BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY
AT HARVARD UNIVERSITY

Research Publication No. 2018-6
September 25, 2018

Artificial Intelligence & Human Rights: Opportunities & Risks

Filippo A. Raso
Hannah Hilligoss
Vivek Krishnamurthy
Christopher Bavitz
Levin Kim

23 Everett Street • Second Floor • Cambridge, Massachusetts 02138
+1 617.495.7547 • +1 617.495.7641 (fax) • http://cyber.law.harvard.edu/ •
cyber@law.harvard.edu

BERKMAN
KLEIN CENTER
FOR INTERNET & SOCIETY
AT HARVARD UNIVERSITY

Artificial Intelligence & Human Rights:

Opportunities & Risks

September 25, 2018

Filippo Raso
Hannah Hilligoss
Vivek Krishnamurthy
Christopher Bavitz
Levin Kim

Table of Contents

Background & Context 3

Summary of Findings 4

1 Introduction 7

2 What is Artificial Intelligence? 10

3 What are Human Rights? 12

4 Identifying the Human Rights Consequences of AI 14

5 AI's Multifaceted Human Rights Impacts 17

5.1 Criminal Justice: Risk Assessments 20

5.2 Access to the Financial System: Credit Scores 26

5.3 Healthcare: Diagnostics 32

5.4 Online Content Moderation: Standards Enforcement 37

5.5 Human Resources: Recruitment and Hiring 42

5.6 Education: Essay Scoring 47

6 Addressing the Human Rights Impacts of AI: The Strengths and Limits of a Due Diligence-Based Approach 52

7 Conclusion 58

8 Further Reading 59

Background & Context[1]

This report explores the human rights impacts of artificial intelligence ("AI") technologies. It highlights the risks that AI, algorithms, machine learning, and related technologies may pose to human rights, while also recognizing the opportunities these technologies present to enhance the enjoyment of the rights enshrined in the Universal Declaration of Human Rights ("UDHR"). The report draws heavily on the United Nations Guiding Principles on Business and Human Rights ("Guiding Principles") to propose a framework for identifying, mitigating, and remedying the human rights risks posed by AI.

Readers wishing to better understand the often-paradoxical human rights impacts of the six current AI applications that are detailed in this report are invited to explore a series of interactive visualizations that are available at **ai-hr.cyber.harvard.edu**.

[1] We would like to express our appreciation to our Berkman Klein colleagues Amar Ashar, Ryan Budish, Daniel Dennis Jones, Rob Faris, Jessica Fjeld, Sarah Newman and Casey Tilton for their helpful suggestions throughout the course of this project; to our interns and research assistants Sam Bookman, Daniel Chase, Christina Chen, Areeba Jibril, Adam Nagy, and Marianne Strassle for their assistance in finalizing this report; and to Urs Gasser and Jonathan Zittrain, respectively the Executive Director and Faculty Chair of the Berkman Klein Center, for their visionary leadership of our Center and of the Ethics and Governance of Artificial Intelligence Fund.

We are grateful to Kim Albrecht, Solon Barocas, Dinah PoKempner, Mark Latonero, An Xiao Mina, Brian Root, Maria Sapignoli, Seamus Tuohy, and Andrew Zick for consulting with us at various stages of this project and helping us refine our analytical methodology.

Finally, we wish to thank the Government of Canada for its sponsorship of this project, and in particular thank Tara Denham, Salahuddin Rafiqhuddin, Philippe-André Rodriguez, Marketa Geislerova, Maroussia Levesque, Asha Mohidin Siad and Jennifer Jeppsson of Global Affairs Canada for their consistent support of this project.

The views expressed in this report are those of the authors alone and do not reflect those of the Government of Canada or of the Berkman Klein Center for Internet & Society at Harvard University.

Summary of Findings

A Human Rights-based Approach to AI's Impacts

The ongoing dialogue regarding the ethics of artificial intelligence (AI) should expand to consider the human rights implications of these technologies.

