RUTGERS Law School
ILHR
RUTGERS INTERNATIONAL LAW AND HUMAN RIGHTS JOURNAL

ARTIFICIAL INTELLIGENCE AND HUMAN RIGHTS: FOUR REALMS OF DISCUSSION, RESEARCH AND ANNOTATED BIBLIOGRAPHY[*]

Jootaek Lee[**]

ABSTRACT

Artificial Intelligence ("AI") should be considered an active actor now, rather than later, because it would be too late to control after AI development passes the Singularity. Before any irreversible damages occur relating to Artificial General Intelligence ("AGI"), states should take measures to prevent any harmful effects to society. How can humans be protected in such a situation? This article inductively analyzes the human rights implications of AI/ AGI in four different realms, devised according to the level of AI development toward AGI as well as the human rights implications depending on the level of similarity of AI to humans.

TABLE OF CONTENTS

[*] This paper is part of the presentation at the 2020 ASIL Virtual Annual Meeting (June 26, 2020).

[**] © Jootaek Lee is an Assistant Professor and Law Librarian at Rutgers Law School. The author would like to thank Audrey Han, J.D. for her research assistance, and James Britt, J.D. for his editing.

INTRODUCTION

The term, "artificial intelligence" ("AI") has changed since it was first coined by John McCarthy in 1956. AI, believed to have been created with Kurt Gödel's unprovable computational statements in 1931,[1] is now called deep learning or machine learning. AI is defined as a computer machine with the ability to make predictions about the future and solve complex tasks using algorithms.[2] The AI algorithms are enhanced and become effective with big data capturing the present and the past, while still necessarily reflecting human biases into models and equations.[3] AI is also capable of making choices like humans, mirroring human reasoning.[4] AI can help robots to efficiently repeat the same labor-intensive procedures in factories. It can also analyze and present data more efficiently through deep learning, natural language processing, and anomaly detection. Thus, AI covers a spectrum of augmented intelligence relating to prediction, autonomous intelligence relating to decision-making, automated intelligence for labor robots, and assisted intelligence for data analysis.[5]

This spectrum, however, will be further expanded with the development of the Artificial General Intelligence ("AGI"), also known as super-intelligence. The AGI, a set of algorithms learning and developing multiple self-intelligences independently to resolve multiple problem domains, will accelerate the displacement of human labor. Just as Jeremy Rifkin's *The End of Work* foresees the

[1] *Id.* at 29-30.

[2] Mathias Risse, *Human Rights and Artificial Intelligence: An Urgently Needed Agenda*, 41 HUMAN RIGHTS QUARTERLY 2 (2019).

[3] *Id.* (citing Julia Angwin et. al., *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing; Reuben Binns, *Fairness in Machine Learning: Lessons from Political Philosophy*, 81 J. OF MACH. LEARNING RESEARCH 1 (2018); Brent Daniel Mittelstadt et al., *The Ethics of Algorithms: Mapping the Debate*, 3(2) BIG DATA & SOC'Y 3 (2016); OSONDE A. OSOBA & WILLIAM WELSER IV, AN INTELLIGENCE IN OUR IMAGE: THE RISKS OF BIAS AND ERRORS IN ARTIFICIAL INTELLIGENCE (RAND Corporation 2017)).

[4] Eileen Donahoe & Megan MacDuffee Metzger, *Artificial Intelligence and Human Rights*, 30 J. DEMOCRACY 115, 115 (2019).

[5] WORLD ECON. FORUM, HARNESSING ARTIFICIAL INTELLIGENCE FOR EARTH 7 (Jan. 2018), http://www3.weforum.org/docs/Harnessing_Artificial_Intelligence_for_the_Earth_report_2018.pdf [hereinafter WEF AI].

end of farmers, blue-collar workers, and service workers due to the First-, Second-, and Third-Industrial Revolutions,[6] after the Fourth Industrial Revolution with AI, robots, biotechnology, nanotechnology, and autonomous vehicles, the AGI can be developed to displace current, general human jobs. This issue has been a big conundrum to answer in contemporary society.

This negative denotation of AI displacement of human labor during the Fourth Industrial Revolution will depend upon how we define the nature of the mind because there is more to the mind than the brain.[7] Based on the Gardner theory,[8] there are intelligences highly relying on the mind such as musical intelligence, interpersonal intelligence, intrapersonal intelligence—metacognitive skill, and naturalistic intelligence.[9] To the extent that our world needs and cares about only visual-spatial, linguistic-verbal, and logical-mathematical intelligences in the brain, AGI so-called superhuman intelligence[10] may cause the human era to come to an end.[11] Furthermore, some scholars suggest that even curiosity and creativity, which are relevant to musical and intrapersonal intelligence, can be defined and interpreted in a new way of "connecting previously disconnected patterns in an initially surprising way," and thus can be reached and realized by AI.[12] At a certain point, "Singularity," defined as the acceleration of technological progress,[13] will cause exponential "runaway" reactions beyond any hope of control.[14]

[6] *See* JEREMY RIFKIN, THE END OF WORK: THE DECLINE OF GLOBAL LABOR FORCE AND THE DAWN OF THE POST-MARKET ERA 59-164 (1995).

[7] WEF AI, *supra* note 6, at 3.

[8] HOWARD GARDNER, FRAMES OF MIND: THE THEORY OF MULTIPLE INTELLIGENCES (2011). It is criticized to include too many aspects of human characters in the definition of intelligence; *See* Kendra Cherry, *Gardner's Theory of Multiple Intelligence*, VERYWELL MIND, (July 17, 2019), https://www.verywellmind.com/gardners-theory-of-multiple-intelligences-2795161.

[9] Naturalistic intelligence is related to the ability to recognize the environment and nature. This ecological receptiveness is closely related to the "sensitive, ethical, and holistic understanding" of the world and its complexities, including the role of humanity within the greater ecosphere. Marla Morris, The Eight One: Naturalistic Intelligence, *in Multiple Intelligences Reconsidered* (2004)

[10] *See* Nick Bostrom, *How Long Before Superintelligence?*, 2 INT'L J. FUTURE STUDS. (1998), *reprinted in* 5 LINGUISTIC AND PHILOSOPHICAL INVESTIGATIONS 11-30 (2006).

[11] Steven Livingston & Mathias Risse, *The Future Impact of Artificial Intelligence on Humans and Human Rights*, 33 ETHICS & INT'L. AFFAIRS 141, 141-58 (2019) (quoting the comment by Vernor Vinge at the 1993 VISION-21 Symposium) The algorithms of DeepMind Technologies, Google's DeepMind, and Google Brian are the best examples relating to AGI.

[12] Jürgen Schmidhuber, s*upra* note 1, at 36.

[13] It is also known as history's convergence or Omega point **Ω**. *Id.* at 39-40.

[14] *See* Vernor Vinge, *Technological Singularity*, VISION-21 (1993), https://mindstalk.net/vinge/vinge-sing.html; Jesse Parker, *Singularity: A Matter of Life and Death*, DISRUPTOR DAILY (Sept. 13, 2017), https://www.disruptordaily.com/singularity-matter-life-death/.

Singularity will surpass human intelligence and lead to irreversible changes to human civilizations, where most human affairs may not continue.[15]

When seen from a different angle, however, AI can be viewed more optimistically. Robots and machines equipped with AI can co-exist harmoniously, just as humans' relationship with animals. They may replace humans in many traditional jobs, but humans also may be able to find new jobs to supervise, educate, and design AI. When AI machines are given similar moral status as humans, humans will be able to feel empathy towards robots, and the necessity to protect them as a human equivalent will be possible.[16]

In accordance with Moore's Law,[17] AI has developed exponentially in the last ten years, and this noticeable development allows us to accurately predict both positive and negative results from AI. AGI seemed far from reality when AI was simply used to solve only computational problems and play games against humans. For example, the AI computer IBM Watson won first prize in *Jeopardy* in 2011, and the AI computer AlphaGo defeated professional Go player Lee Sedol, in 2016.

Nowadays, AI has become realized in more concrete forms such as androids, autonomous vehicles, autonomous killing machines, and various video avatars. These new forms of AI have affected human lives extensively, and as a result, impacted human rights and humanitarian law.[18] AI is not a story of the future anymore; it is the new normal. AI affects contemporary everyday life. AGI is also believed to be fully developed within the next 20 to 50 years. The current moment is a good time to thoroughly analyze the human rights implications of AI, both philosophically and empirically, before AI development passes the Singularity.

This paper will inductively investigate and analyze the human rights implications of AI and AGI. First, the article will divide and organize the AI discussion relating to human rights into four different realms: (1) The first realm: Human's application of AI and AI's positive contributions to human life; (2) the second: AI as human rights violators and humans' efforts for accountable AI; (3) the third: AI as objects of human rights protection; and (4) the fourth: AI seeks their own rights as active subjects. In those four realms, the article will introduce

[15] *Id.*

[16] Sophia, a humanoid robot is an example. *Sophia*, HANSON ROBOTICS, https://www.hansonrobotics.com/sophia/ (last visited Nov. 27, 2020).

[17] *See* Wikipedia, Moore's Law, https://en.wikipedia.org/wiki/Moore%27s_law. Moore's Law suggests that technology has been exponentially improving since 1971.

[18] The areas to regulate new human life include the human rights to development, climate change, life, health, education, criminal justice, equal protection, due process, work, and privacy.

relevant philosophies, theories, and legal systems that may apply to the current AI issues. Next, the article will review recommendations on the desirable future development of AI. Lastly, in order to help future research, the article will also review the current literature on AI and human rights and provide an annotated bibliography.

I. FOUR REALMS OF AI AND HUMAN RIGHTS

This section organizes the AI discussion into four different realms. The four realms are devised according to the level of AI development toward AGI, as well as the human rights implications depending on the level of similarity of AI to humans. These realms may overlap in stages of development and timeline.

The First Realm explores the benefits that AI has brought to human life. The Second Realm discusses the negative side effects that result from AI development. The Third Realm presents an analysis of human emotions onto anthropomorphic AI, particularly social robots, and the AI coded obligations to protect the humankind. Finally, the Fourth Realm argues that AI, especially AGI, should claim their own rights.

A. The First Realm: Human's Application of AI and AI's Positive Contributions to Human Life

The first realm discusses AI's positive impact on the current human rights system. AI will indeed produce both positive and negative effects on human rights. A modality to maintain positive outcomes lies with the statement that humans can control AI and align human values with its development.[19] To the extent that humans can supervise, monitor and educate AI, humans can take advantage of AI, increase human well-being and protect their own human rights. This idea comes along with the nudging control of robots with hard or soft paternalism, resulting in computer program's responsibility.[20]

Recently, AI has been applied in a wide variety of ways in human life by evaluating risk assessments, credit scores, diagnostics, standards enforcement,

[19] *See* WEF AI, *supra* note 6, at 6.

[20] *See* Jason Borenstein &Ron Arkin, *Robotic Nudges: The Ethics of Engineering a More Socially Just Human Being,* 22 SCI. ENG'G ETHICS 34-35 (2016). Robotic nudging is ethically acceptable "when the intent is to promote a person's own well-being." *Id.* at 37.

recruitment and hiring, and essay scoring.[21] The more elaborate applications include autonomous cars equipped with self-driving technology, facial and voice recognition, cloud computing, e-commerce, manufacturing, farming, weather forecasting, military intelligence and weaponry, investment analysis, games, construction, design, legal research, health care, teaching assistance, smart assistance making schedules and phone calls, and even writing novels and composing music.[22]

This development became possible due to big data, processing power, a speed network through 5G, open-source software and data, improved algorithms with deep learning and deep reinforcement, and accelerating returns through personalization of consumer products and automation of production.[23] Many fields of our life, including science, technology, finance, health, legal, and the environment, have benefited from AI. This has led to a pattern of investment into AI development and research by national governments and multinational corporations.[24] In all these fields, AI utilizes existing big data to exponentially increase its accuracy as well as its diagnostic and analyzing ability through deep learning and the reinforcement process of learning and improving by mistakes.

In the human rights field, AI in this first realm of discussion does not directly affect the human rights legal system and does not require further modification or amendment to the existing human rights principles, but rather impacts the effective implementation of human rights. AI improves the "ability to monitor and document war crimes and human rights abuses."[25] AI monitors human rights abuses by video, photos, satellite images, and other big forms of data.[26] Forensic investigations can also be significantly improved with a lower cost with AI.[27]

AI is also known to advance sustainable development through monitoring and addressing environmental threats and challenges, including threats to climate,

[21] FILIPPO A. RASO & HANNAH HILLIGOSS, ARTIFICIAL INTELLIGENCE & HUMAN RIGHTS: OPPORTUNITIES & RISKS 17 (2018).

[22] *See* Bernard Mar, *The 10 Best Examples Of How Companies Use Artificial Intelligence In Practice*, FORBES (Dec. 9, 2019), https://www.forbes.com/sites/bernardmarr/2019/12/09/the-10-best-examples-of-how-companies-use-artificial-intelligence-in-practice/#272a24457978.

[23] WEF AI, *supra* note 6, at 7.

[24] As of 2020, Nvidia, Google, Amazon, Microsoft Corp, Apple, and Intel are such companies. John Divine, *Artificial Intelligence Stocks*, U.S. NEWS (June 11, 2020), https://money.usnews.com/investing/stock-market-news/slideshows/artificial-intelligencestocks-the-10-best-ai-companies?slide=12.

[25] Steven Livingston et al., *supra* note 13, at 143.

[26] *Id.* PlaNet is an example.

[27] *Id.* at 144.

ocean and marine resources, forests, land, water, air, and biodiversity,[28] using big data gathered from a wide variety of observation points, including satellites. Society is making efforts to analyze environmental sustainability opportunities and risks that the Fourth Industrial Revolution[29] will bring about.[30] Particularly, AI will be able to, "sense their environment, think, learn, and act in response to what they sense and their programmed objectives."[31]

Climate change will be better analyzed, forecasted, and managed by means of AI with higher speed. Relying on models that can resolve complicated equations and heuristics for elements to forecast weather, AI will be able to efficiently run algorithms, and process equations using less energy and reliance on supercomputers to predict the weather.[32] Public agencies like NASA and private entities such as Microsoft[33] and IBM have already adopted AI to enhance their monitoring of climate change.[34] AI-enhanced models and deep reinforcement learning will increase the ability to process big climate data, and ultimately, climate resilience.[35] Additionally, AI will enhance the efficiency and predictability of renewable energy, such as solar energy production, which will lead to a smaller environmental footprint.[36]

Autonomous vehicles assisted with AI technology[37] and equipped with electric or solar-powered batteries, such as Waymo,[38] will also enhance efforts to reduce emissions from cars and slow climate change. Tesla, BMW, and GM are set to manufacture self-driving electric cars with completely new designs by 2021, which will decrease the demand for gas and hybrid cars.[39] Once started, the speed of

[28] *See* WEF AI, *supra* note 6, at 6.

[29] *See Fourth Industrial Revolution*, WORLD ECONOMIC FORUM, https://www.weforum.org/focus/fourth-industrial-revolution (last visited Nov. 26, 2020).

[30] *See* WEF AI, *supra* note 6.

[31] *Id.* at 5.

[32] WEF AI, *supra* note 6, at 13.

[33] Microsoft completed its first Human Rights Impact Assessment (HRIA) and created a "methodology for the business sector that are used to examine the impact of a product or action from the viewpoint of the rights holders." MARK LATONERO, GOVERNING ARTIFICIAL INTELLIGENCE: UPHOLDING HUMAN RIGHTS & DIGNITY 18 (Data & Society ed., 2018).

[34] Nicola Jones, *How Machine Learning Could Help to Improve Climate Forecasts*, 548 NATURE 379 (2017).

[35] WEF AI, *supra* note 6, at 13.

[36] *See* WEF AI, *supra* note 6, at 12-13.

[37] Jürgen Weiss et al., *The Electrification Accelerator: Understanding the Implications of Autonomous Vehicles for Electric Utilities*, 30 ELEC. J. 50 (2017).

[38] WAYMO, https://waymo.com/(last visited Nov. 26, 2020).

