HARVARD Kennedy School
CARR CENTER
for Human Rights Policy

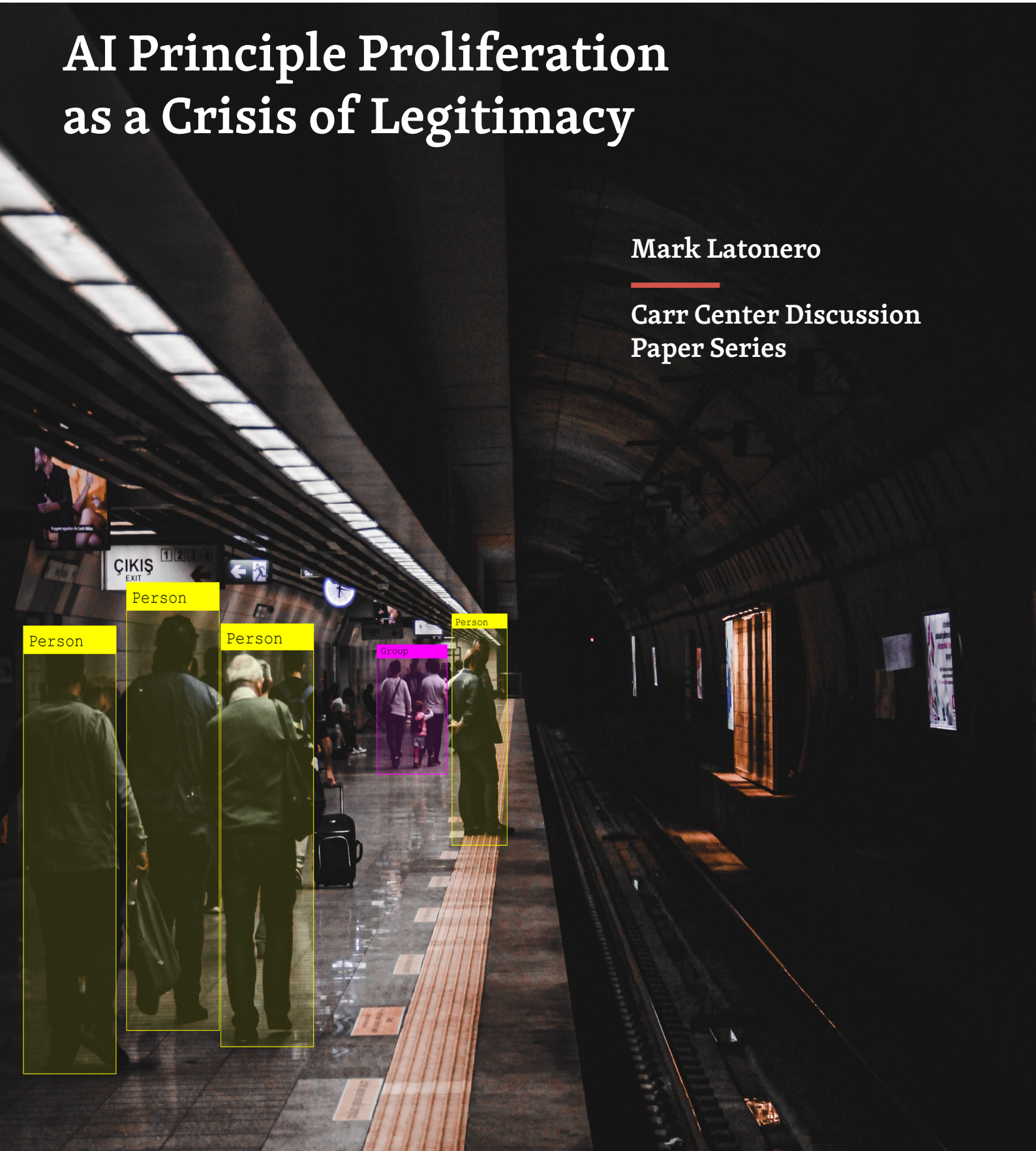
FALL 2020
ISSUE 2020 - 011

CARR CENTER FOR HUMAN RIGHTS POLICY
HARVARD KENNEDY SCHOOL

AI Principle Proliferation as a Crisis of Legitimacy

Mark Latonero

Carr Center Discussion Paper Series



AI Principle Proliferation as a Crisis of Legitimacy

Mark Latonero

Fellow, Harvard Kennedy School, Carr Center for Human Rights Policy

August, 2020

Mark Latonero is a fellow at Harvard Kennedy School, Carr Center for Human Rights Policy. He is also a senior associate at the Center for Strategic and International Studies, an affiliate at Data & Society, and serves as a consultant for the UN Office for the High Commissioner for Human Rights. Previously, he was a research professor at the University of Southern California where he led the Technology and Human Trafficking Initiative. His research examines the risks, harms, and benefits of emerging technologies, particularly for vulnerable populations in human rights and humanitarian contexts. Recent research publications include *Governing Artificial Intelligence: Upholding Human Rights & Dignity; Digital Identity in the Migration and Refugee Context; Refugee Connectivity: A Survey of Mobile Phones, Mental Health, and Privacy at a Syrian Refugee Camp in Greece*; as well as *The New York Times* op-ed "Stop Surveillance Humanitarianism." He has led field research in over a dozen countries, and his projects have been supported by the Open Society Foundations, Humanity United, USAID, UNICEF, and Microsoft, among others. Mark completed his PhD at USC's Annenberg School for Communication and was a postdoctoral scholar at the London School of Economics.

Table of Contents

Introduction 2

Crisis of Legitimacy 2

Human Rights 5

Catastrophe 6

Introduction

While Artificial Intelligence is a burgeoning field today, there is a growing concern about the mushrooming of proposed principles on how AI should be governed. Recent years have seen numerous AI principles emanating from academia, government, civil society, and private sector actors. To name but a few examples: In 2017, academics who gathered for a conference on "Beneficial AI" created the Asilomar AI principles that cover research, ethics and values, and long-term issues, which over 5,000 individuals have endorsed.