International human rights law provides a universally accepted framework for considering, evaluating, and ultimately redressing the impacts of artificial intelligence on individuals and society.

Since businesses are at the forefront of developing and implementing AI, the United Nations Guiding Principles on Business and Human Rights are especially salient in ensuring that AI is deployed in a rights-respecting manner.

Determining Impacts

We propose that the best way to understand the impact of AI on human rights is by examining the difference, both positive and negative, that the introduction of AI into a given social institution makes to its human rights impacts. We take this view for two reasons:

1. Determining the human rights impacts of AI is no easy feat, for these technologies are being introduced and incorporated into existing social institutions, which are not rights-neutral.

2. Each application of AI impacts a multitude of rights in complicated and, occasionally, contradictory ways. Exploring these relationships within use cases allows for more nuanced analysis.

Measuring Impacts

Current implementations of AI impact the full range of human rights guaranteed by international human rights instruments, including civil and political rights, as well as economic, cultural, and social rights.

Privacy is the single right that is most impacted by current implementations of AI. Other rights that are also significantly impacted by current AI implementations include the rights to equality, free expression, association, assembly, and work. Regrettably, the impact of AI on these rights has been more negative than positive to date.

The positive and negative impacts of AI on human rights are not distributed equally throughout society. Some individuals and groups are affected more strongly than others, whether negatively or positively. And at times, certain AI implementations can positively impact the enjoyment of a human right by some while adversely impacting it for others.

Addressing Impacts

Addressing the human rights impacts of AI is challenging because these systems can be accurate and unfair at the same time. Accurate data can embed deep-seated injustices that, when fed into AI systems, produce unfair results. This problem can only be addressed through the conscious efforts of AI systems designers, end users, and ultimately of governments, too.

Many of the existing formal and informal institutions that govern various fields of social endeavor are ill-suited to addressing the challenges posed by AI. Institutional innovation is needed to ensure the appropriate governance of these technologies and to provide accountability for their inevitable adverse effects.