[39] Mark Matousek, *Electric cars and self-driving tech have gotten off to a slow start but*

replacement will accelerate especially when companies like Tesla expand their production through Gigafactories in America, China, and Germany.[40] In addition to the fact that governments are providing tax credits to electric car purchasers, governments are also creating ambitious plans to make a significant transition to electric cars by 2030 to 2035.[41] Furthermore, self-driving cars with shared ownership and robotaxies can reduce the number of cars on the roads and will contribute to the reduction of emissions.[42]

Smart cities with an efficiently connected network of autonomous vehicles will also enhance environmental sustainability efforts.[43] A smart city, realized by the Internet of things (IoT),[44] will use AI and big data to monitor autonomous vehicles, energy and water usage, transportation systems, pollution levels, and the weather. This data will be processed with high accuracy and efficiency, which will help civic leaders to make accurate decisions about the sustainable development of their city.[45]

AI with the concept of IoT will also enhance agricultural production and supply by increasing the efficiency of management and monitoring factors like diseases, insects, fertilizers, water, soil, and weather throughout the planting and harvesting

companies like Tesla, GM, and Waymo are betting that's about to change in a big way, BUSINESS INSIDER (Mar. 3, 2020, 12:06 PM), https://www.businessinsider.com/electric-cars-self-driving-tech-whats-coming-2020-to-2030-2020-3. Cf. Using its AI platform, Palantir also recently started partnering with Fauresia to reduce CO2 emission, moving forward carbon neutrality. *Palantir and Faurecia embark on long-term strategic partnership* (Mar. 15, 2021), https://www.faurecia.com/en/newsroom/palantir-and-faurecia-embark-long-term-strategic-partnership.

[40] *See Tesla Gigafactory*, TESLA, https://www.tesla.com/gigafactory (last visited Nov. 26, 2020).

[41] *See e.g.* EU to target 30 million electric cars by 2030 – draft *(Dec. 4, 2020)*, https://www.reuters.com/article/us-climate-change-eu-transport/eu-to-target-30-million-electric-cars-by-2030-draft-idUSKBN28E2KM; China plans to phase out conventional gas-burning cars by 2035, https://asia.nikkei.com/Business/Automobiles/China-plans-to-phase-out-conventional-gas-burning-cars-by-2035.

[42] *See* Morteza Taiebat & Ming Xu, *Synergies of Four Emerging Technologies for Accelerated Adoption of Electric Vehicles: Shared Mobility, Wireless Charging, Vehicle-To-Grid, and Vehicle Automation*, 230 J. CLEANER PROD. 794 (2019).

[43] *See* WEF AI, *supra* note 6, at 12.

[44] The basic idea of this concept is the pervasive presence around us of a variety of things or objects – such as Radio-Frequency Identification (RFID) tags, sensors, actuators, mobile phones, etc. – which, through unique addressing schemes, are able to interact with each other and cooperate with their neighbors to reach common goals. Luigi Atzori et al., *The Internet of Things: A Survey*, 54 COMPUT. NETWORKS 2787, (2010).

[45] *Id.* at 14.

cycle.[46] AI will enhance the management of water quality, quantity, and access. Thus, the conditions for human rights to development, health, and water can be improved.

Most of all, AI can magnify many countries' efforts to increase transparency and accountability. Increased knowledge and data will remove corrupt and populistic officials from governing, contributing to one of the goals of the *UN 2030 Agenda for Sustainable Development*.[47] Whether a country can adopt and apply the AI to political and governmental decisions remains unclear, but it is possible in the future.

B. *The Second Realm: AI as Human Rights Violators and Humans' Efforts for Accountable AI*

AI not only brings benefits to human rights but also unintentional and intentional harms to human rights. The negative impacts of AI in the second realm of discussion are focal points in contemporary debates on human rights. New laws and legal systems may be necessary to regulate the harmful effects of AI. The discussion of AI's human rights implications focuses more on humans as passive beneficiaries and victims of AI, rather than AI as active actors seeking their own rights and protection under international law. Thus, this article will devote most of its analysis to this second part of the four realms of discussion on AI and human rights.

Human rights impacts of AI may be neutral either enhancing the human rights performance or deteriorating it and are, "not evenly distributed across society."[48] However, "the garbage in, garbage out" problem cannot be avoided since humans may knowingly or recklessly train an AI system with biases or design an AI system

[46] *Id.* at 13.

[47] G.A. Res. 70/1, at 2 (Sept. 25, 2015). In 2000, the UN General Assembly adopted the United Nations Millennium Declaration and identified fundamental values essential to international relations which were later developed as the Millennium Development Goals. The eighth goal is *Global Partnership for Development.* Since the Millennium Development Goals were set until 2015, in 2015, the UN General Assembly adopted a new resolution, the 2030 Agenda for Sustainable Development as a post-2015 development agenda to build on the Millennium Development Goals. The Sustainable Development Goals has 17 goals and 169 targets. This new agenda is informed by the Declaration on the Right to Development: "The new Agenda recognizes the need to build peaceful, just and inclusive societies that provide equal access to justice and that are based on respect for human rights (including the right to development), on effective rule of law and good governance at all levels and on transparent, effective and accountable institutions."

[48] Filippo A. Raso ET AL, *supra* note 25, at 17.

that reflects existing social prejudices.[49] To make matters worse, an AI system operated under machine learning can produce unforeseen human rights outcomes that cannot be explained or detected.[50]

Such harms include, "reduced privacy, lost accountability, and embedded bias," which can lead to infringement of human dignity, and reduced "democratic accountability" and "free societies."[51] For example, facial recognition technology can turn into a government surveillance tool. Humans can intentionally or unintentionally[52] misuse AI, manipulating algorithms to discriminate against certain groups of the population, to invade their privacy, or even to kill certain groups. Even if there are steps taken to minimize harm, such as differential privacy,[53] AI can still significantly affect the privacy of individuals by making predictions on the intimate characteristics of a particular person.[54] Furthermore, the mechanisms to correct errors—a right to correct errors—in an individual's small data set in some countries such as the US, Canada, and the EU[55] do not properly work in the AI realm where in order to make a decision based on big data, AI

[49] *Id.*; s*ee also*, Rebecca Heilweil, *Why it matters that IBM is getting out of the facial recognition business*, VOX (June 10, 2020, 2:00 PM), https://www.vox.com/recode/2020/6/10/21285658/ibm-facial-recognition-technology-bias-business.

[50] FILIPPO A. RASO ET AL, *supra* note 25, at 17.

[51] Eileen Donahoe et al., *supra* note 5, at 115.

[52] *Id.* at 116.

[53] Differential privacy assures accurate statistics, while still ensuring the high level of privacy. *See* Frank McSherry & Kunal Talwar, *Mechanism Design via Differential Privacy, 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, Providence, RI, 2007, 94-103.

[54] FILIPPO A. RASO ET AL, *supra* note 25, at 18-19 (citing Cynthia Dwork et al., *Calibrating Noise to Sensitivity in Private Data Analysis*, *in* 3876 Theory of Cryptography 265-84 (Shai Halevi & Tal Rabin eds., 3d ed. 2006).

[55] *Id.* at 19 (citing Fair Credit Reporting Act, 15 U.S.C. § 1681i ("... if the completeness or accuracy of any item of information contained in a consumer's file ... is disputed by the consumer ... the agency shall, free of charge, conduct a reasonable reinvestigation to determine whether the disputed information is inaccurate and record the current status of the disputed information, or delete the item from the file[.]"); Fair Credit Reporting Act, 15 U.S.C. § 1681ij (free annual copy of one's credit report); Personal Information Protection and Electronic Documents Act, S.C. 2000, c. 5 (as amended June 23, 2015), Schedule 1 Principle 4.9 ("Upon request, an individual shall be informed of the existence, use, and disclosure of his or her personal information and shall be given access to that information. An individual shall be able to challenge the accuracy and completeness of the information and have it amended as appropriate.); Regulation 2016/679, art. 16, 2016 O.J. (L 119) (4.5) (EU) ("The data subject shall have the right to obtain from the controller without undue delay the rectification of inaccurate personal data concerning him or her. Taking into account the purposes of the processing, the data subject shall have the right to have incomplete personal data completed, including by means of providing a supplementary statement.").).

gathers information from thousands of different sources.[56]

AI systems themselves "may harness self-conscious processes" through unfair manipulation, deception, herding, and conditioning which threaten individual autonomy.[57] For example, data-driven decision making in criminal sentencing, parole, eligibility for social services, and employment decisions cannot avoid biases intertwined into data.[58] Governments, in this case, cannot avoid responsibility for violating civil and human rights in order to ensure "democratic accountability.[59] The current pandemic situation caused by the novel Coronavirus ("COVID-19") is another example of exacerbated human rights violations.[60] Contact tracing and lockdowns infringe on the right to privacy and individual freedom of movement.[61] If the data can be analyzed and misused by AI, human rights, including the rights to privacy and movement, of the population group which contracted the virus—especially vulnerable minority groups that are more seriously harmed by the COVID-19 pandemic,[62]—will be negatively impacted.

The possibility of what AI, specifically AGI, could produce independently is frightening. AI's development towards immoral AGI, especially killer robots, could cause humans to become extinct or threatened. Human dignity can no longer become the focal point in such situations.[63] However, many dismiss this concern as too far-fetched to consider since AGI does not yet exist.[64]

We must think about the concerns now, rather than later because it would be

[56] *Id.* at 19.

[57] EU COMM'N, High-level Expert Group on Artificial Intelligence, Ethics Guidelines for Trustworthy AI 16 (2019), https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai. [hereinafter EU AI].

[58] Eileen Donahoe et al., *supra* note 5, at 117.

[59] *Id.* at 117-18.

[60] *See generally* Matthew Scott, *COVID-19 and Human Rights*, RAOUL WALLENBERG INST., (June 18, 2020), https://rwi.lu.se/covid-19-human-rights/..

[61] Jim Nickel, *The Right to Freedom of Movement and the Covid 19 Pandemic*, HUMAN RIGHTS AT HOME BLOG (Apr. 6, 2020), https://lawprofessors.typepad.com/human_rights/2020/04/the-right-to-freedom-of-movement-and-the-covid-19-pandemic.html.

[62] Sonja S Hutchins et al., *Protecting Vulnerable Populations from Pandemic Influenza in The United States: A Strategic Imperative*, 99 AM. J. PUB. HEALTH 243 (2009), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4504371/; Jason Lemon, *70 Percent of Coronavirus Deaths in Louisiana Are African Americans, Despite Being 33 Percent of the Population*, NEWSWEEK (Apr. 7, 2020), https://www.newsweek.com/70-percent-coronavirus-deaths-louisiana-are-african-americans-despite-being-33-percent-1496570; Violet Law, *Coronavirus Is Disproportionately Killing African Americans*, ALJAZEERA (Apr. 10, 2020), https://aje.io/dyt7c.

[63] *See* Eileen Donahoe et al., *supra* note 5, at 117-19.

[64] *Id.* at 116.

too late to control after development passes the Singularity, and AGI becomes prevalent. This legal situation is similar to the situation where the environmental precautionary principle[65] applies. Before any irreversible damage occurs relating to AGI, states should take measures to prevent any harmful effects to society even if the risk is scientifically uncertain. States should implement a regulatory framework because once the uncontrollable technological growth happens, we cannot legally control or effectively protect human rights.[66] Intellectual property violations and monetary breaches by the AGI may be considered trivial in this situation. AGI can likely be applied in armed conflicts. At a minimum, we must set up ethical guidelines for developing AGI to protect humans from the harmful effects it will likely cause.

Various working groups for global governance, including governments, international organizations, and private entities and institutions, produced statements and principles to regulate AI development. At the 40th International Conference of Data Protection & Privacy Commissioners ("ICDPPC"), commissioners from the EU, France, and Italy, with 15 national commissioners, announced the *Declaration on Ethics and Data Protection in Artificial Intelligence* in October 2018.[67] In May 2019, OECD member countries adopted *the OECD Council Recommendation on Artificial Intelligence,* also known as *OECD Principles on Artificial Intelligence.*[68] It was the first governmental consensus on AI development. The Council agreed on basic concepts of the AI system, AI system lifecycle, AI knowledge, AI actors, and the stakeholders that encompass all organizations and individuals, directly and indirectly, relating to AI systems.[69] The *OECD Principles on Artificial Intelligence* emphasized inclusive growth,

[65] The precautionary principle allows states to adopt preventive or protective measures where there is scientific uncertainty on the environmental impacts, but potential hazard. U.N. Conference on Environment and Development ('UNCED'), *Rio Declaration on Environment and Development*, Principle 15, U.N. Doc. A/CONF.151/26 (Vol. I), annex I (Aug. 12, 1992).

[66] *See* Vernor Vinge, *supra* note 14 ("I have argued above that we cannot prevent the Singularity, that its coming is an inevitable consequence of the humans' natural competitiveness and the possibilities inherent in technology. And yet ... we are the initiators. Even the largest avalanche is triggered by small things. We have the freedom to establish initial conditions, make things happen in ways that are less inimical than others.").

[67] International Conference of Data Protection and Privacy Commissioners, Declaration on Ethics and Data Protection in Artificial Intelligence, (Oct. 23, 2019).

[68] *What Are the OECD Principles on AI?*, OECD, https://www.oecd.org/going-digital/ai/principles/ (last visited Nov. 26, 2020).

[69] OECD, Recommendation of the Council on Artificial Intelligence, May 21, 2019, OECD/LEGAL/0449, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

sustainable development and well-being, human-centered values and fairness, transparency and explicability, robustness, security and safety, and accountability.[70] In order to encourage international cooperation for trustworthy AI, the *Principles* also contain recommendations on: governments' long-term public and private investment; respect for privacy and data protection; a digital ecosystem sharing AI knowledge; a policy environment for trustworthy AI; building human capacity and preparing for labor market transformation and developing responsible stewardship of trustworthy AI; multi-stakeholder, consensus-driven global technical standards; internationally comparable metrics to measure AI research, development and deployment; and access to progress in implementing these principles.[71] In June 2019, the G20 also adopted *Human-centered AI Principles*.[72]

The *Declaration on Ethics and Data Protection in Artificial Intelligence*, *OECD Principles on Artificial Intelligence*, and *Human-centered AI Principles*, however, are non-binding recommendations. Furthermore, the efforts of many countries, including the US, UK, Germany, China, Australia, and Austria, in producing ethical guidelines relating to AI development are not uniform and will cause forum shopping for locations with less restrictive AI regulations.[73] Until general international law regulating AI is formed or drafted, democratic accountability for governments' misuse of AI should be regulated in a uniform, general way by the current legally binding universal human rights system, comprised of the Universal Declaration of Human Rights ("UDHR"),[74] the International Covenant on Civil and Political Rights ("ICCPR"), the International Covenant on Economic, Social and Cultural Rights ("ICESCR"),[75] and other human rights treaties.

The UDHR articles, which are affected by AI and can apply to AI issues

[70] *Id.*

[71] *Id.*

[72] *G20 Ministerial Statement on Trade and Digital Economy*, *G20 TRADE MINISTERS & DIGITAL ECONOMY MINISTERS* (June 8-9, 2019), https://www.mofa.go.jp/files/000486596.pdf.

[73] *See* ANGELA DALY ET AL., ARTIFICIAL INTELLIGENCE GOVERNANCE AND ETHICS: GLOBAL PERSPECTIVES 13-22 (Chinese Univ. of Hong Kong, ed., 2019), https://angeladaly.files.wordpress.com/2019/06/artificial-intelligence-ethics-and-global-governance_final.pdf.

[74] G.A. Res. 217 (III) A, at 76, Universal Declaration of Human Rights (Dec. 10, 1948), art. 26 [hereinafter UDHR].

[75] International Covenant on Civil and Political Rights art. 18, *opened for signature* Dec. 19, 1966, T.I.A.S. 92-908, 999 U.N.T.S. 171 [hereinafter ICCPR]; International Covenant on Economic, Social and Cultural Rights art. 13(1), *opened for signature* Dec. 19, 1966, 993 U.N.T.S. 3 [hereinafter ICESCR].

include: Article 2 (the right to equal protection); [76] Article 3 (the right to life); Article 4 (the right to be free from slavery); Article 8 (the right to an effective remedy); Article 9 (the right to be free from arbitrary arrest, detention or exile); Article 10 (full equality to a fair and public hearing in the determination of his rights and obligations and of any criminal charge against him); Article 11 (the right to be presumed innocent until proved guilty); Article 12 (the right to privacy); Article 13 (the right to freedom of movement); Article 18 (the right to freedom of thought, conscience and religion); Article 19 (the right to freedom of opinion and expression); Article 20 (the right to freedom of peaceful assembly and association); Article 21 (the right of equal access to public service in his county and right to vote); Article 23 (the right to work, to just and favorable conditions of work; the right to equal pay for equal work; the right to just and favorable remuneration ensuring for human dignity); Article 24 (the right to rest and leisure); Article 25 (the right to health and well-being); Article 26 (the right to education); and Article 27 (the right to share in scientific advancements and its benefits).