[1] In 2018, Google released its AI Principles with its CEO saying, "if you don't have at least some stated principles, then you can't even figure out when you're breaking the rules."[2] In 2019, the European Commission's High-Level Expert Group on Artificial Intelligence outlined 7 requirements to realize Trustworthy AI.[3] And in 2020, researchers from Harvard's Berkman Klein Center (2020) analyzed a non-exhaustive set of 36 AI principles across sectors in an attempt to make sense of the space.[4]

Floridi and Cowls have described this state of affairs as "principle proliferation" – a problem that does more to confuse and overwhelm than provide clarity to the field.[5] This discussion paper agrees with the problem identified by Floridi and Cows and goes a step further. It explores the argument that principle proliferation signals a crisis of legitimacy that could have negative ramifications on a global scale. The paper posits that human rights could serve to stabilize AI governance, particularly if framed as an anchor to guide AI usage that could avert both everyday and catastrophic social harms.

Crisis of Legitimacy

The flourishing of principles can be seen as positive at first blush. If leaders in policy, business, and academia are rapidly iterating on AI principles and norms to live by, it may indicate they consider the global impact of AI as an urgent and serious concern. It may be that we need to let a hundred flowers bloom to unleash the plurality of ideas and perspectives needed to address the ethical quandaries raised by fast-moving advancements in AI. The cluster of technologies that comprise what we call AI is inextricably embedded in social and political dynamics such that no one profession is the expert and no one institution is the authority. This has led to exciting, cross-disciplinary work, evidenced by computer scientists, social scientists, legal scholars, and philosophers building knowledge at novel conferences, like the ACM's conference on Fairness, Accountability and Transparency in AI/Machine Learning.

