to discriminate against trade union members in automated remuneration and promotion processes, or programmes that – deliberately – favour certain ethnic groups in the distribution of social services. Both cases constitute acts of discrimination and depict severe

---

[78] 'Germany Starts Enforcing Hate Speech Law', *BBC* (1 January 2018), https://www.bbc.com/news/technology-42510868 (accessed 30 August 2019).

[79] Frank La Rue, 'Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression', United Nations General Assembly Human Rights Council (16 May 2011), https://www2.ohchr.org/english/bodies/hrcouncil/docs/17session/A.HRC.17.27_en.pdf (accessed 12 September 2019).

[80] David Kaye, 'Mandate of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression', Open Letter to Office of the High Commissioner for Human Rights (1 June 2017), https://www.ohchr.org/Documents/Issues/Opinion/Legislation/OL-DEU-1-2017.pdf (accessed 27 November 2019).

[81] Floridi et al, note 4.

[82] Compare Asimov's Law: 'A robot may not injure a human being or, through inaction, allow a human being to come to harm'.

https://doi.org/10.1017/bhj.2019.28 Published online by Cambridge University Press

violations of the right to equal and fair treatment. While the legislative power can establish legal provisions and screening mechanisms for preventing human rights violations under its jurisdiction, human rights violations intended outside of its scope are far more difficult to prevent.

Recent developments in some countries give cause for concerns that the combination of AI with big data might strengthen the surveillance mechanisms of 'rogue states' or terrorist organizations.[83] One example that underscores the practical relevance of non-maleficence is the Chinese government's crackdown on ethnic minorities. Recently, government surveillance has expanded and has been combined with the use of AI to access biodata and DNA databases, and is linked for facial recognition.[84] The Chinese government's measures also raise questions concerning legitimacy, as some of these applications seem to be used in connection with crackdowns on parts of the Chinese opposition or on national minorities.[85] According to a report by the *New York Times* in 2019,[86] Chinese start-ups have built algorithms to monitor ethnic Muslims in China's Xinjiang province, which has sparked criticism from several NGOs as well as governmental[87] and international organizations.[88] The existence of camera systems aimed at tightening control over citizens, combined with detainment camps and face-recognition tools, intensify the matter's urgency.[89] From a human rights perspective, the events taking place in Xinjiang constitute a violation of human rights as the state's interventions conflict with proportionality considerations and the legitimate use of state power. The fact that the members of ethnic minorities are generally placed under suspicion, the violations of the right to physical integrity and equal treatment (such as article 18/19/21/29 of the Counterterrorism Law of the PRC, or the 'Regulation on De-Extremification') and the low legal threshold for detaining individuals seem to confirm this view.[90]

Such cases naturally entail governance implications for companies supplying technologies to repressive regimes, because AI qualifies for dual-use technology.

---

[83] Ibid.

[84] Uyghur Human Rights Project, 'China's Repression and Internment of Uyghurs: U.S. Policy Responses', *House Committee on Foreign Affairs: Subcommittee on Asia and the Pacific* (26 September 2018), https://docs.house.gov/meetings/FA/FA05/20180926/108718/HHRG-115-FA05-Wstate-TurkelN-20180926.pdf (accessed 27 November 2019).

[85] Lily Kuo, '"If You Enter a Camp, You Never Come Out": Inside China's War on Islam', *The Guardian* (11 January 2019), https://www.theguardian.com/world/2019/jan/11/if-you-enter-a-camp-you-never-come-out-inside-chinas-war-on-islam (accessed 27 November 2019).

[86] Paul Mozur, 'One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority', *New York Times* (14 April 2019), https://www.nytimes.com/2019/04/14/technology/china-surveillance-artificial-intelligence-racial-profiling.html (accessed 27 November 2019).

[87] Federica Mogherini, 'Speech by HR/VP Mogherini at the Plenary Session of the European Parliament on the State of the EU–China Relations' (11 September 2018), https://eeas.europa.eu/headquarters/headquarters-homepage/50337/speech-hrvp-mogherini-plenary-session-european-parliament-state-eu-china-relations_en (accessed 26 November 2019). See German Parliament Document 19/5544 (November 2018).

[88] United Nations Human Rights Office of the High Commissioner, 'Committee on the Elimination of Racial Discrimination reviews the report of China' (13 August 2018), https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=23452&LangID=E (accessed 27 November 2019).

[89] Adrian Zenz, 'Thoroughly Reforming Them Towards a Healthy Heart Attitude: China's Political Re-education Campaign in Xinjiang' (2019) 38 *Central Asian Survey* 1, 102.