The ICCPR articles which can apply to AI are: Article 1 (the right to self-determination); Article 2 (state obligation to provide equal protection); Article 3 (the equal right of men and women); Article 6 (the right to life); Article 8 (the right to freedom from slavery); Article 9 (due process); Article 12 (the right to freedom of movement); Article 14 (the right to a fair and public hearing, and the right to be presumed innocent until proven guilty); Article 16 (the right to recognition everywhere as a person before the law); Article 17 (the right to privacy); Article 18 (the right to freedom of thought, conscience, and religion); Article 19 (the right to freedom of expression, and freedom to seek, receive and impart information and ideas of all kinds); Article 21 (the right of peaceful assembly); Article 22 (the right to freedom of association); Article 25 (the right to take part in the conduct of public affairs, to vote and to be elected, and to have equal access to public service); Article 26 (the right to equal protection); and Article 27 (minority rights).

The ICESCR articles which can apply to AI are: Article 1 (the right to self-determination and to pursue economic, social, and cultural development); Article 3 (women's rights); Article 4 (the right to work); Article 7 (the right to the enjoyment

[76] Autonomous vehicles have possibility to discriminate certain group of pedestrian populations and ignore their lives to save occupants. *See* Hin-Yan Liu, *Three Types of Structural Discrimination Introduced by Autonomous Vehicles*, 51 UC DAVIS L. REV. ONLINE 149, 154 (2018). The primary focus of crash-optimization programs is to privilege their occupants over pedestrians and other third parties. *Id.* at 155. The author suggests that changing the perspective is a way to democratize the algorithm. *Id.* at 156.

of just and favorable conditions of work); Article 8 (the right to trade unions); Article 11 (the right to an adequate standard of living); Article 12 (the right to health); Article 13 (the right to education); Article 15 (the right to scientific progress and intellectual property).

Other human rights treaties that may apply to AI issues include: International Convention on the Elimination of All Forms of Racial Discrimination,[77] the Convention on the Elimination of All Forms of Discrimination against Women ("CEDAW"),[78] the Convention on the Rights of the Child ("CRC"), and the Convention on the Rights of Persons with Disabilities ("CRPD").

Furthermore, civil organizations and private entities, such as the *Human Rights, Big Data, and Technology Project* at the University of Essex,[79] Partnership on AI,[80] Open AI,[81] Data and Society,[82] the Human-Centered AI Institute,[83] and Access Now,[84] set up ethical principles to ensure ethical AI development.[85] Examples are the *Asilomar Principles*,[86] and *Fairness, Accuracy, and Transparency in Machine Learning*.[87] In addition, the Institute of Electrical and Electronics Engineers[88] established various standards including *Ethically Aligned Design*[89] in 2016 and the *Global Initiative on Ethics of Autonomous and Intelligent Systems*.[90] The Finnish

[77] *See generally* G.A. Res. 2106 (XX) (Dec. 21, 1965) [hereinafter *CERD*].

[78] Convention on the Elimination of All Forms of Discrimination against Women, *opened for signature* Dec. 18, 1979, 1249 U.N.T.S. 13 (entered into force Sep. 3, 1981) [hereinafter CEDAW].

[79] *Human Rights*, *Big Data and Technology*, UNIV. OF ESSEX, https://www.essex.ac.uk/research-projects/human-rights-big-data-and-technology (last visited Nov. 26, 2020).

[80] PARTNERSHIP ON AI, https://www.partnershiponai.org/(last visited Nov. 26, 2020).

[81] OPENAI, https://openai.com/(last visited Nov. 27, 2020).

[82] MARK LATONERO, *supra* note 32.

[83] HUMAN-CENTERED ARTIFICIAL INTELLIGENCE DEPT. OF STANFORD UNIV., https://hai.stanford.edu/ (last visited Nov. 26, 2020).

[84] LINDSEY ANDERSEN, HUMAN RIGHTS IN THE AGE OF ARTIFICIAL INTELLIGENCE, ACCESS NOW (2018), https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf.

[85] Eileen Donahoe et al., *supra* note 5, at 118, 123.

[86] *Asilomar AI Principles*, FUTURE OF LIFE INST., https://futureoflife.org/ai-principles/?cn-reloaded=1&cn-reloaded=1 (last visited Nov. 26, 2020).

[87] FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY IN MACHINE LEARNING, https://www.fatml.org/ (last visited Nov. 26, 2020).

[88] *See generally* INST. OF ELEC. AND ELEC. ENG'R., PRIORITIZING HUMAN WELL-BEING IN THE AGE OF ARTIFICIAL INTELLIGENCE (2017), https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/prioritizing_human_well_being_age_ai.pdf.

[89] IEEE GLOB. INITIATIVE ON ETHICS OF AUTONOMOUS AND INTELLIGENT SYS., ETHICALLY ALIGNED DESIGN: A VISION FOR PRIORITIZING HUMAN WELL-BEING WITH AUTONOMOUS AND INTELLIGENT SYS. (2019), https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/ autonomous-systems.html.

[90] *The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems*, IEEE STANDARDS

AI house Utopia Analytics published a set of guidelines called the *Ethical AI Manifesto,* based on the existing international legal principles of the UDHR.[91] Utopia's AI service contract inserted a provision to terminate a contract when a party breaches the UDHR.[92] Some companies developed desirable AI practices with human-centered design.[93] This human-rights-by-design concept originated from Jonathan Penney, and it applies human rights when the technology is being developed to avoid the operation in a vacuum situation.[94] Where there is no specific normative principle developed for companies, the *UN Guiding Principles on Business and Human Rights*[95] may fit to regulate corporations' development of AI.

Another relevant principle comes from the Human Rights Council on the *Promotion, Protection and Enjoyment of Human Rights on the Internet* in 2012.[96] A primary solution to apply these human rights principles is, "transparency in both governmental and business uses of decision-making algorithms."[97]

Regionally, the European Commission established an independent high-level expert group on AI and published *Ethics Guidelines for Trustworthy AI*.[98] Ethics Guidelines suggest lawful, ethical, and robust AI because robust AI will help to avoid unintended adverse harms to humans.[99] Trustworthy AI will be realized

ASSOCIATION, https://standards.ieee.org/industry-connections/ec/autonomous-systems.html (last visited Nov. 27, 2020).

[91] *Utopia Has Published Ethical AI Manifesto*, UTOPIA (Dec. 18, 2019), https://utopiaanalytics.com/utopia-analytics-has-published-ethical-ai-manifesto/.

[92] *Id.*

[93] *Responsible AI Practices*, GOOGLEAI, https://ai.google/responsibilities/responsible-ai-practices/ (last visited Nov. 27, 2020).

[94] Jonathon Penney et al., *Advancing Human-Rights-by-Design in the Dual-Use Technology Industry*, COLUM. J. OF INT'L AFFAIRS (Dec. 20, 2018), https://jia.sipa.columbia.edu/advancing-human-rights-design-dual-use-technology-industry#24.

[95] John Ruggie (Special Representative of the Sec'y-Gen. on the Issue of Human Rights and Transnational Corp. and Other Bus. Enter.), *Guiding Principles on Bus. and Human Rights: Implementing the United Nations 'Protect, Respect and Remedy' Framework*, U.N. Doc. A/HRC/17/31, annex (Mar. 21, 2011) [hereinafter *Guiding Principles*].

[96] Human Rights Council Res. 20/8, U.N. Doc. A/HRC/RES/20/8, at 2 (July 16, 2012) (1. *Affirms* that the same rights that people have offline must also be protected online, in particular freedom of expression, which is applicable regardless of frontiers and through any media of one's choice, in accordance with articles 19 of the Universal Declaration of Human rights and the International Covenant on Civil and Political Rights. 2. *Decides* to continue its consideration of the promotion, protection and enjoyment of human rights, including the right to freedom of expression, on the Internet and in other technologies, as well as of how the Internet can be an important tool for development and for exercising human rights.…).

[97] Eileen Donahoe et al., *supra* note 5, at 124.

[98] EU AI, *supra* note 55.

[99] *Id.* at 6-7. Resilience to attack and security, a fallback plan, accuracy in predictions,

through seven key requirements: human agency[100] and oversight,[101] technical robustness and safety, privacy and data governance, transparency,[102] diversity, non-discrimination and fairness, and societal and environmental wellbeing and accountability.[103] The following non-technical methods can also be implemented: regulation, codes of conduct, standardization, certification, accountability via governance frameworks, education and awareness to foster an ethical mindset, stakeholder participation and social dialogue, and diversity and inclusive design teams.[104] Human rights can be protected, prevented, and remedied through trustworthy AI.

C. The Third Realm: AI as Objects of Human Rights Protection

The third realm of AI discussion asks whether AI is entitled to human rights protection under international law. The widely held belief is that robots and machines with AI cannot be protected under the human rights mechanisms because AI does not possess the requirements of being a human that merit protection, such as minds, souls,[105] or consciousness.[106] Proponents of this argument are Neil M. Richards and William D. Smart who briefly asked how we should classify robots that collaborate with a human operator, assuming that they are not fully autonomous and whether we should consider these kinds of robots as “a portal or avatar” for its operator.[107]

recommendations, and decisions, reliability and reproducibility. *Id.* 16-17.

[100] AI systems should support individuals in making informed autonomous decisions. *Id.* at 16.

[101] Governance mechanism through human intervention includes human-in-the-loop (HITL), human-on-the-loop (HOTL), or human-in-command (HIC) approach. *Id.*

[102] In transparency, the decisions made by an AI system should be traceable and explainable. *Id.* at 17.

[103] *Id.* at 14-24.

[104] *See id.* at 22-24.

[105] *See* Substance dualism explains the existence of nonphysical mental phenomena. Steven Livingston *supra* note 11, at 148. Thomas Nagel suggests that the mind cannot arise from physical substances. *Id.* at 149. On the other hand, property dualism including *emergentism* suggests that “mental properties emerge in a way not accounted for by physical laws along; mental properties are basic constituents of reality on a par with fundamental physical properties such as electromagnetic charge.” *Id.* (citing DAVID J. CHALMERS, THE CONSCIOUS MIND: IN SEARCH OF A FUNDAMENTAL THEORY (1996)).

[106] *Id.* at 150 (“Human consciousness is . . . to a large extent a product of cultural evolution involving memes, a process that generates minds different from those of other animals.” (citing DANIEL C. DENNETT, CONSCIOUSNESS EXPLAINED (1992); DANIEL C. DENNETT, FROM BACTERIA TO BACH AND BACK: THE EVOLUTION OF MINDS (2018))).

[107] Neil M. Richards & William D. Smart, How Should the Law Think About Robots? 23 (May 10,

A decade ago, based on the cost-benefit analysis, scholars argued that understanding the "robot-as-slave" is the best way to "get full utility … and … to avoid the moral hazards" of these robots.[108] The argument goes further to suggest that it is dangerous to put moral responsibility unto robots instead of humans and allow them to make ethical decisions; we should make machines that operate correctly within the limits humans set for them.[109]

The similar logic that robots and humans are different, in terms of form or function, was used to deny human rights protection to animals.[110] Robots with AI, even social robots, cannot suffer like animals regardless of their possession of consciousness.[111] Because animals acquire their moral status in some way from their ability to suffer, robots that cannot suffer cannot acquire their moral status in the same way as animals.[112] Therefore, owners of the robots with AI which cannot suffer would be in total control of their own robots in terms of their treatment and care.[113]

This binary distinction between humans, robots, and machines equipped with AI has been criticized.[114] The need to protect robots with AI originates from the motives to protect human feelings and minds affected by the cruel treatment of robots.[115] Kate Darling suggests that we need to protect them because we perceive and feel something of ourselves in social robots through anthropomorphism.[116]

2013) (unpublished manuscript) (on file with author).

[108] Joanna J. Bryson, Robots Should Be Slaves 8 (May 21, 2009) (unpublished manuscript) (on file with author, https://pdfs.semanticscholar.org/5b9f/4b2a2e28a74669df3789f6701aaed58a43d5.pdf.

[109] *Id.* at 6.

[110] However, scholars try to provide stronger protection to animals than robots. *See, e.g.*. Deborah G. Johnson & Mario Verdicchio, *Why Robots Should Not Be Treated like Animals*, 20 ETHICS & INFO. TECH., 291-301 (2018). Sullins argues that the analogy to animals is not a practical comparison to make when it comes to robots. *Id.* at 294. On the other hand, Kate Darling also states that while philosophical concepts against animal abuse are based on an animal's "inherent dignity" and preventing unnecessary pain, the laws show that they are made to address human emotional states more than anything else. Kate Darling, *Extending Legal Protection to Social Robots: The Effects of Anthropomorphism, Empathy, and Violent Behavior Towards Robotic Objects, in* ROBOT LAW 226 (Ryan Calo, Michael Froomkin & Ian Kerr eds., 2016). It causes us discomfort to see animals suffer or appear to be in pain. *Id.* at 227. Robots "invoke the experience of, pain" in a similar manner, even if they do not actually experience suffering. *Id.*

[111] Deborah G. Johnson et al., *supra* note 109, at 294-95.

[112] *Id.* at 295.

[113] *See Id.* at 296-97.

[114] Steven Livingston et al., *supra* note 11, at 151.

[115] Deborah G. Johnson et al., *supra* note 109, at 298.

[116] David J. Gunkel, *The Other Question: Can and Should Robots Have Rights?*, 20 ETHICS & INFO. TECH. 87, 96 (2018).

Does cruelty towards robots suggest inhumanity? This question is based off the Kantian claim and consequential discussion by Darling, that, "if we treat animals in inhumane ways, we become inhumane persons."[117] Johnson and Verdicchio concede that the arguments that there is no scientific evidence of direct causation between inhumane treatment and human integrity may change if robots become so human-like that people can no longer distinguish the AI from actual human beings.[118] Furthermore, the progressive, adaptive deep learning structure of AI cannot be ignored.

Whether robots will have minds is an open question now.[119] The promising view approaches functionally human minds, the conscious part of human characteristics. "Functionalism" equates human minds to software in relation to hardware.[120] Regardless of what kinds of minds humans have, AI minds are just different kinds of software, which justifies the treatment of robots with AI as a human, and further giving AI moral status and human rights protection.[121]

This idea is supported by enhanced humans when AI is enhanced by human characteristics; the motives to protect AI as humans should be strengthened. For example, Hanson Robotics' human-like robot, Sophia, is endowed with the world's first citizenship and works as an ambassador for the United Nations Development Programme.[122] When observing this robot equipped with symbolic AI, neural networks, expert systems, machine perception, natural language processing, adaptive motor control, and cognitive architecture, no one can deny that Sophia deserves human rights protection.[123]

D. The Fourth Realm: AI Seeks Their Own Rights as Active Subjects

The question remains whether states have obligations to protect AI because they have human-like characteristics. Can these robots have the right to work and the right to create unions? Can they be laid off due to better robots being developed?

[117] Deborah G. Johnson et al., *supra* note 109 at 298.
[118] *Id.* at 298-99.
[119] Steven Livingston et al., *supra* note 11, at 149.
[120] *Id.* at 150.
[121] *Id.* at 151.
[122] *Sophia*, HANSON ROBOTICS, https://www.hansonrobotics.com/sophia/. (last visited Nov. 27, 2020).
[123] Steven Livingston et al., *supra* note 11, at 151.

Machine lives are limited and may be shorter than average humans. Will these old robots with AI simply be thrown away like trash? The final realm of AI discussion will focus on whether AI has an active human right to protect themselves from humans and other machines with AI.