However, principle proliferation can be an indicator of a more worrying situation. At a time of increasing anxiety about the social implications of AI, we know not where to turn for guidance on what principles to choose from and why we should follow them. The sheer number of principles makes us unsure who to trust, who sets the rules of the road, and who would enforce them.

AI, like other emergent technologies in the past, is assumed to have globally transformative qualities.[6] Yet AI narratives come with an important twist: We are as aware of the risks and harms of AI as its benefits. While AI theorists ponder the future of superintelligent AI replete with dire warning of

[1] "AI Principles." Future of Life Institute, 2017. https://futureoflife.org/ai-principles/?cn-reloaded=1.

[2] "Artificial Intelligence at Google: Our Principles." Google, 2018. https://ai.google/principles/.

[3] "Ethics Guidelines for Trustworthy Artificial Intelligence (AI)." European Commission High-Level Expert Group on Artificial Intelligence, 2019. http://ec.europa.eu/futurium/en/ethics-guidelines-trustworthy-ai/stakeholder-consultation-guidelines-first-draft#Top; The 7 elements are: Ensure that the development, deployment and use of AI systems meets the seven key requirements for Trustworthy AI: (1) human agency and oversight, (2) technical robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, non-discrimination and fairness, (6) environmental and societal well-being and (7) accountability.

[4] Fjeld, Jessica, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI." SSRN. February 14, 2020. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3518482; The report found 8 key themes in the 36 AI principles document: privacy, accountability, safety & security, transparency and explainability, fairness and non-discrimination, human control of technology, professional responsibility, and promotion of human values.

[5] Floridi, Luciano, and Josh Cowls. "A Unified Framework of Five Principles for AI in Society · Harvard Data Science Review." Harvard Data Science Review. June 14, 2019. Accessed June 19, 2020. https://hdsr.mitpress.mit.edu/pub/l0jsh9d1; It should be noted that the authors suggest 5 core principles for ethical AI: beneficence, non-maleficence, autonomy, justice, and explicability.

[6] For example, at the 1893 Word's Fair, Westinghouse luminated the city Chicago with Tesla's scientific achievements, which reportedly enthralled the global community.

existential risks to human kind,[7] AI's negative social impacts are already happening in the present. Pioneering researchers have demonstrated the ways automated systems are discriminating against the poor or people of color. [8]

Given the uncertainty, looking at metaphors for AI may be instructive since, as Watson (2016) argues, "[m]etaphors have always helped with introducing new technologies in our everyday lives and finding ways to familiarize ourselves with the novelty."[9] Given the alarming consequences, the emergence of AI may be like the development of nuclear energy wherein the promise of a vast energy supply is accompanied by radioactive waste, conflagration, and potential catastrophes like Chernobyl.

Stark and Hoffman discuss facial recognition as the "plutonium of AI" due to its inherent deleterious impacts:

> "[R]isks of these technologies vastly outweigh the benefits, in a way that's reminiscent of hazardous nuclear technologies. That is why the metaphor of plutonium is apt. Facial recognition, simply by being designed and built, is intrinsically socially toxic, regardless of the intentions of its makers; it needs controls so strict that it should be banned for almost all practical purposes."[10]

Such metaphors reflect the current reality where the use of AI systems is multiplying across social domains – from policing to hiring to health care – relying on unexplainable black-box processes that can scale social harms to millions of people faster than our institutions are equipped to respond or correct.[11] With risks such as these, AI principles would need to be accompanied with political legitimacy in order to regulate and constrain policies, practices, and behaviors. Political leaders cannot agree on what to do about AI harms. As with AI principles, we are awash in national AI strategies. There are over 30 different national strategies on AI as well as international strategies from groups like the EU and the G7. Many of these strategies are focused on the economic impact of AI technologies and the relative competitive advantage this may, or may not, bring rather than risk. The U.S. has historically led discussions on internet governance and, more recently, on balancing the opportunities of big data analytics with values like non-discrimination and privacy.[12] Yet today, the U.S. government is largely missing in pressing debates on AI governance. This adds to the sense that the institutions we have relied on to guide on technology and protect us from potential harms are failing.