[90] Rian Thum, 'China's Mass Internment Camps Have No Clear End in Sight', *Foreign Policy* (22 August 2018), https://foreignpolicy.com/2018/08/22/chinas-mass-internment-camps-have-no-clear-end-insight/ (accessed 27 November 2019).

https://doi.org/10.1017/bhj.2019.28 Published online by Cambridge University Press

The dual-use character applies explicitly for facial- and voice-recognition software, the collection of biodata and predictive policing databases. According to the UNGPs, companies need therefore to consider how these technologies will be used by the end user, and whether they need to establish country-specific due diligence measures to prevent cases of misuse. This matters also from a compliance perspective, as human rights violations connected in the broader supply chain of enterprises could already fall into the scope of extraterritorial legislation including the UK Modern Slavery Act or the US Global Magnitzky Act.

Nevertheless, the dual-use character of AI might be helpful for companies to detect human rights violations. Modern technologies have decisively supported the work of journalists, activists and scholars in their research on the ongoing events in Xinjiang and elsewhere.[91] A concrete case in which AI has had a positive impact on human rights includes the use of blockchain in the supply chain of cobalt, to help companies and governments evaluate whether the material has been involved in human rights issues.[92] Reports have also underscored AI's role in detecting and fighting financial crime and money laundering as well as in checking sanctions.[93]

Although the use of algorithms and of blockchain technologies might be central to tracing human rights violations in the future and increase the pressure on companies to comply with human rights standards when operating abroad, using these technologies might reach certain factual limitations. Countries with larger economic size and stronger political weight are more likely to resist sanction mechanisms imposed on them and have more capabilities to circumvent individually imposed sanctions. Effectively preventing these actions is therefore not only a matter of AI-specific legislation but also a question of foreign policy and enforcement of international human rights norms.

VII. Conclusion

In this article, we examined the responsibilities of corporate actors to the enforcement and realization of human rights standards in the context of artificial intelligence. In general, we found that the use of AI – even in life-and-death decisions – does not principally conflict with the principle of moral self-determination. Nevertheless, the handling of AI needs to be based on broader human rights considerations. The general application of AI in irreversible and dilemmatic situations, for example, needs to comply with universal and abstract rules, which might evolve from an open and democratic discourse.

In a further step, we delineated challenges for companies concerning human rights conduct posed by artificial intelligence. According to the beneficence criterion of AI ethics, AI has the aim to benefit human beings and the common good. This idea might be

[91] Ibid; Zenz, note 88.

[92] Compare Adam J. Sulkowski, 'Blockchain, Business Supply Chains, Sustainability, and Law: The Future of Governance, Legal Frameworks, and Lawyers?' (2019) 43 *Delaware Journal of Corporate Law* 2, 303–345.

[93] Ellen Zimiles and Tim Mueller, 'How AI Is Transforming the Fight against Money Laundering', *World Economic Forum* (17 January 2019), https://www.weforum.org/agenda/2019/01/how-ai-can-knock-the-starch-out-of-money-laundering/ (accessed 27 November 2019).

https://doi.org/10.1017/bhj.2019.28 Published online by Cambridge University Press

linked to the progressive realization of human rights according to the International Covenant on Economic Social and Cultural Rights as well as the framework of UN Social Development Goals. We have found many cases, which show the use of AI can be conducive to the realization of human rights, particularly when it comes to health, combating poverty and education, and that companies can play a positive role in offering AI solutions that address human rights relevant issues.

In other contexts, where the actions of companies might conflict with human rights, the handling of AI use cases necessitates the establishment of governance mechanisms preventing human rights violations. In the case of data input, the application of AI is generally unproblematic, as long as all involved parties consent to the conditions of its use and as long as the use complies with legislative and ethical standards. The consent principle, however, might be challenged given the power asymmetry between company and the consumer. This applies specifically for the interaction between enterprises and public authorities. Another area of potential human rights violations are biases leading to illicit discrimination. Here companies need to adhere to quality standards and to transparently illustrate corporate decision making and the risks of human rights violations, which result from the use of AI. Moreover, the use of AI is questionable in some cases and environments, particularly when it comes to citizen participation or the far-reaching transfer of decision-making powers. In order to prevent infringements of the right to opinion, companies need to develop preventive measures and to build a company internal infrastructure which aims to remedy human rights violations. Finally, the last issue related to AI and human rights relates to actors who use AI to violate human rights. In contrast to previous enforcement mechanisms, AI's advantages are that the results are more reliable and yield faster results. However, the enforcement of human rights against state actors poses major challenges to the internationally acting enterprises, as international law is increasingly emphasizing companies' responsibilities.

When assessing these aspects of the topic of AI and human rights, we discover that corporate decision makers are confronted with complex challenges ranging from transparency, data quality issues and supply chain management. The input and output of AI, the users and how the technologies are used are completely different from the location of the regulatory possibilities and therefore require measures appropriate to the situation.