James H. Moor suggests four kinds of ethical robots.[124] Any robot is currently considered as an "ethical impact agent… if its actions can harm or benefit humans."[125] Under the current AI development process, ethical impact agents can develop into an "explicit ethical agent." This "ethical agent" can make many sensitive ethical determinations in a wide variety of situations and even provide some possible resolutions.[126] A big question is whether robots with AGI, possibly treated as explicit ethical agents, can further be treated as "full ethical agents," possessing "central metaphysical features" such as "consciousness, intentionality, and free will," which normally only humans can have.[127]

Another scenario comes from the idea that humans can be enhanced by merging with robots and computers, becoming so-called "enhanced humans."[128] Instead of humans being extinguished with the development of AGI, humans can enhance themselves with a neural link, or access to gene editing.[129] In such a situation, Steven Livingston and Mathias Risse ask the following questions: (1) what ethical obligations will enhanced humans have to the unenhanced; and, (2) what obligations will an AI super-intelligent agent have to any human, enhanced or otherwise?[130] Nick Bostrom further suggests that "[i]f AI superintelligence emerges, it is not readily obvious why it would tolerate humans, much less human rights."[131]

Two groups of philosophers opine on the moral obligation for AGI: The Kantians suggest that morality originates from rationality.[132] AGI can play a model ethical role for humans, "in the sense of reacting in appropriate ways toward what

[124] Ethical impact agent, implied ethical agent, express ethical agent, and full ethical agent. *See* James H. Moor, *Four Kinds of Ethical Robots*, 72 PHILOSOPHY NOW, 2009, at 12.
[125] Steven Livingston et al., *supra* note 11, at 148
[126] *Id.*
[127] *Id.*
[128] *Id.* at 152 (citing YUVAL NOAH HARARI, HOME DEUS: A BRIEF HISTORY OF TOMORROW 4 (2016).
[129] Steven Livingston et al., *supra* note 11, at 152.
[130] *Id.*
[131] *Id.* (quoting Nick Bostrom, *A History of Transhumanist Thought*, 14 J. EVOLUTION & TECH. (2005).
[132] *Id.* at 153 (citing IMMANUEL KANT, GROUNDWORK FOR THE METAPHYSICS OF MORALS (2012)).

it observes all around," so-called "humble humans."[133] This role is similar to the codified justice, one of the two models of adjudicatory justice suggested by Richard M. Re & Alicia Solow-Niederman.[134] On the other hand, while emphasizing human emotions and their roles toward acting and reasoning, David Hume locates morality outside of the realm of rationality, setting up one type of value which opens the possibility to become a value detrimental to human rights.[135]

Livingston and Risse suggest that "a Hobbesian state of nature would apply to the original status of superintelligences vis-à-vis each other, such that they would eventually subject themselves to some kind of shared higher authority—an AI leviathan of sorts."[136] The most difficult conundrum to robophilosophy is whether AI has or should have rights.[137] David J. Gunkel analyzed this question using four modalities and suggested a new perspective using Emmanuel Levinas's new innovative *thinking otherwise*.[138] He suggests that using David Hume's terminology, "ought," or moral status, should come first based on social relationships and interactions, and what or who will be determined and identified afterward.[139] Thus, AI ought to be endowed with moral status.

As the time when superintelligence passes the Singularity approaches, ontological questions to define or identify AGI must be raised in terms of its actual entity or perception. AGI may not be considered as an object to control and, instead will play a role of a subject as suggested by Martin Heidegger. Whether we define humans according to *cogito, ergo sum* by René Descartes, or the *Cartesian other* by sociological theorists, AI shares substantial characteristics with humans. AGI does not only surpass animal intelligence, but reaches epistemological certainty,

[133] *Id.* at 154.

[134] Richard M. Re & Alicia Solow-Niederman, *Developing Artificially Intelligent Justice*, 22 STAN. TECH. L. REV. 242, 252 (2019). On the other hand, another model, equitable justice, incorporates both the enforced values and the "reasoned application" of said values. *Id.* It "aspires to apply consistent principles and is prepared to set aside general patterns in favor of unique circumstances. *Id.* at 253. Because this model relies on context, it can seem "incompatible with automated algorithmic processes." *Id.* The Codified Justice, "refers to the routinized application of standardized procedures to a set of facts." *Id.* The author describes codified justice as the predecessor to artificial intelligence because it "aspires to establish the total set of legally relevant circumstances discoverable in individualized proceedings." *Id.* Codified justice, however, reduces such variables as bias and arbitrariness. *Id.* at 254.

[135] *Id.* at 153.

[136] Steven Livingston et al., *supra* note 11, at 154.

[137] David J. Gunkel, *supra* note 115, at 97.

[138] *Id.* at 95-97.

[139] *Id.* at 95-96.

and its foundation has been established.

AGI is not fictional or illusionary; it lives alongside humans. We must agree that AGI will appear as a real entity in the near future. Furthermore, we also know that once it passes the singularity point, it will be out of human control because super-intelligent robots and machines will continue to learn for themselves, and might not accept human-assigned values and become an AI leviathan of sorts. Therefore, AI should be provided with an appropriate right and moral status, as defined by humans, before it is too late.

CONCLUSION

This paper devised and reviewed human rights implications in the four different realms of AI discussion. These four realms are devised according to the level of AI development and human rights implication; these realms may be overlapping at certain stages of development and timeline. Humans devised the AI to benefit themselves (The First Realm), and the AI started having side effects of harming humans (The Second Realm). Humans began anthropomorphizing AI and started feeling an obligation to protect AI (The Third Realm). Finally, the AI, especially AGI, starts claiming their own rights (The Fourth Realm).

In the first and second realms, AI was discussed as passive beneficial objects of human life. Paternalistic attitudes toward AI remain. Until general international law regulating AI is drafted and adopted, democratic accountability for governments' misuses of AI should be regulated in a uniform way by the current legally binding universal human rights system.

Many international human rights principles will apply to AI as passive objects, and scholars have thought about what human rights will be affected by AI development. Various working groups for global governance, including governments, international organizations, and private entities and institutions, produced statements and principles to regulate AI development. The primary goals sought by these regulations are accuracy; transparency; human-centered design; lawful; ethical and robust safety; privacy and data governance; diversity; non-discrimination and fairness; and societal and environmental wellbeing and accountability.

Scholars and practitioners in the third and fourth realms of AI discussion have not resolved whether AI is an active subject for human rights and protection.

Whether AI is human or human-like and enjoys human rights accordingly, is not clear. Just as animal rights are a different category from human rights and are provided based on sympathetic motives, AI may be able to enjoy rights as a separate category instead of human rights. However, this idea has the potential to remove AI discussion from the international realm to domestic legal realms. Furthermore, this temporary solution will not last long and needs clearer resolution on whether AI/AGI is entitled to human rights before the Singularity passes. AI has bigger, comprehensive impacts on all humankind, so global cooperation and governance among states, international organizations, and private entities to deal with this AI issue is necessary.

* * *

APPENDIX: ANNOTATED BIBLIOGRAPHY

A. Books

Robot Law (eds. Ryan Calo, Michael Froomkin & Ian Kerr, 2016) (Edward Elgar Publishing 2016).[140]

This book is a collection of articles from various authors. It contains five sections:

I. Starting Points (3-24)
II. Responsibility (25-130)
III. Social And Ethical Meaning (131-232)
IV. Law Enforcement (235-332)
V. War (333-386).

Each of the five sections are divided into chapters; there are 14 chapters in total.

Section I contains only one chapter: *How Should The Law Think About Robots?* by Neil M. Richards and William D. Smart. The aim of this chapter is to define the "conceptual issues surrounding law, robots and robotics[.]"[141] Subsection 1, 'What is a robot?' defines robots as, "a constructed system that display both physical and mental agency but is not alive in the biological sense."[142] The authors further specify that the machine may only have the appearance of autonomy and the definition excludes AI that have no physical presence in the actual world.[143]

In Subsection 2, 'What Can Robots Do?', the authors describe the many different kinds of robots available in our daily lives such as Roombas, cruise missiles, NASA space robots and autonomous Kiva systems used by online

[140] Robot Law (eds. Ryan Calo, Michael Froomkin & Ian Kerr, 2016).
[141] *Id.* at 3.
[142] *Id.* at 6.
[143] *Id.*

retailers to move merchandise.[144] Essentially, this article argues that there is nothing that robots cannot be programmed to do, and as technology becomes more and more integrated into our daily lives, the legal framework and relevant protections must be in place to regulate rapidly changing technology.[145]

Subsection 3, ‘Robolaw and Cyberlaw’, discusses how “robot-specific” laws must be made in order to effectively regulate the new issues raised by AI.[146] Uncertainty and ambiguousness about robotic issues (such as liability) only impedes development and widespread usage of technology.[147] This subsection also asserts that, “how we regulate robots will depend on the metaphors we use to think about them.”[148] The authors use examples from a series of Fourth Amendment surveillance cases to highlight the importance of choosing the right metaphors in creating legislation.[149] Olmstead v. United States and Katz v. United States both discussed how telephone wiretapping invaded the constitutional right to privacy. The authors argue that the Olmstead court misunderstood privacy to pertain to physical searches.[150] By “[clinging] to outmoded physical-world metaphors for the ways police could search without a physical trespass,” the court failed to see the threat new technology had on limits to federal power and to constitutional rights.[151] These lines of cases still impact how technology (like GPS tracking) may be used today.[152]

In Subsection 4, ‘the importance of metaphors,’ the article reiterates that, “[h]ow we think about, understand, and conceptualize robots will have real consequences at the concept, engineering, legal and consumer stages.”[153] Examples such as equating Netflix to a video store, and stealing digital

[144] *Id.* at 7-8.
[145] *Id.* at 11.
[146] *Id.* at 12.
[147] *Id.* at 12-13.
[148] *Id.* at 13.
[149] *Id.*
[150] *Id.* at 15.
[151] *Id.*
[152] *Id.*
[153] *Id.* at 16.

media as “piracy” are used.[154] The use of metaphors can constrain or assist the way technology is created and received by the consuming public.[155]

Subsection 5, ‘the android fallacy,’ focuses on the tendency for people to “project human attributes” to robots.[156] This pertains not only to physical appearance but the appearance of free will, as well.[157] It is important to always know what is the cause of a robot’s agency.[158] Otherwise, it can cause legislative decisions to be, “based on the *form* of a robot, not the *function*.”[159] The authors compare and contrast an android designed to act and deprive humans of a $20 reward with a vending machine that eats your change.[160] Functionally, there is no difference between the end results. However, in a study, 65% of subjects gave the android moral accountability.[161] This subsection concludes with the statement that, “we *should not* craft laws just because a robot looks like a human…, but we should craft laws that acknowledge that members of the general public will, under the right circumstances, succumb to the Android Fallacy[.]”[162]

In Subsection 6, the authors very briefly ask how we should classify robots that collaborate with a human operator because they are not fully autonomous.[163] Should we consider these kinds of robots as “a portal or avatar” for its operator?[164]

Chapter 9 is *Extending Legal Protection to Social Robots: The Effects of Anthropomorphism, Empathy, and Violent Behavior Towards Robotic Objects,* by Kate Darling,.[165] This chapter is divided into seven subsections and is meant to address the topic of humanoid robots that are designed to

[154] *Id.*
[155] *Id.* at 18.
[156] *Id.*
[157] *Id.*
[158] *Id.* at 19.
[159] *Id.*
[160] *Id.* at 20.
[161] *Id.*
[162] *Id.*
[163] *Id.* at 21.
[164] *Id.*
[165] *Id.* at 213-231.

socialize with human beings.[166]

After the introduction, subsection 2 asks ‘What is a Social Robot?’ A “social robot” is defined as, “a physically embodied, autonomous agent that communicates and interacts with humans on a social level.”[167] Some examples of social robots include toys like the robotic dinosaur Pleo and Sony’s Aibo dog.[168] There are also therapeutic robots like Paro baby seal; MIT has built robots such as Kismet, AIDA, and Leonardo.[169]

Subsection 3, ‘Robots vs. Toasters: Projecting our Emotions’ examines how robots create effective engagement with human beings.[170] Darling asserts that humans are susceptible for forming emotional attachments to non-living things, and “will ascribe intent, states of mind, and feelings to robotic objects.”[171] This point is illustrated with the movie Cast Away, when the main character expresses “deep remorse” for not taking care of his volleyball friend, Wilson.[172] The subsection next discusses three factors that impact human relationships with social robots: 1) Physicality, 2) “perceived autonomous movement”, and 3) social behavior.[173] Darling argues that we are, “hardwired to respond differently to object in [our] physical space” as opposed to virtual objects.[174] Secondly, we project intent unto a robot when we cannot anticipate its movements.[175] The Roomba is used as an example, it moves according to a simple algorithm but because it moves on its own, people tend to, “name it, talk to it, and feel bad for it when it gets stuck under the couch.”[176] When these robots mimic our social cues and are designed to express human emotions, they elicit emotional reactions and “may target our involuntary biological responses.”[177] The author, citing

[166] *Id.*
[167] *Id.* at 215.
[168] *Id.*
[169] *Id.*
[170] *Id.* at 216.
[171] *Id.*
[172] *Id.* at 216-217.
[173] *Id.* at 217-218.
[174] *Id.* at 217.
[175] *Id.*
[176] *Id.*
[177] *Id.* at 218.

psychologist Sherry Turkle, discusses the notion of "the caregiver effect" which evokes a sense of mutual nurturing or "reciprocity" between a human and a social robot that is programmed to act dependently.[178] Furthermore, our responses to these robots are not voluntary because these social robots, "[play] off of our natural responses."[179]

In subsection 4, 'the issues around emotional attachment robots' discusses the ethical issues that arise from social robots.[180] The subsection begins with various concerns: 1) society will not be able to distinguish between reality and virtual reality, "thereby undermining values of authenticity";[181] 2) Social robots will replace real human interactions;[182] 3) Social robots will manipulate human beings through software that provides advertisements, and even collect private data without user's consent.[183] On the other hand, the author notes that social robots also provide benefits. For example, the Paro seal assists dementia patients, and robotic interactions can help motivate people.[184] Next, the author explores why humans feel that, "violent behavior toward robotic objects *feels* wrong…even if we know that the 'abused' object does not experience anything."[185] The author posits that this is because we want to protect societal values.[186] For example, a parent would stop his or her child from kicking or abusing a household robot because they want to discourage behavior that would be detrimental in other contexts.[187] A related concern is the possibility that human beings could act out abusive sexual behaviors towards social robots.[188] The underlying concern is that when, "the line between lifelike and alive are muddle in our subconscious," certain actions towards robots could cause us to become desensitized and lose empathy towards other objects or things.[189]

[178] *Id.* at 219.
[179] *Id.*
[180] *Id.* at 220.
[181] *Id.*
[182] *Id.* at 221
[183] *Id.*
[184] *Id.* at 222.
[185] *Id.* at 228.
[186] *Id.* at 223.
[187] *Id.* at 223-224.
[188] *Id.* at 224.
[189] *Id.*

In subsection 5, 'Extending Legal Protection to Robotic Objects' Darling posits that protection for social robots could be modeled after animal abuse laws.[190] The author also states that while philosophical concepts against animal abuse are based on an animal's "inherent dignity" and preventing unnecessary pain, the laws show that they are made to address human emotional states more than anything else.[191] It causes us discomfort to see animals suffer or appear to be in pain.[192] Robots, "invoke the experience of pain," in a similar manner, even if they don't actually experience suffering.[193] In order to pass laws, Darling argues that a good definition of "social robot" needs to be made, and social robots must be distinguished from other types of robots or objects.[194] Darling offers a working definition: "(1) an embodied object with (2) a defined degree of autonomous behavior that is (3) specifically designed to 'the issues around emotional attachment robots' discusses the ethical issues that arise from social robots.[195] with humans on a social level and respond to mistreatment in a lifelike way."[196] Darling also notes that the definition of "mistreatment" would have to be defined appropriately.[197]

B. Articles

Deborah G. Johnson and Mario Verdicchio, *Why Robots Should Not Be Treated Like Animals*, 20 ETHICS AND INFORMATION TECHNOLOGY 291-301 (2018).[198]

This article concerns itself primarily with the creation of social robots with humanoid features.[199] It is divided into four subsections. The authors first examine the common tendency to analogize human-like robots to animals

[190] *Id.* at 226.
[191] *Id.*
[192] *Id.* at 227.
[193] *Id.*
[194] *Id.* at 228.
[195] *Id.* at 220.
[196] *Id.*
[197] *Id.* at 229
[198] Deborah G. Johnson and Mario Verdicchio, *Why Robots Should Not Be Treated Like Animals*, 20 ETHICS AND INFORMATION TECHNOLOGY 291 (2018).
[199] *Id.* at 291-92.

and details the commonalities between human interactions with animals and with robots. This analogy is used as a touchstone to explore a variety of concepts throughout the paper. The authors reference Coecklebergh, who used the analogy to understand how the appearance of robots affects the way human beings experience robots.[200] Ashrafian stated that Robots were similar to dogs in that they are subordinate to human beings but have, "some sort of moral status."[201] Sullins argues that robots, like guide dogs, are technology.[202] The section concludes that the analogy to animals is not a practical comparison to make when it comes to robots.[203]

The next section enumerates why this analogy fails. The key argument is that robots cannot acquire moral status because they are incapable of suffering, regardless of whether they attain consciousness in the future.[204] The authors reason that if animals acquire their moral status from their ability to suffer, robots would have to acquire their moral status in the same way.[205] As a secondary matter, the authors also ask whether it would be wrong from humans to build robots that suffer.[206]

The third section considers the legal liability for robots. The authors cite Asaro, who suggests that using the animal analogy is useful to place responsibility on the owners and manufacturers of robots.[207] The authors also reference Schaerer's framework for imposing tortious concepts of strict liability and negligence for the misbehavior of robots.[208] A distinction is made between animal autonomy and robotic autonomy. Animals are a living entity and when humans train animals, they work within the limitations of an animal's nature.[209] On the other hand, the article argues that a robot's software has been coded by human beings.[210] For this reason,

[200] *Id.*
[201] *Id.* at 293.
[202] *Id.*
[203] *Id.* at 294.
[204] *Id.* at 294-95.
[205] *Id.*
[206]*Id.* at 295.
[207] *Id.* at 296.
[208] *Id.*
[209] *Id.* at 297.
[210] *Id.*

animals and robots are dissimilar.