Human rights could serve to stabilize AI governance, particularly if framed as an anchor to guide AI usage that could avert both everyday and catastrophic social harms.

The lack of U.S. policy on AI led to one of biggest tech companies in the world, Microsoft, essentially pleading for government regulation of facial recognition.[13] Corporate self-regulation indeed has its limits. Empirical research from Buolamwini and Gebru had already demonstrated that all commercial facial recognition technology from the major companies were biased and discriminatory towards people of color.[14]

---

[7] See Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. New York: Oxford University Press, 2017.

[8] See Buolamwini, Joy and Gebru, Timnit. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." Proceedings of ACM Conference on Fairness, Accountability, and Transparency, 81:1–15, 2018; Eubanks, Virginia. *Automating Inequality: How High-tech Tools Profile, Police, and Punish the Poor*. New York: Picador, St. Martins Press, 2019.

[9] Watson, Sarah M. "Data is the New '___.': on the Industrial Metaphors of Big Data" DIS Magazine, 2018. http://dismagazine.com/discussion/73298/sara-m-watson-metaphors-of-big-data/; also see Hwang, Tim and Levy, Karen. ""The Cloud' and Other Dangerous Metaphors: Contemporary ideas about data and privacy are tied up inextricably with language choices." The Atlantic, January 20, 2015 https://www.theatlantic.com/technology/archive/2015/01/the-cloud-and-other-dangerous-metaphors/384518/.

[10] Stark, Luke and Hoffmann, Anna Lauren. "Data Is the New What? Popular Metaphors & Professional Ethics in Emerging Data Culture," Journal of Cultural Analytics, 2019.

[11] See Noble, Safiya. *Algorithms Of Oppression: How Search Engines Reinforce Racism*. New York University Press, 2018; Pasquale, Frank. *The Black Box Society: the Secret Algorithms That Control Money and Information*. Harvard University Press, 2015

[12] See The White House. "Big Data: Seizing Opportunities Preserving Values Memo." February, 2015. https://obamawhitehouse.archives.gov/sites/default/files/docs/20150204_Big_Data_Seizing_Opportunities_Preserving_Values_Memo.pdf.

[13] Smith, Brad. "Facial recognition technology: The need for public regulation and corporate responsibility." Microsoft, July, 13, 2018. https://blogs.microsoft.com/on-the-issues/2018/07/13/facial-recognition-technology-the-need-for-public-regulation-and-corporate-responsibility/.

[14] Buolamwini and Gebru, 2018 ibid. Also see Snow, Jacob. "Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots." ACLU, July 26, 2018. https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28.

If human rights are not seen as relevant for AI governance, then its relevance is indeed in question. If not human rights, what would take its place?

Yet companies continued to sell technologies to law enforcement and other government agencies. It took widespread protests against the brutal killing of George Floyd in the hands of Minneapolis police on May 25, 2020 to motivate some of the major companies to change. In June 2020, IBM announced it would cease its facial recognition program entirely in response to the Black Lives Matter protest.[15] Soon after, Microsoft and Amazon announced they would suspend selling facial recognition products to the police.[16] Not all big tech companies have followed suit and if lesser known companies remain unencumbered they may fill the void in providing similarly flawed technologies that may exacerbate entrenched racial discrimination and undue violence by the state.

Some governmental regulation has emerged. For example, the city of San Francisco has banned facial recognition for government agencies like the police.[17] Other cities like Oakland and Sommerville have banned facial recognition use as well.[18]But the inherent problem with technology regulation at the local or national levels is that it protects only those within its jurisdiction. For example, a tech company may not be able to sell facial recognition technology to police in San Francisco but could do so for any other municipalities or to federal agencies.