The Path Forward

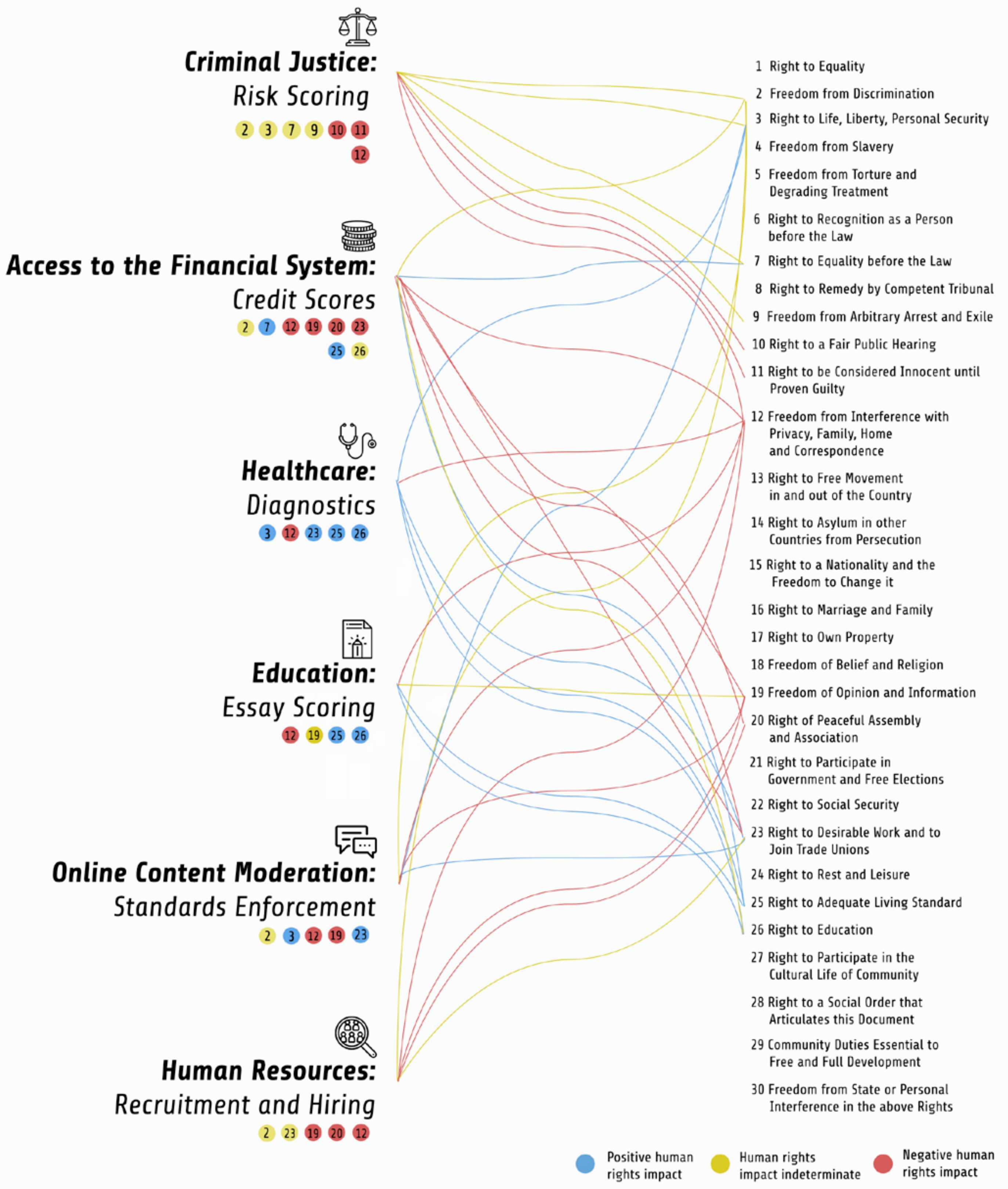
Human rights due diligence by businesses can help avoid many of the adverse human rights impacts of AI.

Non-state grievance and remedy mechanisms can provide effective redress for some, but by no means all, of the inevitable adverse impacts that AI will produce.

Governments have an important role to play in creating effective mechanisms to remedy the adverse human rights impacts of AI.

The role of government is essential to addressing the distributive consequences of AI by means of the democratic process.

ARTIFICIAL INTELLIGENCE & HUMAN RIGHTS



Criminal Justice: *Risk Scoring*
🟡2 🟡3 🟡7 🟡9 🔴10 🔴11 🔴12

Access to the Financial System: *Credit Scores*
🟡2 🔵7 🔴12 🔴19 🔴20 🔴23 🔵25 🟡26

Healthcare: *Diagnostics*
🔵3 🔴12 🔵23 🔵25 🔵26

Education: *Essay Scoring*
🔴12 🟡19 🔵25 🔵26

Online Content Moderation: *Standards Enforcement*
🟡2 🔵3 🔴12 🔴19 🔵23

Human Resources: *Recruitment and Hiring*
🟡2 🟡23 🔴19 🔴20 🔴12

1 Right to Equality
2 Freedom from Discrimination
3 Right to Life, Liberty, Personal Security
4 Freedom from Slavery
5 Freedom from Torture and Degrading Treatment
6 Right to Recognition as a Person before the Law
7 Right to Equality before the Law
8 Right to Remedy by Competent Tribunal
9 Freedom from Arbitrary Arrest and Exile
10 Right to a Fair Public Hearing
11 Right to be Considered Innocent until Proven Guilty
12 Freedom from Interference with Privacy, Family, Home and Correspondence
13 Right to Free Movement in and out of the Country
14 Right to Asylum in other Countries from Persecution
15 Right to a Nationality and the Freedom to Change it
16 Right to Marriage and Family
17 Right to Own Property
18 Freedom of Belief and Religion
19 Freedom of Opinion and Information
20 Right of Peaceful Assembly and Association
21 Right to Participate in Government and Free Elections
22 Right to Social Security
23 Right to Desirable Work and to Join Trade Unions
24 Right to Rest and Leisure
25 Right to Adequate Living Standard
26 Right to Education
27 Right to Participate in the Cultural Life of Community
28 Right to a Social Order that Articulates this Document
29 Community Duties Essential to Free and Full Development
30 Freedom from State or Personal Interference in the above Rights