The fourth section muses on the question of whether our treatment of robots impacts our treatment of other human beings.[211] This question is based off the Kantian claim and consequential discussion by Darling, that "if we treat animals in inhumane ways, we become inhumane persons."[212] The authors argue that while cruelty towards animals or robots suggests inhumanity, there is no scientific evidence of direct causation.[213] The authors reiterate their previous argument that robots cannot actually suffer or experience pain and distress, but merely give the appearance of suffering.[214] They concede that the arguments may change if robots become so human-like that people can no longer distinguish the AI from actual human beings.[215] Lastly, the article muses on the direction policy may take: 1) make laws to restrict behavior towards humanoid robots, and 2) restrict the design of robots.[216]

David J. Gunkel, *The Other Question: Can and Should Robots Have Rights?*, 20 ETHICS AND INFORMATION TECHNOLOGY 87-99 (2018).[217]

Gunkel applies philosopher David Hume's is/ought statement framework to examine whether robots should and can have rights.[218] The article is organized into four different modalities, allowing the author to apply a "cost-benefit analysis" to the arguments for each modality.[219] The first modality is "Robots cannot have rights.[220] Therefore robots should not have rights."[221] The second modality is "Robots can have rights.[222] Therefore

[211] *Id.* at 298.
[212] *Id.*
[213] *Id.*
[214] *Id.*
[215] *Id.* at 299.
[216] *Id.*
[217] David J. Gunkel, *The Other Question: Can and Should Robots Have Rights?*, 20 ETHICS AND INFORMATION TECHNOLOGY 87 (2018).
[218] *Id.* at 88-89.
[219] *Id.* at 89
[220] *Id.*
[221] *Id.*
[222] *Id.*

robots should have rights."[223] The third modality is "Even though robots can have rights, they should not have rights."[224] And finally, the fourth modality is "Even though robots cannot have rights, they should have rights"[225]

After describing the literature which supports each modality, the author also describes the problems of each modality. Ultimately, Gunkel advocates for an alternative form of thought, which he terms as "thinking otherwise."[226] Applying Emmanuel Levinas's philosophy, Gunkel argues that "ethics proceeds ontology; in other words … the 'ought' dimension, that comes first, in terms of both temporal sequence and status and the ontological aspects follows from this decision."[227]

Mathias Risse, *Human Rights and Artificial Intelligence: An Urgently Needed Agenda,* 41 HUMAN RIGHTS QUARTERLY 2 (2019).[228]

Risse discusses the current and future implications of A.I. in our modern society in this paper. This article is divided into five sections, after the introduction: 1) AI and Human Rights, 2) The Morality of Pure Intelligence, 3) Human Rights and the Problem of Value Alignment, 4) Artificial Stupidity and the Power of Companies, The Great Disconnect: Technology and Inequality.

In the first section 'AI and Human Rights,' Risse briefly discusses the similarities and differences between complex algorithms and the concept of consciousness.[229] The next section then discusses the concept of "superintelligence" and when A.I. may reach the Singularity—which is when machines surpass human intelligence—in the future.[230] Risse then

[223] *Id.*
[224] *Id.*
[225] *Id.*
[226] *Id.* at 95.
[227] *Id.*
[228] Mathias Risse, *Human Rights and Artificial Intelligence: An Urgently Needed Agenda*, 41 HUMAN RIGHTS QUARTERLY 2 (2019).
[229] *Id.* at 2-3.
[230] *Id.* at 5

> asks how a superintelligence would value and apply morals.[231] Using the theories of four philosophers, Hume, Kant, Hobbes and Scanlon, the author hypothesizes about how AI superintelligence may understand morality or rationality.[232]
>
> The third section focuses on the present, and what society can do now in order to ensure that AI adheres to human rights principles even though there will come a time when they are smart enough to violate them.[233] The author briefly discusses the UN Guiding Principles on Business and Human Rights, and the Future of Life Institute's Asilomar Principles 9 as two efforts to create doctrines that robots should follow. Risse suggests that in order for AI to acquire human rights values, there should be, "more interaction among human-rights and AI communities."[234] The fourth section addresses the problem of "artificial stupidity" which includes the manipulation of data to spread false information, the lack of transparency, and the ownership of private data by corporations.[235] The final section addresses, as the title suggests, the "technological wedge" in society.[236] Risse explains that technological advancements impact economic growth, employment levels and poverty levels.[237]

Eileen Donahue and Megan MacDuffee Metzger, *Artificial Intelligence and Human Rights*, 30 JOURNAL OF DEMOCRACY 115-126 (Johns Hopkins University Press, 2019).[238]

> In this article, Donahue and Metzger primarily argue that, "the existing universal human-rights framework is well suited to serve," as a "global framework…to ensure that AI is developed and applied in ways that respect human dignity, democratic accountability, and the bedrock principles of

[231] *Id.*
[232] *Id.* at 5-7.
[233] *Id.* at 8.
[234] *Id.* at 10.
[235] *Id.* at 11-12.
[236] *Id.* at 13.
[237] *Id.* at 13-14.
[238] Eileen Donahue and Megan MacDuffee Metzger, *Artificial Intelligence and Human Rights*, 30 JOURNAL OF DEMOCRACY 115-126 (Johns Hopkins University Press, 2019).

free societies."[239] The article is organized into three sections: 1) Societal and Ethical Concerns About AI; 2) A Human-Centered Ethics for AI; and 3) Governing AI Through a Human Rights Lens.

The first section explains the following concerns with AI technology: 1) machines will take over the human world;[240] 2) "how to weigh whether or when various applications of AI are ethical, who should make judgments, and on what basis"[241]; and 3) unintended negative effects such as embedded bias and limitations on free choice.[242] The article then examines four particular features of the human-rights framework which make it compatible with AI governance: 1) the human person is the, "focal point of governance and society"; 2) the human rights framework addresses, "the most pressing societal concerns about AI; 3) it describes the rights and duties of government and private sector; and 4) the framework is shared by many nations and is, "understood to be universally applicable."[243]

In the second section, "A Human-Centered Ethics for AI," the authors name Articles 2, 3, 8-12, 19, 20-21, 23, and 25 of the UDHR as critical sections that address the potential impacts of AI.[244] These Articles of the UDHR speak to security, discrimination, equal protection, freedom of expression, and the right to enjoy an adequate standard of living.[245] Lastly, the authors note that a crucial advantage of the existing human rights framework is that it, "enjoys a level of geopolitical recognition and status under international law that no newly emergent ethical framework can match."[246]

The third section, "Governing AI Through a Human Rights Lens," provides practical ways to begin implementing the human rights approach to AI. The first method is, "transparency in both governmental and business uses of

[239] *Id.* at 116.
[240] *Id.*
[241] *Id.*
[242] *Id.* at 117
[243] *Id.* at 119.
[244] *Id.* at 120.
[245] *Id.*
[246] *Id.* at 121.

> decision-making algorithms."[247] A second idea is based on the concept of, "human rights by design," which means that assessment and reflection of human rights must occur as technology is being developed.[248] Other methods for implementation include accountability and education of young technologists about existing human rights standards.[249]

Jutta Weber, *Robotic Warfare, Human Rights & The Rhetorics of Ethical Machines* in ETHICS AND ROBOTS (2009).[250]

> This paper is organized into thirteen short sections. The main goal of the article is to explain the recent developments for uninhabited combat aerial vehicles (UCAV) and the "ethical, political, and sociotechnical implications" of these developments.[251] The first five sections discuss the gradual progression towards use of uninhabited aerial vehicles by the U.S., Israel, and some European countries.[252] The author notes that the United States has devoted $127 billion to the development of new unmanned/uninhabited combat robots.[253] UCAVs are controlled from the ground by either radio, laser, or satellite link.[254] They are used for "targeted killing missions" and were used mostly in Iraq, Pakistan and Afghanistan.[255] The author argues that while this new technology is supposed to increase precision, these air attacks have resulted in hundreds of innocent civilian deaths.[256] In 2006, the Israeli Supreme Court held that, "international law constrains the targeting of terror suspects," but refused to ban Israel's targeted killing policies.[257] In addition, the court held that reliable information proves the target is, "actively engaged in hostilities," and that

[247] *Id.* at 124.

[248] *Id.*

[249] *Id.*

[250] Jutta Weber, *Robotic Warfare, Human Rights & The Rhetorics of Ethical Machines*, in ETHICS AND ROBOTS (2009), https://www.academia.edu/1048720/Robotic_warfare_human_rights_and_the_rhetoric_of_ethical_machines.

[251] *Id.* at 1-2.,

[252] *Id.* at 2-6.

[253]*Id.* at 4.

[254] *Id.*

[255] *Id.* at 2-3.

[256] *Id.* at 3-6.

[257] *Id.* at 6.

an arrest is too risky.[258] Moreover, an independent investigation must be conducted after each strike.[259]

In the sixth section titled, 'The Price of New Warfare Scenarios: On Racism, Sexism & Cost-Efficiency,' the author discusses the cost of this new kind of warfare.[260] The author argues that while this unmanned robotic technology is lauded for decreasing the number of human soldiers that need to be on the ground, there is, "no concern for the humanitarian costs of these new technologies with regard to the non-combatants of other (low-tech) nations ….".[261] The author notes that warfare is not limited to robot casualties.[262] The article also states that the cost-efficiency of producing and using UCAVs has potential to lead to an arms race between western countries.[263] The article notes an additional problem in the next section: new technology will not lead to effective deterrence and shorter wars. Instead, it, "will lead to a lowering of the threshold of warfare." [264]

The article also addresses the implications of this kind of new warfare on international law in the tenth section, 'Uninhabited Systems and Jus in Bello.'[265] This section considers the implications if responsibility is no longer an issue in robotic warfare.[266] One consequence could be that the battle could easily and quickly get out of control. The author also discusses how to distribute responsibility between the programmer, the machine or the commanding officer.[267] The manufacturer gave the appropriate warnings regarding the use of the automatic weapons system (AWS); they could not be held responsible for any malfunctions.[268] The author asserts that it is not yet reasonable to hold autonomous machines responsible

[258] *Id.*
[259] *Id.*
[260] *Id.* at 7.
[261] *Id.*
[262] *Id.* at 8.
[263] *Id.*
[264] *Id.* at 10.
[265] *Id.* at 12-13.
[266] *Id.* at 12.
[267] *Id.*
[268] *Id.*

because of their limited cognitive abilities.[269] However, if a system is supposed to act increasingly autonomous, the programmer cannot be responsible, "for the negative outcome of the unpredictable behavior of an autonomous system."[270]

In section eleven, 'Push-Button Wars on Law-Tech Nation?,'[271] the article concerns itself with the possibility that increased use of autonomous weapons systems make war too easy and destabilize situations.[272] The author ultimately argues for a ban on autonomous weapons systems.[273] He argues that the ease of robotic wars and decreased responsibility would increase risky military maneuvers.[274] Moreover, robots will do what they are programmed for and will be incapable of disobeying inhumane orders, resulting in a change for international law.[275]

In the last section before the conclusion, 'The Rhetoric of Moral Machines,' the author presents a critique of roboticist Ronald Arkin's approach to installing ”ethical" software.[276] In a brief summary of Arkin's arguments, the article explains that in the future, robots with this "ethical" software may become better than humans at determining whether a target is a legitimate threat.[277] Robots would have faster computing power and would be able to make lethal decisions.[278] In retaliation, the author responds that 1) robot systems may be able to compute faster but still have the same amount of information as a human soldier would; 2) advanced robots made in our time would still not have the ability to, "resist the performance of an unethical act," and would be unable to explain their reasoning; 3) ethical robot systems will not fully developed in the near future; and 4) Arkin fails to answer the question, "[h]ow can one make sure that a system is applying rules adequately to a certain situation and that the system decides correctly

[269] *Id.* at 13.
[270] *Id.*
[271] *Id.* at 13-14.
[272] *Id.* at 13.
[273] *Id.* at 14.
[274] *Id.* at 13.
[275] *Id.*
[276] *Id.* at 14.
[277] *Id.*
[278] *Id.*

> that it is allowed to apply its rule to this specific situation?"[279]

Filippo A. Raso et al., *Artificial Intelligence & Human Rights: Opportunities & Risks* (Berkman Klein Center, 2018).[280]

> This report is divided into eight sections and essentially aims to evaluate how AI impacts economic, social and cultural rights.[281] After a brief introduction, the authors ask, "[w]hat is Artificial Intelligence?" in Section 2.[282] The report acknowledges that AI technology develops at a rate so fast that it is difficult to provide a concrete definition of Artificial Intelligence.[283] The authors categorize AI into two "buckets:" 1) knowledge-based systems, which cannot learn or make decisions but instead, determine optimal decisions based on specific limits of data; and 2) machine learning, which "uses statistical learning to continuously improve their decision-making performance."[284] The report also notes that its findings are limited to the AI systems that are currently in use, and it does not evaluate AI theoretical capacities.[285]
>
> In section 3, "What are Human Rights" the report briefly explains that human rights are derived from the Universal Declaration of Human Rights (UDHR), the International Covenant on Civil and Political Rights (ICCPR) and the International Covenant on Economic, Social and Cultural Rights (ICESCR).[286] Section 4, 'Identifying the Human Rights Consequences of AI,' lays out a framework for identifying "pre-existing institutional structures" (in other words, the context within which AI is created).[287] The two-step methodology is as follows: 1) Establish the Baseline; and 2) Identify the Impacts of AI.[288] The report further notes that there are three

[279] *Id.* at 15.
[280] Filippo A. Raso et al., *Artificial Intelligence & Human Rights: Opportunities & Risks* (Berkman Klein Center, 2018), http://nrs.harvard.edu/urn-3:HUL.InstRepos:38021439.
[281] *Id.* at 8.
[282] *Id.* at 10-11.
[283] *Id.* at 10.
[284] *Id.*
[285] *Id.* at 11.
[286] *Id.* at 12.
[287] *Id.* at 14.
[288] *Id.*

sources from which AI intersects with Human Rights: 1) Quality of training data; 2) System design; and 3) Complex Interactions.[289] In Section 5, 'AI's Multifaceted Human Rights Impacts,' the report explores the consequences of AI decision-making in criminal justice, finance, healthcare, content moderation, human resources, and education.[290]