Piecemeal regulation is no match for a technology that is poised to scale across the planet without regard to boundaries. The result we might see is global AI arbitrage, or forum shopping, where a buyer intent on using facial recognition to suppress peaceful protest simply purchases from a company willing to develop the technology in a country without a ban. This has led to calls for a global moratorium on facial recognition.[19]

Under what principles would be legitimate enough to guide a global set of actors towards value driven decision making in the AI space?

Human Rights

In the midst of principle proliferation is a resurgence of interest in human rights and the role it can play in global AI governance. I have written how the concerns about AI's social implications have brought about a search for a "north star" for the values that will guide its development towards a common good. International human rights provides a global formulation of those values.[20] Human rights gives us a language for communicating about social risks and harms grounded in rights like privacy, freedom of expression, assembly, non-discrimination, equality, and dignity.[21] If an AI system has the possibility of infringing on people's ability to exercise these and other fundamental rights, it constitutes a high-stakes domain that demands a governing framework to guide decisions about the development and deployment of that system. These global standards can be found in what is known as the international bill of rights: the Universal Declaration on Human Rights; the international treaty on Civil and Political Rights; and the treaty on Economic, Social, and Cultural Rights.[22] The large majority of the world's states have signed or ratified these treaties and have integrated human rights laws into their own national laws and constitutions.

Human rights is far from perfect and not without its critics.[23] As a legal and political framework, it is dependent on nation states that can ignore or dismiss their obligations

[15] Arvind Krishna. "IBM CEO's Letter to Congress on Racial Justice Reform." June 8, 2020. https://www.ibm.com/blogs/policy/facial-recognition-susset-racial-justice-reforms/.

[16] Matt O'Brien. "Microsoft joins Amazon, IBM in pausing face scans for police." Associated Press, June 11, 2020. https://apnews.com/e5dfcb8c0b003c1134137d33add4c301.

[17] Conger, Kate, Fausset, Richard and Kovaleski, Serge F. "San Francisco Bans Facial Recognition Technology." The New York Times, May 14, 2019. https://www.nytimes.com/2019/05/14/us/facial-recognition-ban-san-francisco.html.

[18] See Ordinance: Banning the usage of facial technology surveillance in Somerville (Massachusetts), 2019. http://somervillecityma.iqm2.com/Citizens/Detail_LegiFile.aspx?ID=20991&highlightTerms=facial%20recognition%20technology.

[19] Kind, Carly. "We need a moratorium on use of facial recognition technology." Financial Times, Sep 8, 2019. https://www.ft.com/content/6ecf7150-cfbd-11e9-b018-ca4456540ea6.

[20] Latonero, Mark. "Governing Artificial Intelligence: Upholding Human Rights and Dignity." Data & Society Research Institute, 2018. https://datasociety.net/wp-content/uploads/2018/10/DataSociety_Governing_Artificial_Intelligence_Upholding_Human_Rights.pdf.

[21] See Van Veen, Christian and Cath, Corinne. "Artificial Intelligence: What's Human Rights Got To Do With It?" Data & Society, Points, May 14, 2018. https://points.datasociety.net/artificial-intelligence-whats-human-rights-got-to-do-with-it-4622ec1566d5.

[22] Universal Declaration of Human Rights. https://www.ohchr.org/EN/UDHR/Pages/Language.aspx?LangID=eng; International Covenant on Civil and Political Rights. https://www.ohchr.org/en/professionalinterest/pages/ccpr.aspx; International Covenant on Economic, Social and Cultural Rights. https://www.ohchr.org/en/professionalinterest/pages/cescr.aspx.

[23] See, e.g., Moyn, Sam. *Not Enough: Human Rights in an Unequal World*. Belknap, 2017.

to protect individuals' rights. Still, as Sikkik argues, human rights carries international legitimacy that has proved effective, albeit unevenly, in directing policy and actions of governments in the past.[24] Human rights can be effective in the short term, like alleviating the immediate impact on victims of abuse, as well in the long term, like destabilizing authoritarian regimes.[25] And empirical research has found that prosecuting political leaders for past human rights violations can affect the behavior of political leaders globally.[26]

According to Donahue and Metzger, human rights fulfills the need for a "framework that can claim global buy-in and that addresses the roles and responsibilities for both government and the private sector when it comes to accountability for the impact of AI-based decisions."[27] Other sets of principles and values may have more nuanced relevance for AI systems in particular. Yet, simply put, when it comes to options for present day benchmarks with the international legitimacy for addressing harms from AI, human rights may be as good as it gets.