🔵 Positive human rights impact
🟡 Human rights impact indeterminate
🔴 Negative human rights impact

1. Introduction

Artificial intelligence ("AI") is changing the world before our eyes. Once the province of science fiction, we now carry systems powered by AI in our pockets and wear them on our wrists. Vehicles on the market can now drive themselves, diagnostic systems determine what is ailing us, and risk assessment algorithms increasingly decide whether we are jailed or set free after being charged with a crime.

The promise of AI to improve our lives is enormous. AI-based systems are already outperforming medical specialists in diagnosing certain diseases, while the use of AI in the financial system is expanding access to credit to borrowers that were once passed by. Automated hiring systems promise to evaluate job candidates on the basis of their bona fide qualifications, rather than on qualities such as age or appearance that often lead human decision-makers astray. AI promises to allow institutions to do more while spending less, with concomitant benefits for the availability and accessibility of all kinds of services.

Yet AI also has downsides that dampen its considerable promise. Foremost among these is that AI systems depend on the generation, collection, storage, analysis, and use of vast quantities of data—with corresponding impacts on the right to privacy. AI techniques can be used to discover some of our most intimate secrets by drawing profound correlations out of seemingly innocuous bits of data. AI can easily perpetuate existing patterns of bias and discrimination, since the most common way to deploy these systems is to "train" them to replicate the outcomes achieved by human decision-makers. What is worse, the "veneer of objectivity" around high-tech systems in general can obscure the fact that they produce results that are no better, and sometimes much worse, than those hewn from the "crooked timber of humanity."

These dystopian possibilities have given rise to a chorus of voices calling for the need for Fairness, Accountability, and Transparency in Machine Learning ("FAT" or "FAT/ML"). Advocates of this approach view the response to AI's potential problems in terms of ethics. For example, the Institute of Electrical and Electronics Engineers—the world's largest technical professional body that plays an important role in setting technology standards—has published an influential treatise on *Ethically Aligned Design* that suggests that "the full benefit of these technologies will be attained only if they are aligned with our defined values and ethical principles."[2] In a similar vein, the governments of France and India have recently released discussion papers to frame their national strategies on AI that embrace an ethics-based approach to addressing the social impacts of these technologies.[3]

During the pendency of this project, however, several influential actors have come to recognize the

[2] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, "EthicallyAligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems." Version 2. http://standards.ieee.org/develop/indconn/ec/autonomous_ systems.html.

[3] For France's strategy, see: Cédric Villani, "For a Meaningful Artificial Intelligence: Towards a French and European Strategy" (AI For Humanity), accessed June 22, 2018, https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf. For India's, see Amitabh Kant, "National Strategy for Artificial Intelligence" (NITI Aayog, June 2018), www.niti.gov.in/writereaddata/files/document_publication/NationalStrategy-for-AI-Discussion-Paper.pdf. Although both papers are substantial works that are each over 100 pages long, they barely mention the concept of human rights.

value of examining the challenges around AI from a human rights perspective.[4] This incipient conversation on AI and human rights has already produced two significant documents. One is the Toronto Declaration on Protecting the Rights to Equality and Non-Discrimination in Machine Learning Systems ("Toronto Declaration"), which was opened for signatures on May 16, 2018.[5] As its full title suggests, the Toronto Declaration highlights the potential adverse effects of machine learning on rights to equality and non-discrimination and calls for the development of effective remedial mechanisms for all those who are adversely affected by these systems.[6] The other is Global Affairs Canada's Draft Strategy Paper on the Human Rights and Foreign Policy Implications of AI, which examines how AI can impact the rights to equality, privacy, free expression, association, and assembly, and suggests ways that these impacts can be redressed.[7]