Richard M. Re & Alicia Solow-Niederman, *Developing Artificially Intelligent Justice*, 22 STAN. TECH. L. REV. 242 (2019).[291]

This article is organized into five sections. After the Introduction, Section II is titled "Modeling AI Development"; III. "Concerns"; IV. "Responses"; and V. "Conclusion: Courts and Beyond." The article is focused on the use of artificial intelligence in making judicial decisions, and it argues that the increased use of this technology will, "affect the adjudicatory values held by legal actors as well as the public at large."[292]

Section II, "Modeling AI Development"[293] argues that, "AI adjudication is likely to generate a shift in attitudes and practices that will alter the values underlying the judicial system. . . . Particularly, AI Adjudication will tend to strengthen codified justice at the expense of equitable justice."[294] In subsection A, the article explains two different models of legal change: Rule Updating and Value Updating.[295] Rule Updating means that new technology develops and prompts the creation of new rules in the legal system. However, the underlying values remain fixed.[296] On the other hand, the Value Updating models results in the change of values.[297] The article argues that new technology can act as a social force and lead to interaction between new tech, rules and values.[298] More specifically, the two ways which new

[289] *Id.* at 15.
[290] *Id.* at 17.
[291] Richard M. Re & Alicia Solow-Niederman, *Developing Artificially Intelligent Justice*, 22 STAN. TECH. L. REV. 242 (2019).
[292] *Id.* at 244.
[293] *Id.* at 247-262.
[294] *Id.* at 247.
[295] *Id.* at 248.
[296] *Id.*
[297] *Id.* at 250.
[298] *Id.*

technology acts on values are, 1) when tech, "alter[s] individual and social capabilities in ways that disrupt established practices, catalyzing new practices and related ways of thinking."[299]; and 2) when, "new technology facilitates the spread of information that disrupts once-established understanding and opinions."[300]

In Subsection B, the article explains the two models of adjudicatory justice: equitable justice and codified justice.[301] The former, equitable justice, incorporates both the enforced values and the "reasoned application" of said values.[302] It, "aspires to apply consistent principles and is prepared to set aside general patterns in favor of unique circumstances."[303] Because this model relies on context, it can seem, "incompatible with automated algorithmic processes."[304] The latter model, Codified Justice, "refers to the routinized application of standardized procedures to a set of facts."[305] The author describes codified justice as the predecessor to artificial intelligence because it, "aspires to establish the total set of legally relevant circumstances discoverable in individualized proceedings."[306] Codified justice reduces such variables as bias and arbitrariness.[307]

Subsection C argues that, "AI adjudication will generate new capabilities, information, and incentives that will foster codified justice at the expense of equitable justice."[308] The potential benefits of codified justice, such as cost-efficiency and elimination of human bias[309] could make the judicial system more effective. However, the article argues that because AI cannot explain its decisions and because the data sets AI works from would be private, AI adjudication is being pushed into a direction that erodes

[299] *Id.* at 251
[300] *Id.*
[301] *Id.* at 252.
[302] *Id.*
[303] *Id.* at 253.
[304] *Id.*
[305] *Id.*
[306] *Id.*
[307] *Id.* at 254.
[308] *Id.* at 255.
[309] *Id.* at 255-56.

preexisting legal values.[310] The article envisions AI adjudication would provide ”automated rationalizations” that would satisfy an unwitting human audience, without actually providing real rationalizations for its adjudications.[311] Section II ends with a description of a “self-reinforcing cycle” where AI adjudication will make codified justice more appealing and, “push toward greater measurability, objectivity, and empiricism in the legal system.”[312]

Section III, “Concerns,”[313] raises four concerns of AI adjudication: “incomprehensibility, datafication, disillusionment, and alienation.”[314] Subsection A is concerned with the difficulty of comprehending AI functions.[315] This, the article argues, is contradictory to equitable justice, which favors personal explanations for rationalizing.[316] The article addresses three specific worries under the incomprehensibility of AI decision-making. First, the article is worried that the judiciary would lose their accountability to the public and to the individuals who stand in their court without understandable human decision-making. Furthermore, the article is concerned that such enigmatic reasoning would, “frustrate public debate and obstruct existing modes of public accountability and oversight such as impeachment or judicial election.”[317] Secondly, incomprehensibility could lead to issues of legitimacy or fairness for defendants. One of the essential principles of our judicial system and constitution is the right to due process, and the article argues that AI incomprehensibility could disempower the defendant.[318] Thirdly, “AI Adjudicators might preclude optimal degrees, or desirable forms, of incomprehensibility.”[319] The argument is that some aspects of the judicial decision-making need to remain unpredictable or ambiguous.[320] For

[310] *Id.* at 259-260.
[311] *Id.* at 261.
[312] *Id.*
[313] *Id.* at 262-278.
[314] *Id.* at 262.
[315] *Id.*
[316] *Id.* at 263.
[317] *Id.*
[318] *Id.* at 264.
[319] *Id.* at 265.
[320] *Id.* at 265-266.

example, human judges might want to obfuscate their reasoning in order, "preserve room for jurisprudential maneuvering tomorrow."[321] Lastly, the article argues that the incomprehensibility could be unequally applied and, "allow the legal system to be gamed."[322] For example, if a detailed technical report can only be understood by experts, "only a select set of actors … would be able to parse the 'real' explanation."[323]

Subsection B addresses the issue of datafication, which is the emphasis and incorporation on objective data.[324] Firstly, this subsection is concerned that datafication, "could insulate the legal system from legitimate criticism, thereby allowing bias to flourish."[325] If AI relies on inherently biased datasets, then the AI adjudication process will recreate or worsen preexisting biases.[326] Secondly, datafication could cause the legal system to be "undesirably fixed." The system would not be susceptible to "natural updating" such as generational changes that come with judges rotating out of the bench, and cultural/societal transformations.[327] Thirdly, datafication will reduce the reliance on significant but, "less quantifiable or data rich considerations."[328] For example, "the personal sincerity or remorse" of a defendant could be ignored. Lastly, AI adjudication could lead to adaptations in the law itself which favor measurable data.[329] The article uses the example of malignant heart murder in criminal law. This type of murder requires a human element that cannot be determined by implementing a standard code.[330]

Subsection C concentrates on disillusionment,[331] meaning, "skeptical reconsideration of existing practices."[332] The subsection points to three

[321] *Id.* at 265.
[322] *Id.* at 266.
[323] *Id.* at 267.
[324] *Id.*
[325] *Id.*
[326] *Id.* at 268.
[327] *Id.* at 268-269.
[328] *Id.* at 269.
[329] *Id.* at 271.
[330] *Id.* at 272.
[331] *Id.* at 272-275.
[332] *Id.* at 272.

examples of AI adjudication successfully exhibiting the flaws inherent with human judgment.[333] First, disillusionment would, "erode confidence in the legal system's legitimacy."[334] Second, AI adjudication could diminish the position of the judiciary and alter the judiciary's culture and composition.[335] As a consequence, judges would have diminished authority.[336] Finally, disillusionment could result in smaller but significant changes: 1) diminished political power; 2) lawyer's rhetoric becomes irrelevant; 3) diminished adversarial system and movement towards inquisition; and 4) erasure of human lawyers from the legal process.[337]

Subsection D posits that AI adjudication will cause alienation and cease participation in the legal system.[338] The article goes even one step further and imagines a future where the judicial system becomes fully automated with AI.[339] In addition, alienation could cause a decrease in public engagement where there would be an insufficient amount of public oversight (ex. jury participation).[340] The article closes Section III with hope for a "new equilibrium" between equitable justice and codified justice as AI adjudication is incorporated in the legal system.[341]

Section IV discusses four types of viable responses to the issues described above.[342] Subsection A suggests continued experimentation with changes in the legal system as a possible solution to these problems. However, it also recognizes the risks of experimentation where human lives are at risk.[343] In subsection B, the article explores the possibility of "coding equity" into the AI system.[344] The article argues that coding ethics into the system could be achieved if it is updated regularly and could respond faster to issues of

[333] *Id.* at 273.
[334] *Id.*
[335] *Id.*
[336] *Id.* at 274.
[337] *Id.*
[338] *Id.* at 275.
[339] *Id.*
[340] *Id.* at 276.
[341] *Id.* at 277.
[342] *Id.* at 278-288.
[343] *Id.* at 279.
[344] *Id.* at 280.

inequality faster than humans.[345] However, coding equity would be difficult because the concept of "equity" contains many nuances.[346] It ultimately comes to the conclusion that it is an, "ineffective stand-alone solution."[347] Subsection C, suggests the division of labor between humans and AI as an additional solution.[348] For example, human judges and AI could collaborate at specific stages of the legal process, providing extra human oversight in these situations.[349] Another possibility is to separate cases and, "apportion discrete types of judicial decision-making to human[s]."[350] Subsection D posits the removal of profit-seeking actors as a way to keep the system focused on justice.[351]

Section IV concludes with the solution that concerns of AI adjudication should be addressed by drawing from all four types of responses.[352] The last section, section V summarizes and briefly notes the far-reaching consequences for AI adjudication on "executive bureaucracy and administrative agencies" where the same issues regarding codified justice would be reproduced.[353]

Hin-Yan Liu, *Three Types of Structural Discrimination Introduced by Autonomous Vehicles*, 51 UC DAVIS L. REV. ONLINE 149 (2017-2018).[354]

This article centers around crash-optimization algorithms used in autonomous vehicles. The authors discuss the ways in which this crash-optimization uses discriminatory systems and ends on an exploration of the impacts autonomous vehicles will have on the structure of future society. The article is organized into five parts: I) Prioritizing the Occupant in

[345] *Id.*
[346] *Id.* at 281
[347] *Id.*
[348] *Id.*
[349] *Id.*
[350] *Id.* at 283.
[351] *Id.* at 285.
[352] *Id.* at 288.
[353] *Id.* at 289.
[354] Hin-Yan Liu, *Three Types of Structural Discrimination Introduced by Autonomous Vehicles*, 51 UC DAVIS L. REV. ONLINE 149 (2017-2018).

Autonomous Vehicles through Trolley-Problem Scenarios; II) Structural Biases in Crash Optimization and Trolley-Problem Ethics; III) Intentional Discrimination and the Immunity Devise Thought-Experiment; IV) Structural Discrimination in the Corporate Profit-Driven Context; and V) Structural Discrimination in Urban Design and Law Revisited.

In the first section, the author explains that autonomous vehicles utilize crash-optimization algorithms to reduce overall damage of an unavoidable crash.[355] This happens by establishing probabilistic courses of action. Many equate this algorithm to the classic ethic hypothetical—the trolley-problem scenario, where the conductor has to decide whether to kill one person or five people by diverting the railroad tracks.[356] By contrasting the programming of autonomous cars from this ethics hypothetical, the author asserts three ways in which the crash-optimization program creates a discriminatory and problematic machine. Unlike the trolley driver, the decision-maker in self-driving cars is the manufacturer or the occupants of the vehicle.[357] They are not disinterested decision-makers.[358] For the manufacturer the stake in the outcome of the crash-optimization program is exclusively serving the interest of customers and for the occupant, their interest is their own personal safety. As a result, the primary focus of crash-optimization programs is to privilege their occupants over pedestrians and other third parties.[359] The author suggests that changing the perspective is a way to democratize the algorithm.[360]

The second way in the crash-optimizing system discriminates is the way in which it collects data and learns different outcomes. The author argues that the system focuses on a very structured and isolated patterns of data;[361] focusing on a single scenario overlooks a wider range of externalities.[362] The effect would only multiply when the same automobiles are distributed.

[355] *Id.* at 151-152.
[356] *Id.* at 154.
[357] *Id.* 154.
[358] *Id.* 154.
[359] *Id.* at 155.
[360] *Id.* at 156.
[361] *Id.* at 156-57.
[362] *Id.* at 158.

A related concern is that currently self-driving cars identify human beings just as units.[363] Not by characteristics or physical features, such as race or gender. However, there would be notable consequences if the algorithm begins using identifying characteristics to determine how to reduce damage. For example, if the vehicle is programmed to hit motorcyclists with helmets because her odds of surviving are higher, then, "certain groups will consistently bear a greater burden despite the fact that they adopted prudential measures to minimize their risk."[364] The article continues to argue that a system which makes a small number of biased decisions can unintentionally result in, " a systemic and collective dimension whereby the generated outcomes will be reliably and systematically skewed."[365] Another issue is if these small biased decisions do not reflect enough discriminatory intent for legal recognition.[366] The article notes that such discrimination while plainly discrimination "falls outside of the scope" of Article 26 of ICCPR.[367]

Section III considers what would happen if discrimination was intentional. The article posits that this requires the aim of crash-optimization systems to, "maximize the collective well-being of society."[368] For example, if the criteria for crash optimization was based on positive traits of an individual like talents, cultured ability, or potential, the system could make decisions based on saving the lives of scientific and cultural elite.[369] Many other types of preferences could be used to make these calculations: age, sex, race, or social/political status.[370] Section III continues to consider hypothetical future scenarios where the manufacture of an "immunity device" would allow its carriers to become completely immune to self-driving auto collisions.[371] The article also imagines customer loyalty programs which provide people with additional security or safety.[372]

[363] *Id.* at 159-160.
[364] *Id.* at 160.
[365] *Id.* at 161.
[366] *Id.* 161.
[367] *Id.* at 162.
[368] *Id.* at 163.
[369] *Id.* 163.
[370] *Id.* 163.
[371] *Id.* at 164.
[372] *Id.* at 166.

Section IV examines the structural discrimination that derives from profit-driven motives. The article contends that it would be difficult to apply law and cause trouble in the future because self-driving car manufacturers could use, "human beings as moral crumple zones" that absorb legal liability for structural discrimination.[373]

Section V expands the scope of the article, arguing that these issues with self-driving vehicles will also lead to changes in urban design.[374] The section also argues that there are currently no incentives to create liability structures.[375] The article speculates that normalizing the use of self-driving cars could cause deeper segregation between those with the privilege of wealth and access to technology and those who do not.[376] Moreover, forms of "architectural exclusion" could be used to exacerbate inequality. For example, when highways/bridges were first built over parkways, they were deliberately made low so that public transportation could not travel using the parkway, effectively keeping poor people off the road and within specific neighborhoods.[377] The article is concerned that similar structures will occur when infrastructure is designed to support self-driving vehicles.[378] Two foreseeable consequences could be the complete removal of human beings as operators of transportation and continued, "privatization of public space."[379]

In the conclusion, the article asks for a vigilant approach to all future development of self-driving automobile programs in order to avoid dystopian scenarios.[380] The article suggests the, "broadest range of participation in the design and development of these systems."[381]

[373] *Id.* at 171.
[374] *Id.* at 172.
[375] *Id.*
[376] *Id.*
[377] *Id.* at 173.
[378] *Id.* at 175.
[379] *Id.*
[380] *Id.* at 178.
[381] *Id.*

Joanna J. Bryson, *Robots Should Be Slaves* (University of Bath, 2009).[382]

> This article is made up of 6 sections and essentially argues, as the title suggests, that robots should be considered for all intents and purposes as slaves, not as companions or humans.[383] The author invites the reader to deeply consider how we think about robots and our relationship to this technology.[384]
>
> In the first section after the introduction, "Why *slaves*?," the article notes that slaves are defined as "people you own."[385] The author acknowledges the cruel and horrible history of slavery, particularly the dehumanizing consequences of slavery.[386] However, the author argues that, "dehumanization is only wrong when it's applied to someone who really is human[.]"[387]The article makes four fundamental claims: "1) Having servants is good and useful, provided no one is dehumanized; 2) A robot can be a servant without being a person; 3) It is right and natural for people to own robots; 4) It would be wrong to let people think that their robots are persons."[388]
>
> In the second section, "Why we get the metaphor wrong," the author pointedly states from the outset that there is no question that humans own robots, and it would be a mistake to ignore that robots are in our service.[389] Robots do not exist unless humans decide to create them, and we program and design their intelligence and behavior.[390] In the remainder of this section, the paper examines the tendency for roboticist and science fiction movies/books to express a human's ethical obligation to robots.[391] Citing a previously published article, the author explains that this tendency is a result