There are a number of recent examples that incorporate human rights in AI governance. Harvard Berkman Klein Center's mapping of 36 sets of AI principles found that human rights are a dominant theme among them. Google's AI principles states the company will not pursue AI technologies "whose purpose contravenes widely accepted principles of international law and human rights."[28] Microsoft has conducted a human rights impact assessment for its AI programs.[29] Even Facebook has included human rights in their global community standards.[30] IEEE, the world's largest professional community of engineers, has stated that human rights is a first principle of ethical AI design.[31] And governmental bodies like the Council of Europe have anchored their policies in human rights law. These initiatives indicate a recognition of the value of human rights as a ready-made set of global values for AI governance.

But the above examples do not represent an overwhelming consensus and there are no guarantees human rights will emerge globally as the minimum set of normative values for AI governance. And therein lies the risk. If human rights are not seen as relevant for AI governance, then its relevance is indeed in question. If not human rights, what would take its place? Under what circumstances would make the world act as one around a global agreement of AI principles?

Catastrophe

AI systems acting autonomously in ways that significantly diverges from human values has been said to present potentially catastrophic risks to humanity.[32] For this paper, we should consider that a sustained crisis of legitimacy will exacerbate the unrestrained use of AI systems by governments, companies and other stakeholders, which also enhances the risk of catastrophe.

Piecemeal regulation is no match for a technology that is poised to scale across the planet without regard to boundaries.

[24] Sikkink, Kathryn. *Evidence for Hope: Making Human Rights Work in the 21st Century*. Princeton University Press, 2017.

[25] Sikkink, Kathryn. "The Effectiveness of U.S. Human Rights Policy, 1973–1980." The International Dimensions of Democratization: Europe and the Americas, edited by Laurence Whitehead, 93–124. Oxford: Oxford University Press, 1996.

[26] Sikkink, Kathryn and Kim, Hun Joon. "The Justice Cascade: The Origins and Effectiveness of Prosecutions of Human Rights Violations." Annual Review of Law and Social Science 9:1, 269-285, 2013.

[27] Donahoe, Eileen and Metzger, Megan MacDuffee. "Artificial Intelligence and Human Rights." Journal of Democracy, Volume 30, Number 2, April 2019, pp. 115-126.

[28] Artificial Intelligence at Google: Our Principles. https://ai.google/principles/.

[29] Microsoft. (2018). Salient Human Rights Issues Report, 2018. http://download.microsoft.com/download/6/9/2/692766EB-D542-49A2-AF27-CC8F9E6D3D54/Microsoft_Salient_Human_Rights_Issues_Report-FY17.pdf; It should be noted that Microsoft has not released the Human Rights Impact Assessment to the public.

[30] Facebook Community Standards. https://www.facebook.com/communitystandards/.

[31] IEEE. (2020). Ethically Aligned Design, First Edition, 2020.. https://ethicsinaction.ieee.org/.

[32] Russell, Stuart. *Human Compatible: AI and the Problem of Control* . New York: Viking, 2019. Russel, Stuart. "3 principles for creating safer AI." TED (talk), 2017. https://www.youtube.com/watch?v=EBK-a94IFHY; Here I will bracket the problem of whether an AI system might one day be capable of (self) learning about human rights and other value systems in order to act autonomously in accordance with human rights.