This project is rooted in the belief that there is considerable value in adopting a human rights perspective to evaluating and addressing the complex impacts of AI on society. The value lies in the ability of human rights to provide an agreed set of norms for assessing and addressing the impacts of the many applications of this technology, while also providing a shared language and global infrastructure around which different stakeholders can engage.[8]

While there are many different conceptions of human rights, from the philosophical to the moral, we in this project take a legal approach. We view human rights in terms of the binding legal commitments the international community has articulated in the three landmark instruments that make up the International Bill of Rights.[9] This body of law has developed over time with the ratification of new treaties, the publication of General Comments that authoritatively interpret the provisions of these treaties, and through the work of international and domestic courts and tribunals, which have applied the provisions of these treaties to specific cases.

Our project seeks to advance the burgeoning conversation on AI and human rights by mapping the human rights impacts of the current deployment of AI systems in six different fields of endeavor. We strive to move beyond the predominant focus on AI's impact on select civil and political rights, to consider how these technologies are impacting other rights guaranteed by international law—especially economic, social, and cultural rights.

[4] For example, Amnesty International launched a structured initiative on Artificial Intelligence and Human Rights in 2017, while the New York-based Data & Society Research Institute hosted a workshop on Artificial Intelligence and Human Rights in April, 2018. See Sherif Elsayed-Ali, "Artificial Intelligence and the Future of Human Rights," Oct. 19, 2017. https://medium.com/amnesty-insights/artificial-intelligence-and-the-future-of-human-rights-b58996964df5. Mark Latonero, "Artificial Intelligence & Human Rights: A Workshop at Data & Society." May 11, 2018. https://points.datasociety.net/artificial-intelligence-human-rights-a-workshop-at-data-society-fd6358d72149.

[5] Toronto Declaration on Protecting the Rights to Equality and Non-Discrimination in Machine Learning Systems, May 16, 2018. https://www.accessnow.org/cms/assets/uploads/2018/05/Toronto-Declaration-D0V2.pdf.

[6] Ibid.

[7] Digital Inclusion Lab, Global Affairs Canada, "Artificial Intelligence: Human Rights & Foreign Policy Implications." Accessed June 1, 2018. https://docs.google.com/document/d/1fhIJYznWSI70D3TVJ5CgLgHJMJ2HouEZiQ9a_qKbLG0/edit ("GAC Strategy Paper").

[8] Jason Pielemeier, "The Advantages and Limitations of Applying the International Human Rights Framework to Artificial Intelligence," Data & Society: Points, June 6, 2018, https://points.datasociety.net/the-advantages-and-limitations-of-applying-the-international-human-rights-framework-to-artificial-291a2dfe1d8a.

[9] The "International Bill of Rights" is a term to describe the three most important international human rights instruments, namely 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").

In so doing, we suggest what we believe to be the optimal method for identifying the human rights impacts of introducing a particular AI system into a given field of endeavor. Simply put, we believe it is important to recognize that AI systems are not being deployed against a blank slate, but rather against the backdrop of social conditions that have complex pre-existing human rights impacts of their own. This may well appear to be a self-evident truth, but in our view, the existing literature does not adequately consider the impact of these background conditions on the consequences of introducing AI. As a result, human rights impacts, both positive and negative, may be misattributed to AI, contributing to the extreme claims of optimists and pessimists alike about the extent to which AI is changing our lives.

Our report and the accompanying visualizations make clear that AI is already impacting the enjoyment of the full range of human rights–sometimes in paradoxical ways. In the final section, we examine and evaluate how international human rights law generally, and the growing field of business and human rights specifically, can help the developers, users, and regulators of AI systems to address many of these impacts.