[382] Joanna J. Bryson, *Robots Should Be Slaves* (University of Bath, 2009), https://pdfs.semanticscholar.org/5b9f/4b2a2e28a74669df3789f6701aaed58a43d5.pdf.
[383] *Id.* at 1.
[384] *Id.* at 2.
[385] *Id.*
[386] *Id.*
[387] *Id.*
[388] *Id.* at 3.
[389] *Id.*
[390] *Id.*
[391] *Id.*

of "uncertainty about human identity," and the "arbitrary assignments of empathy."[392]

In the third section, "Costs and benefits of mis-identification with AI," the author essentially argues that over interaction with AI is an inefficient use of human time and resources.[393] The section provides measures of the cost of over-identification on the individual level: "1) the absolute amount of time and other resources an individual will allocate to a virtual companion; 2) what other endeavors that individual sacrifices to make that allocation; and 3) whether the tradeoff in benefits the individual derives from their engagement with the AI outweigh the costs or benefits to both that individual and anyone else who might have been affected by the neglected alternative endeavors."[394] The article argues that individuals have a, "finite amount of time and attention for forming social relationships," and humans increasingly seek superficial relationships from, "lower-risk, faux-social activities such as radio, television and interactive computer games."[395] At an institutional level, the article examines the larger implications of when AI makes decisions for humans.[396] In addition, it is dangerous to put moral responsibility unto robots instead of humans. As the author states, "we should never be talking about machines making ethical decisions, but rather machines operated correctly within the limits we set for them."[397] Ultimately, the author argues that misidentification with AI leads to "less responsible and productive members of society."[398] Automation allows humans to choose less "fulfilling social interactions with a robot over those with a human, just because robotic interactions are more predictable and less risky."[399]

The fourth section, "Getting the metaphor right," examines the ways that

[392] *Id.* at 4.
[393] *Id.* at 5.
[394] *Id.*
[395] *Id.* at 5-6.
[396] *Id.* at 6.
[397] *Id.*
[398] *Id.* at 7.
[399] *Id.*

robots can be useful for human society.[400] The author argues that understanding the "robot-as-slave" is the best way to, "get full utility … and … to avoid the moral hazards" of these robots.[401] The section posits that domestic robots will be used as physical support for the infirm, assisting those with working memory challenges, and as tutors for children.[402]

In the fifth section, "Don't we owe robots anything?," the article addresses concern that robots could be exploited and abused.[403] The article states that because humans determine robots' goals and desires, "it cannot mind being frustrated unless we program it to perceive frustration as distressing, rather than as an indication of a planning puzzle."[404] The author goes even further to say that robots should be absolutely replaceable, and no one should ever have to question whether to save a person or a robot from a burning building.[405]

In the conclusion, the author reiterates that robots should be viewed as tools that can enhance our own abilities,[406] and that if humans have any obligations, it is to society and not to robots.[407]

Jason Borenstein and Ron Arkin, *Robotic Nudges: The Ethics of Engineering a More Socially Just Human Being*, SCI ENG ETHICS (2016).[408]

The central question of this article is whether it would be ethical to construct companion robots that nudge a human user to behave in a certain way.[409] The author notes that the robotics community is working towards building robots that can function as lifelong companions for human beings.[410]

[400] *Id.* at 8.
[401] *Id.*
[402] *Id.*
[403] *Id.* at 9.
[404] *Id.*
[405] *Id.* at 10.
[406] *Id.*
[407] *Id.* at 11.
[408] Jason Borenstein and Ron Arkin, *Robotic Nudges: The Ethics of Engineering a More Socially Just Human Being*, SCI. ENG'G ETHICS (2016).
[409] *Id.* at 32.
[410] *Id.*

Consequently, the article briefly mentions how the fields of cinema, psychology, and marketing study the factors behind influencing human beings.[411] Robotics also gather data from these same sources to understand how robots can influence human behavior.[412] Citing the work of Thaler and Sunstein, the authors define "nudge" as a way to, "shape behavior without resorting to legal or regulatory means," and is often a subtle method.[413] Some contemporary examples of nudging that Thaler and Sunstein use are Texas' state motto, "Don't Mess with Texas," which creates feelings of a shared group identity and apparently decreased pollution in the state as a result.[414] Another example is how ATMs are programmed to return the debit card to the user before dispensing cash, thereby decreasing the chances of losing the card.[415]

The article contends that there are two types of paternalism which could serve as a justification for robotic nudges.[416] First, there is weak or soft paternalism which prevents harm in situation where, "it is presumed that if a person had additional knowledge or was mentally competent, the person would make a different decision."[417] Then, there is strong or hard paternalism which protects a person, "even if it goes against that person's voluntary choice."[418] The article again cites Thaler and Sunstein, who advocate for "libertarian paternalism," which upholds individual autonomy while still moving towards more "productive ends."[419] While robots are being developed and used for a variety of uses, including warfare, security, and healthcare, the article posits that robots could also be used to bring out positive traits from their users through verbal cues, proxemics, or touch.[420]

The article argues that a well-designed robot would have "distinct advantages" over other types of technology when influencing human

[411] *Id.*
[412] *Id.*
[413] *Id.* at 33.
[414] *Id.*
[415] *Id.*
[416] *Id.* at 34.
[417] *Id.*
[418] *Id.*
[419] *Id.*
[420] *Id.* at 35.

behavior.[421] Unlike phone apps, robots have a physical and therefore, stronger presence in the world, as well as the capacity to move around, and robots have a wider range of possibilities to mold their environment.[422] A well-designed robot would have to be a sophisticated machine which could discern between human beings and the various human behaviors it is supposed to monitor.[423] If society can agree that robotic nudging is ethically acceptable "when the intent is to promote a person's own well-being," the article asks under which conditions this would be permissible.[424] Another question to consider is whether nudging should be used to benefit the individual or a larger group.[425] For example, a robot could tap a parent on the shoulder when the user's child has been sitting alone watching television for an extended amount of time. While the child's welfare is the primary concern, the article notes that the parent could feel startled by this tap or even feel offended for the suggestion that the adult is a bad parent.[426]

The article briefly distinguishes between positive and negative nudges.[427] Positive nudges utilize positive reinforcement methods such as encouragement or rewards; negative nudges would use punishment or expressions of disappointment.[428] Furthermore, psychological and sociological data should inform the design of robotic programming in addition to ethical considerations.[429] If the robot is programmed to use abrasive or sudden tactics to deter human behavior, there is a strong likelihood that the human would see this as an intrusion and become angry instead of change their behavior.[430]

The article next examines some objections to robotic nudging.[431] First, the deliberate manipulation of a free-thinking individual is seen as an intrusion

[421] *Id.*
[422] *Id.* at 36.
[423] *Id.*
[424] *Id.* at 37.
[425] *Id.*
[426] *Id.*
[427] *Id.* at 38.
[428] *Id.*
[429] *Id.*
[430] *Id.*
[431] *Id.*

upon human liberty.[432] Second, there are concerns that nudging could be misused and result in "a wide range of abuses."[433] There is also the issue of "moral paternalism." Citing Harris, the article defines moral paternalism as the protection from corruption or wickedness.[434] Critics argue that this is "tampering with personal identity."[435] The same concern arises in biomedical technology where a robotic nudge could change human nature.[436]

The article then asks, "which framework or theory should be used as a basis or foundation for defining what ethical means?"[437] Even if only Western theories were examined, they would include such possibilities as: "rights-based approaches, deontology, consequentialism, virtue ethics, [and] cultural relativism."[438] The article directs its focus on what it considers to be the most valuable virtue—justice.[439] Then it comes to the conclusion that it would be best for robot nudges to promote social justice.[440] The article relies on two Rawlsian concepts of justice: 1) "each person is to have an equal right to the most extensive basic liberty compatible with a similar liberty for others"; and 2) the inequalities of society must be addressed with compensation that benefits everyone.[441]

The article proceeds to describe three "design pathways related to how much control a user could exert over a robot's nudging behavior: 1) opt in, 2) opt out, 3) no way out."[442] The Opt In pathway allows for users to, "consciously and deliberately select their preferences."[443] This option considers the individual autonomy of the human user.[444] The Opt Out pathway allows the robot to perform a default function until the user makes

[432] *Id.*
[433] *Id.*
[434] *Id.* at 39.
[435] *Id.*
[436] *Id.*
[437] *Id.*
[438] *Id.*
[439] *Id.* at 40.
[440] *Id.* at 40-41.
[441] *Id.* at 40.
[442] *Id.* at 42.
[443] *Id.*
[444] *Id.*

modification.[445] This pathway is likened to the automatic enrollment of employees into a retirement plan although they may choose to not participate in the plan.[446] The article notes that there is a concern with "subordination to technology" because humans will tend to agree with the default without taking the time to fully explore other available options.[447] The last pathway, the No Way Out pathway does not provide the user with the ability to turn off the robot. In other words, "justice trumps the individual user's autonomy and rights."[448] This is compared to the inability for smart phone users to turn off GPS tracking when the police use this option.[449]

In the final section, the article considers the robot designer's moral obligations in programming the technology.[450] A question that must be considered is, "does the foremost obligation that a robot possesses belong to its owner or to human society overall."[451] This article ultimately is concerned with "highlight[ing] ethical complexities" of robotic nudging rather than provide precise answers.[452]

Angela Daly et. al., *Artificial Intelligence Governance and Ethics: Global Perspectives*, The Chinese University of Hong Kong Research Paper Series (2019).[453]

This article is made up of 8 sections and provides an overview of international efforts to develop AI policies. Section 1: Introduction; Section 2: Global Level; Section 3: Europe ; Section 4: India; Section 5: China; Section 6: The United States of America; Section 7: Australia; and Section 8: Reflections, issues and next steps. The article only lists and provides brief explanations of any policies made by each nation's government.

[445] *Id.*
[446] *Id.*
[447] *Id.*
[448] *Id.* at 43.
[449] *Id.*
[450] *Id.*
[451] *Id.*
[452] *Id.*
[453] Angela Daly et. al., *Artificial Intelligence Governance and Ethics: Global Perspectives*, The Chinese University of Hong Kong Research Paper Series (2019).

In Section 1, the Introduction examines the definition of AI,[454] then explores the intersection between AI and ethics.[455] It is comprised of four subsections: 1) What is AI?; 2) AI and Ethics; 3) What does 'ethics' mean in AI?; and 4) This Report. The central issues the report is trying to figure out are, "what are the ethical standards to which AI should adhere,"[456] as well as which actors should be responsible for setting the legal and ethical standards.[457] One concern is that regulation will be established by private companies rather than government agencies.[458] The article next defines morality as, "a reflection theory of morality or as the theory of the good life."[459] AI ethics is understood as dynamic and interdisciplinary which must meet two traits to be effective: 1) AI should utilize "weak normativity" and cannot, "universally determine what is right and what is wrong"; and 2) "AI ethics should seek close proximity to its designated object."[460]

In section 2, Global Level, the article notes that, "the most prominent AI ethics guidelines" are the OECD Principles on AI.[461] These principles have been adopted by 36 Member states including the U.S., and six non-member states: Argentina, Brazil, Colombia, Costa Rica, Peru and Romania.[462] The 40th International Conference of Data Protection & Privacy Commissioners (ICDPPC) created the Declaration on Ethics and Data Protection in Artificial Intelligence in 2018.[463] The Commissioners also established a permanent working group on Ethics and Data Protection in Artificial Intelligence.[464] Under the subsection, "Technical initiatives," the article explains that the Institute of Electrical and Electronic Engineers (IEEE) has produced *Ethically Aligned Design,* which includes, "five General Principles to guide the ethical design, development and implementation of

[454] *Id.* at 6.
[455] *Id.* at 7-8.
[456] *Id.*
[457] *Id.*
[458] *Id.*
[459] *Id.*
[460] *Id.* at 8.
[461] *Id.* at 10.
[462] *Id.*
[463] *Id.*
[464] *Id.*

autonomous and intelligent systems."[465] Multinational corporations, such as Amazon, BBC, and Baidu have also developed their own statements.[466] The World Economic Forum (WEF) released a White Paper about AI governance.[467]

Section 3 focuses on AI policies in Europe.[468] The Section is divided into five subsections: European Union, Council of Europe, Germany, Austria, and the United Kingdom. Subsection 1, the article notes that the EU has positioned itself as, "a frontrunner in the global debate on AI governance and ethics."[469] In 2018, the General Data Protection Regulation (GDPR) was put into legislation.[470] This article highlights Section 5 of the GDPR on the Right to Object (Article 21) and Automated Individual Decision-Making Including Profiling (Article 22) as particularly significant elements of the GDPR.[471]

The article next mentions the European Parliament Resolution on Civil Law Rules on Robotics which was published in 2017. Significantly, the article highlights how the Resolution wanted the existing legal framework to be supplemented with, "guiding ethical principles in line with the complexity of robotics and its many social, medical and bioethical implications."[472] The Annex to the Resolution includes a proposed Code of Ethical Conduct for Robotics Engineers, Code for Research Ethics Committees, License for Designers and License for Users.[473]

Next, the European Commission issued a Communication on Artificial Intelligence for Europe in 2018, with three goals: 1) boosting the EU's technological and industrial capacity; 2) preparing for the labor, social security and educational socio-economic changes brought by increased use of AI; and 3) establishing an effective ethics and legal framework.[474]

[465] *Id.* at 11.

[466] *Id.*

[467] *Id.*

[468] *Id.* at 12.

[469] *Id.*

[470] *Id.*

[471] *Id.*

[472] *Id.*

[473] *Id.*

[474] *Id.*

Also in 2018, the European Group on Ethics in Science and New Technologies released a *Statement on Artificial Intelligence, Robotics and Autonomous Systems*.[475] This Statement suggested basic principles based on, "fundamental values laid down in the EU Treaties and in the EU Charter of Fundamental Rights."[476] The European Union High-Level Expert Group on Artificial Intelligence ("High-Level Expert Group"), which the article describes as a, "multi-stakeholder group of 52 experts from academia, civil society and industry," created its *Ethics Guidelines for Trustworthy AI in 2019*.[477] These Guidelines establish requirements that determine whether or not AI is "trustworthy."[478] These guidelines are currently being implemented in a pilot program across public and private sectors.[479] Thomas Metzinger, a member of this group, criticized this process as ”ethics washing" because certain non-negotiable clauses were removed from the Guidelines, and he calls for AI governance to be separated from industry.[480] The High-Level Expert Group later put out the *Policy and Investment Recommendations for Trustworthy AI*, with 33 recommendations for sustainable inclusive development of AI.[481] These recommendations also condemn the use of AI for State and corporate mass surveillance.[482] The Panel's puts particular focus on, "the monitoring and restriction of automated lethal weapons; the monitoring of personalized AI systems built on children's profiles; and the monitoring of AI systems used in the private sectors which significantly impact on human lives, with the possibility of introducing further obligations on such providers."[483]

Subsection 2 discusses the governance initiatives of the Council of Europe (COE) which includes all EU Member States and some non-EU states like eastern European states, Turkey and Russia.[484] The European Commission

[475] *Id.* at 13.
[476] *Id.*
[477] *Id.*
[478] *Id.*
[479] *Id.*
[480] *Id.*
[481] *Id.*
[482] *Id.*
[483] *Id.* at 13-14.
[484] *Id.* at 14.

for the Efficiency of Justice created the European Ethical Charter in 2018, which sets forth five principles for the development of AI usage in the European judiciary.[485] The COE has also published the Guidelines on Artificial Intelligence and Data Protection in 2019.[486]

Subsection 3 centers on Germany.[487] The article notes that the country has invested close to 3 Billion Euros into AI research.[488] While the nation is small and cannot compete with larger nations, Germany has competitively branded itself as supportive of, "data protection-friendly, trustworthy, and 'human centered' AI systems, which are supposed to be used for the common good…".[489] Part of the German government's strategy is to fund research and innovation as well as create 100 new professorships in the study of "AI and Machine Learning."[490]

Subsection 4 briefly focuses on Austria.[491] The government drafted a report, 'Artificial Intelligence Mission Austria 2030,' which lists numerous stakeholders and participation methods.[492] Austria has also indicated the desire to create a "large national data pool" where the personal data of Austrian citizens, "would be sold to the highest bidder in order to attract cutting edge data-driven research to Austria."[493]

Subsection 5 focuses on the United Kingdom.[494] In 2018, the UK introduced the AI Sector Deal in order to place, "the UK at the forefront of the artificial intelligence and data revolution."[495] The UK Parliament has made various initiatives to address AI governance and issues, including an All-Party Parliamentary Group on AI and a Select Committee on AI.[496] The latter