It is instructive to remember that 71 years ago, our contemporary system of international human rights was born out of catastrophe. It took a conflagration on the scale of WWII and the Holocaust to mobilize a vote of nations at the newly formed United Nations to adopt the Universal Declaration of Human Rights. A central aspect of the UN itself is to prevent another global catastrophe as the second UN Secretary General said: "the United Nations was not created in order to bring us to heaven, but in order to save us from hell."[33]

I will not speculate on what kind of future AI catastrophe might force the world's governments and stakeholders to develop an agreement on AI governance. There are plenty of studies of existential risk of superintelligence and even more science fiction narratives to choose from replete with well-trodden tropes like a self-aware general AI killing large swaths of humanity.[34]

Yet one need only look at the present day for a sense of the dystopian arising from unrestrained government or corporate use of AI systems. The Chinese Government is using AI-driven tools like mobile phone tracking apps, facial recognition, and data command centers to imprison, surveil, and "re-educate" the Muslim population of one million Uighurs in Xinjiang.[35] Through government and tech company partnerships, the situation in Xinjiang represents a large-scale AI assisted program of social control, oppression, and subjugation. As Chinese companies are exporting these technologies to countries like Malaysia and Zimbabwe, we need to be on guard about AI systems being used to exacerbate human rights violations on a global scale.

Myanmar provides another example, where the presence of Facebook became a promising sign that the country was opening up after decades of military dictatorship. However, the platform's news feed algorithm and lack of company oversight facilitated hate speech and incitement to violence against the Rohingya Muslim minority in the country.[36] In 2018, independent UN investigators found that Facebook was used to facilitate genocide, which forced over 700,000 Rohingya to flee to neighboring Bangladesh.[37]

The existing human rights framework not only provides the normative lens to understand the human impact of these two examples after the fact. Human rights provides mechanisms for accountability for the perpetrators of these acts in the political and legal spheres. And human rights provides the standards upon which to assess the potential human impacts of AI systems in the future.

Considering principle proliferation as a crises of legitimacy is a wake-up call for the urgent need to find a collective agreement for fundamental human values that will govern AI. This paper has argued that the existing human rights framework can serve as a baseline for those normative values in order to restrain the most harmful uses of AI. Of course, human rights cannot possibly address all of the sociotechnical concerns arising from AI development, design, and deployment. Therefore, human rights should be thought of as "the floor, not the ceiling" for AI governance.[38] If human rights is positioned as the needed grounding, then the AI governance space can open up to the plurality of principles and norms needed to guide the additional demands of AI in our present and future.

[33] Hammarskjöld, Dag. "Address by Secretary-General Dag Hammarskjöld at University of California Convocation." UC Berkeley, May, 1954. http://digitallibrary.un.org/record/1291161/files/SG-382.PDF.

[34] See Bostrom (ibid) and, generally, the *2001: A Space Odyssey*, *Matrix*, and *Terminator* films.

[35] See Human Rights Watch (2018). "Eradicating Ideological Viruses" China's Campaign of Repression Against Xinjiang's Muslims, 2018. https://www.hrw.org/report/2018/09/09/eradicating-ideological-viruses/chinas-campaign-repression-against-xinjiangs.

[36] See United Nations. "Report of the independent international fact-finding mission on Myanmar." Human Rights Council, September, 12, 2018. https://www.ohchr.org/Documents/HRBodies/HRCouncil/FFM-Myanmar/A_HRC_39_64.pdf/.

[37] Id.

[38] See Vidushi Marda's work in this area, e.g., Marda, Vidushi. "Governance with teeth: How human rights can strengthen FAT [Fairness Accountability Transparency] and ethics initiatives on artificial intelligence." Article 19, April, 2019 https://www.article19.org/wp-content/uploads/2019/04/Governance-with-teeth_A19_April_2019.pdf.

Carr Center Discussion Paper Series

Carr Center for Human Rights Policy
Harvard Kennedy School
79 JFK Street
Cambridge, MA 02138

This publication was published by the Carr Center for Human Rights Policy at the John F. Kennedy School of Government at Harvard University

79 JFK Street
Cambridge, MA 02138

617.495.5819
https://carrcenter.hks.harvard.edu