2. What is Artificial Intelligence?

Despite its expanding presence across many aspects of our lives, there is no widely accepted definition of "artificial intelligence."[10] Instead, it is an umbrella term that includes a variety of computational techniques and associated processes dedicated to improving the ability of machines to do things requiring intelligence, such as pattern recognition, computer vision, and language processing.[11] With such a loose conceptualization and given the rapid growth of technology, it is no surprise that what is considered artificial intelligence changes over time. This is known as the "AI effect" or the "odd paradox": formerly cutting-edge innovations become mundane and routine, losing the privilege of being categorized as AI, while new technologies with more impressive capabilities are labeled as AI instead.[12]

The impossibly large set of technologies, techniques, and applications that fall under the AI umbrella can be usefully classified into two buckets. The first is comprised of *knowledge-based systems*, which are "committed to the notion of generating behavior by means of deduction from a set of axioms."[13] These include "expert systems" which use formal logic and coded rules to engage in reasoning. Such systems, which are sometimes also called "closed-rule algorithms," include everything from commercial tax preparation software to the first generation of healthcare diagnostic decision support algorithms. These systems are good at taking concrete situations and reasoning optimal decisions based on defined rules within a specific domain. They cannot, however, learn or automatically leverage the information they have accumulated over time to improve the quality of their decision-making (unless they are paired up with some of the techniques described below).[14]

The second bucket of technologies uses statistical learning to continuously improve their decision-making performance. This new wave of technology, which encompasses the widely-discussed techniques known as "machine learning" and "deep learning," has been made possible by the exponential growth of computer processing power, the

[10] National Science and Technology Council: Committee on Technology, "Preparing for the Future of Artificial Intelligence," Government Report (Washington, D.C.: Executive Office of the President, October 2016).

[11] One seminal textbook categorizes AI into (1) systems that think like humans (e.g., cognitive architectures and neural networks); (2) systems that act like humans (e.g., pass the Turing test, knowledge representation, automated reasoning, and learning), (3) systems that think rationally (e.g., logic solvers, inference, and optimization); and (4) systems that act rationally (e.g., intelligent software agents and embodied robots that achieve goals via perception, planning, reasoning, learning, communicating, decision-making, and acting), Stuart J. Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, Prentice Hall Series in Artificial Intelligence (Englewood Cliffs, N.J: Prentice Hall, 1995).

[12] Pamela McCorduck, *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence*, 2nd ed. (Natick, MA: A. K. Peters, Ltd., 2004).

[13] Nello Cristianini, "On the Current Paradigm in Artificial Intelligence," *AI Communications* 27, no. 1 (January 1, 2014): 37–43, https://doi.org/10.3233/AIC-130582.

[14] Bruce G. Buchanan, "Can Machine Learning Offer Anything to Expert Systems?," *Machine Learning* 4, no. 3–4 (December 1, 1989): 251–54, https://doi.org/10.1023/A:1022646520981.

massive decline in the cost of digital storage, and the resulting acceleration of data collection efforts.[15] Systems in this category include self-driving vehicles, facial recognition systems used in policing, natural language processing techniques that are used to automate translation and content moderation, and even algorithms that tell you what to watch next on video streaming services. While these systems are impressive in their aggregate capacities, they are probabilistic and can thus be unreliable at the individual level. For example, deep learning computer vision systems can classify an image almost as accurately as a human; however, they will occasionally make mistakes that no human would make—such as mistaking a photo of a turtle for a gun.[16] They are also susceptible to being misled by "adversarial examples," which are inputs that are tampered with in a way that leads an algorithm to output an incorrect answer with high confidence.[17]