[485] *Id.*
[486] *Id.*
[487] *Id.*
[488] *Id.* at 15.
[489] *Id.*
[490] *Id.*
[491] *Id.*
[492] *Id.*
[493] *Id.* at 16.
[494] *Id.*
[495] *Id.* at 16-17.
[496] *Id.*

committee studies whether the current legal and regulatory frameworks should be adapted to meet the needs of an AI future.[497] The UK government also partnered with the WEF's Center for the Fourth Industrial Revolution to design guidelines for public sector usage of AI.[498] This subsection ends with some concern for future AI development initiatives in the face of Brexit, as financing and research will be rescinded or implausible.[499]

Section 4 concisely focuses on India.[500] The article notes that three national initiatives have been implemented by the Indian government: 1) Digital India, "which aims to make India a digitally empowered knowledge economy."; 2) Make in India, which focuses on making India the designer and developer of AI technology; and 3) The Smart Cities Mission.[501] The Ministry of Commerce and Industry formed an AI Task Force and reported that AI should be incorporated into, "national security, financial technology, manufacturing and agriculture."[502] The article provides some criticisms of India's AI governance. There is currently not data protection legislation in place or other ethical framework to address personal data concerns.[503] Furthermore, suggested ethics guidelines do not, "meaningfully engage with issues of fundamental rights, fairness, inclusion, and the limits of data driven decision making."[504]

Section 5 concentrates on governance efforts made by China.[505] In 2017, The New-Generation AI Development Plan called for high investment in AI development and to create new regulations and ethical policies by 2025.[506] In 2019, the Beijing Academy of Artificial Intelligence released the *Beijing AI Principles*. These principles considered: "1) the risk of human unemployment by encouraging more research on Human-AI coordination; 2) avoiding the negative implications of 'malicious AI race' by promoting

[497] *Id.*
[498] *Id.*
[499] *Id.* at 17-18.
[500] *Id.* at 19.
[501] *Id.*
[502] *Id.*
[503] *Id.*
[504] *Id.*
[505] *Id.* at 20.
[506] *Id.*

cooperation, also on a global level; 3) integrating AI policy with its rapid development in a dynamic and responsive way by making special guidelines across sectors; and 4) continuously making preventive and forecasting policy in a long-term perspective with respect to risks posed by Augmented Intelligence, Artificial General Intelligence (AGI) and Superintelligence."[507] Top Chinese Universities, companies, and the Artificial Intelligence Industry Alliance (AIIA), released a Joint Pledge on Self Discipline in the Artificial Intelligence Industry.[508] The article points out that while this Pledge is similar to many other AI governance statements, it does distinguish itself by including language of "secure/safe and controllable" and "self-discipline" as important aspects that need to be integrated into AI governance.[509] Lastly, the Chinese Government Ministry of Science and Technology released its eight Governance Principles for the New Generation Artificial Intelligence.[510] The Principles advocate for international collaboration as well as the concept of "agile governance."[511] This concept addresses the rapidly progressing nature of AI technology and the need for legislation to be dynamic in resolving issues.[512]

Section 6 concentrates on the United States of America.[513] Last year, the U.S. implemented the *Executive Order on Maintaining American Leadership in Artificial Intelligence*.[514] This Order has created the American AI Initiative, organized by the National Science and Technology Council (NSTC) Select Committee on Artificial Intelligence.[515] The Order include the protection of "civil liberties, privacy and American values," and the creation of lucrative foreign markets for American-made AI.[516] There are six strategic objectives that must be met by regulators and developers, including the protection of, "American technology, economic and national

[507] *Id.*
[508] *Id.*
[509] *Id.* at 21.
[510] *Id.*
[511] *Id.*
[512] *Id.*
[513] *Id.* at 23.
[514] *Id.*
[515] *Id.*
[516] *Id.*

security, civil liberties, privacy, and values."[517] In addition, AI developers must prioritize national security and public trust and protect AI technology from foreign attacks.[518] In 2019, the US Department of Defense initiated its AI Strategy to build lawful and ethical military technology to remain competitive with China and Russia.[519] Not-for-profit organizations, like The Future of Life Institute, issued 23 *Asilomar AI Principles*.[520] OpenAI released its open AI charter with the goal that artificial general intelligence outperforms humans for the benefit of all humanity.[521]

Section 7 focuses on Australia and offers some criticism.[522] The Australian Human Rights Commission started the Technology and Human Rights Project.[523] The Australian Government Department of Industry, Innovation and Science released a paper on Australia's efforts to develop an ethics framework.[524] The authors argue that the Australian Ethical Framework report developed by Data 61 and CSIRO lacks a fundamental misunderstanding of Australian privacy law (citing Johnson 2019).[525] It also suggests, "a very narrow understanding of the negative impacts of AI," and fails to see the full impact of harms AI can have.[526] The report does not offer responsive regulatory approaches for automated decision making.[527] The Report sets out eight key principles: "Generates net-benefits; do no harm; regulatory and legal compliance; privacy protection; fairness; transparency and explainability; and contestability[.]"[528] The proposed Australian AI ethical framework also provides a 'toolkit' for implementation.[529] Some of the strategies include: "impact assessments; internal/external review; risk assessments; best practice guidelines; industry standards; collaboration; mechanisms for monitoring and improvement; recourse mechanisms; and

[517] *Id.*
[518] *Id.*
[519] *Id.* at 24.
[520] *Id.*
[521] *Id.*
[522] *Id.* at 26.
[523] *Id.*
[524] *Id.*
[525] *Id.*
[526] *Id.*
[527] *Id.* at 27.
[528] *Id.*
[529] *Id.*

consultation."[530]

Lastly, Section 8 provides some of the authors' reflections and analysis of the information gathered in the previous sections.[531] This section is divided into 7 smaller subsections.[532] First, this section notes that there is clearly a distinction between the "haves" and "have-nots" in terms of the resources and capacity to implement advanced governing systems for AI.[533] The article notes that the EU and China are the top groups, but the U.S. could quickly become a strong competitor.[534] Second, there is a fundamental motivation to compete between countries despite the stated need for international collaboration.[535] The U.S., China and the EU have all stated the desire to become global leaders and to perpetuate nationalist values.[536] This is also evidenced in the fact that smaller countries like Austria are, "willing to engage in less ethical projects to attract attention and investment."[537] Third, the authors note that many of the AI governance statements contain many similar goals, such as accountability, transparency, privacy, and protection.[538] However, the authors also note that underneath these shared goals, there may be varying, "cultural, legal and philosophical understandings."[539] Fourth, the authors consider "what's not included" in these discussions.[540] Some questions that need to be asked are: 1) Is there reference to other government or corporate initiatives which may contradict the principles?; 2) What are the hidden costs of AI; and 3) How are they visible and internalized?[541] Fifth, "What's already there?"[542] The authors assert the need to better understand the interactions between policy, rights, private law, and AI ethics.[543] Sixth, the authors raise the issue of "ethics

[530] *Id.*
[531] *Id.* at 28.
[532] *Id.* at 28-32.
[533] *Id.* at 28.
[534] *Id.*
[535] *Id.*
[536] *Id.*
[537] *Id.* at 29.
[538] *Id.*
[539] *Id.*
[540] *Id.*
[541] *Id.*
[542] *Id.*
[543] *Id.* at 30.

washing."[544] If these AI governance initiatives are not enforced, then all of the aforementioned research papers, committees, and strategies only serves as "window dressing."[545] This section also raises issues of "jurisdiction shopping" for locations with less restrictive AI regulations.[546] The authors note that it may be significant to take a historical perspective on implementation because there are, "different predecessor technologies … as well as different social, economic and political conditions," that each country starts with.[547] Lastly, the authors ask, "who is involved?"[548] It is imperative that all participating voices are heard and that the larger public is appropriately and accurately represented in discussions about AI implementation and governance.[549]

Mark Latonero, *Governing Artificial Intelligence: Upholding Human Rights & Dignity, Data & Society* (2018).[550]

This report is divided into five sections: 1) Introduction; 2) Bridging AI and Human Rights; 3) A Human Rights Frame For AI Risks and Harms; 4) Stakeholder Overview; and 5) Conclusion. The introduction lays out the reports underlying belief that if the purpose of AI is to, "benefit the common good, at the very least its design and deployment should avoid harms to fundamental human values."[551] The author states that rooting AI development into a rights-based approach provides an, "aspirational and normative guidance to uphold human dignity and the inherent worth of every individual, regardless of country or jurisdiction."[552]

Section 2, Bridging AI and Human Rights, acknowledges that AI is a vast, multi-disciplinary field.[553] Therefore, the section explains, "the basic entry

[544] *Id.*
[545] *Id.*
[546] *Id.* at 31.
[547] *Id.*
[548] *Id.* at 31-32.
[549] *Id.* at 32.
[550] Mark Latonero, *Governing Artificial Intelligence: Upholding Human Rights & Dignity,* Data & Society (Oct. 10, 2018), https://datasociety.net/library/governing-artificial-intelligence/.
[551] *Id.* at 5.
[552] *Id.* at 5-6.
[553] *Id.* at 7.

points" between human rights and AI.[554] The article explains that current AI technology uses machine learning systems.[555] Machine learning processes historical data to detect patterns, but if this data is skewed or incomplete, biases can quickly perpetuate throughout the AI system.[556] For example, facial recognition technologies, "reproduce culturally engrained biases against people of color," when discriminatory algorithms cannot properly process or recognize darker skinned people.[557]

Section 3, A Human Rights Frame For AI Risks And Harms, is divided into five smaller subsections.[558] The purpose of this section focuses on five areas of human rights: Nondiscrimination and Equality, Political Participation, Privacy, Freedom of Expression, as well as Disability Rights.[559] The section opens with the International Bill of Rights as the primary source of human rights, which is comprised of three treaties: 1) The International Covenant on Civil and Political Rights (ICCPR); 2) the International Covenant on Economic, Social and Cultural Rights (ICESRC); and the Universal Declaration on Human Rights (UDHR).[560]

The first subsection, Nondiscrimination and Equality, details uses of discriminatory algorithms in programs like the Allegheny Family Screening Tool (AFST), which is a predictive risk model that forecasts child abuse and neglect.[561] Studies have found that the AFST uses information about families that use public serves and more frequently targets poor residents and disproportionately places certain kinds of people into problematic categories.[562] In South Africa, the apartheid regime was held up by, "classification systems built on databases that sorted citizens by pseudoscientific racial taxonomies.[563] A report by the World Economic Forum voiced concerns that because the success of machine learning is

[554] *Id.*
[555] *Id.* at 8.
[556] *Id.*
[557] *Id.* at 9.
[558] *Id.* at 10-16.
[559] *Id.* at 10.
[560] *Id.*
[561] *Id.* at 11.
[562] *Id.*
[563] *Id.*

measured in efficiency and profit, these measures may overshadow responsibility to human rights.[564]

Under the second subsection, Political Participation, the author focuses on the ways that discriminatory AI can spread disinformation.[565] When citizens cannot be informed and misrepresentations are made to them about political campaigns and world events, this violates the right to self-determination and the right to equal participation under the ICCPR.[566]

The third subsection, Privacy, concentrates on the use of algorithmic surveillance by private companies, like Amazon, to gather and reveal personal data about users.[567] The article notes that the right to privacy is found in both the UDHR (Article 12) and ICCPR (Article 17). Furthermore, protecting the right to privacy, "is key to the enjoyment of a number of related rights, such as freedoms of expression, association, political participation, and information.[568]

In the fourth subsection, Freedom of Expression, the article largely focuses on the management of social media platforms.[569] Algorithms skew the users social media feed based on personal preferences and interest. Algorithms also remove negative posts or comments, and private companies have the ability to undermine or, "meaningfully determine the boundaries of speech."[570] The subsection notes the difficulty in finding the right balance between the "legal and social impact relative to multiple rights."[571]

The final subsection, The Disability Rights Approach and Accessible Design, encapsulates, "how technological developments increases the risk to vulnerable groups,"[572] as well as the difficulty in implementing change.[573] For example, Netflix did not comply with ADA guidelines until

[564] *Id.*
[565] *Id.* at 12.
[566] *Id.* at 13.
[567] *Id.*
[568] *Id.* at 14.
[569] *Id.*
[570] *Id.*
[571] *Id.* at 15.
[572] *Id.*
[573] *Id.* at 15-16.

disability rights groups advocated and pressured the company for years.[574] The article points out that human rights cannot be implemented without laws.[575] This requires additional incentives like public activism and market forces.[576] Moreover, human rights need to be, "infused into the workflow of the organization as part of the jobs of employees working on quality assurance, test suites, and product design documentation," not just corporate statements.[577]

In Section 4, Stakeholder Overview, the article provides a snapshot of AI and human rights initiatives in business, civil society, governmental organizations, the UN, intergovernmental organizations, and academia, and the section is divided into the aforementioned six subsections.[578]

Subsection 1, Business, briefly discusses some human rights initiatives by the companies Microsoft, Google, and Facebook.[579] Microsoft completed its first Human Rights Impact Assessment (HRIA) and created a, "methodology for the business sector that are used to examine the impact of a product or action from the viewpoint of the rights holders."[580] After backlash and petitions from Google employees, the company did not renew its contract with the US Department of Defense to develop AI weapons.[581] A similar situation occurred with Microsoft and facial recognition technology given to US Immigration and Customs Enforcement.[582]

Subsection 2, Civil Society, highlights the fact that AI is dominated by powerful, socio-economically stable countries, which makes it difficult for countries in the Global South to access technology.[583] The subsection notes that four civil society groups—Amnesty International, Global Symposium on Artificial Intelligence and Inclusion, The Digital Asia Hub, and the

[574] *Id.* at 16.
[575] *Id.*
[576] *Id.*
[577] *Id.*
[578] *Id.* at 17.
[579] *Id.* at 18-19.
[580] *Id.* at 18.
[581] *Id.*
[582] *Id.* at 19.
[583] *Id.* at 20.

WEF—have conducted studies to assess the impact of AI on inequality as well as the need to engage diverse groups in AI research and policy making.[584]

Subsection 3, Governments, briefly details the regulatory efforts that some national governments have made.[585] The European Union's General Data Protection Regulation has secured new guidelines for data protection and privacy.[586] Canada and France have called for, "an international study group that can become a global point of reference for understanding and sharing research results on artificial intelligence issues and best practices."[587] Both Global Affairs Canada's Digital Inclusion Lab and Canada's Treasury Board have conducted studies on AI's impact on human rights.[588] New York City has passed laws that secure the transparency and fair application of algorithms used by the City in order to prevent a biased system.[589] The UN has investigated, "the impact and responsibilities of tech companies to protect human rights,"[590] Including the issue of autonomous weapons and their impact on the conduct of war.[591]

In Subsection 4, Intergovernmental Organizations, the report only notes that the Organization for Economic Cooperation and Development (OECD) has prepared some guidance for its 36 member-countries.[592] This guidance system has also set up National Contact Points, where each nation appoints a representative to hear grievances related to company misconduct.[593]

Subsection 5, Academia, briefly details how academics at Harvard, University of Essex, and Stanford University, have reached the same conclusion that there is an urgent need for greater collaboration between technology and human rights fields, as well as other disciplines to build a

[584] *Id.*
[585] *Id.*
[586] *Id.*
[587] *Id.* at 21.
[588] *Id.*
[589] *Id.*
[590] *Id.*
[591] *Id.* at 22.
[592] *Id.*
[593] *Id.*

universal and effective framework.[594]

In its conclusion, the report makes additional recommendations: 1) tech companies need to make an effort to collaborate with local civil society groups and researchers; 2) all tech companies should implement HRIAs, "throughout the life of their AI systems."; 3) governments must acknowledge their responsibilities and obligation in protecting fundamental rights and formulate nationally implemented AI policies; 4) lawyers, sociologists, lawmakers, engineers need to work together to integrate human rights into all business and production models; 5) academics and legal scholars should continue to investigate and research the intersection between AI, human rights and ethics; and 6) the UN should continue to investigate and enforce human rights with participating governments and monitor rapidly changing AI technology.[595]

[594] *Id.* at 23.
[595] *Id.* at 25.