In this report, we focus on AI systems from both of these conceptual buckets that "perceive[] and act[]"[18] upon the external environment by "tak[ing] the best possible action in a situation."[19] Simply put, the scope of our report is limited to analyzing those AI systems that automate the making of decisions that were formerly the exclusive province of human intelligence. This view of AI embraces everything from medical diagnostic software that determine what is ailing a patient based on the available evidence, to self-driving vehicles that "decide" whether to steer, accelerate, or brake, millisecond by millisecond. The crucial factor for us is that the system must function and impact the external environment, rather than simply be a theoretical construct that remains under development, to be considered within the scope of our report. Furthermore, we limit our scope to AI technologies that are either currently in use or are far along in the development process; therefore, we do not delve into the realm of artificial general intelligence.[20] We restrict our consideration of AI in this report to those technologies that are being used to make decisions with real-world consequences for the simple reason that these are the technologies that are most likely to have discernible human rights impacts. By contrast, many other strains of AI research remain conceptual for now, and are thus yet to impact human rights.

[15] Gheorghe Tecuci, "Artificial Intelligence," *Wiley Interdisciplinary Reviews: Computational Statistics* 4, no. 2 (2012): 168–80, https://doi.org/10.1002/wics.200.

[16] Adam Conner-Simons, "Fooling Neural Networks w/3D-Printed Objects," MIT Computer Science & Artificial Intelligence Lab (blog), November 2, 2017, https://www.csail.mit.edu/news/fooling-neural-networks-w3d-printed-objects.

[17] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and Harnessing Adversarial Examples," *ArXiv:1412.6572 [Cs, Stat]*, December 19, 2014, http://arxiv.org/abs/1412.6572; A Nguyen, J Yosinski, and J Clune, "Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images," *CVPR*, IEEE, 15 (2015)

[18] Russell and Norvig, *Artificial Intelligence,* 7.

[19] Ibid., 27.

[20] Broadly speaking, an AGI system is one that can perform any task as well as a human can, or a "synthetic intelligence that has a general scope and is good at generalization across various goals and contexts," Ben Goertzel, "Artificial General Intelligence: Concept, State of the Art, and Future Prospects," *Journal of Artificial General Intelligence* 5, no. 1 (December 1, 2014): 1–48, https://doi.org/10.2478/jagi-2014-0001.

3. What are Human Rights?

As noted in the introduction, in this report we adopt a legal conception of human rights. We use the term human rights to refer to those individual and collective rights that have been enshrined first and foremost in the Universal Declaration of Human Rights ("UDHR"), and then further detailed in the International Covenant on Civil and Political Rights ("ICCPR") and the International Covenant on Economic, Social and Cultural Rights ("ICESCR").

The UDHR is the leading statement of the rights that every human being enjoys by virtue of their birth. Although the UDHR was adopted by means of a non-binding U.N. General Assembly resolution,[21] Canada and many other states have long believed that there is an "obligation on states to observe the human rights and fundamental freedoms enunciated in the [UDHR] [that] derives from their adherence to the Charter of the United Nations," which is binding international law.[22]

The ICCPR and the ICESCR, meanwhile, are international treaties that are binding upon those states that have ratified them. These treaties elaborate upon the human rights that were first articulated by the UDHR at the international level, and clarify the duties of states in relation to two categories of rights. Whereas the ICCPR's protections of civil and political rights come into force immediately upon ratification,[23] the ICESCR instead requires states to take measures to progressively realize the economic, social, and cultural rights it protects, having due regard for the state's economic condition and resources.[24]

States shoulder a binding obligation under international law to protect human rights. This includes a duty to respect human rights in their own conduct, and to prevent natural and juridical persons subject to their jurisdiction (including corporations) from committing human rights abuses. These obligations persist even when privatizing the delivery of services that may impact human rights.[25]

Especially since the end of the Cold War, businesses have come to be viewed as having their own responsibilities under international law to respect human rights.[26] The nature and scope of these responsibilities have been articulated most authoritatively in the United Nations Guiding Principles on Business and Human Rights ("UNGP" or "Guiding Principles"). Specifically, the responsibility to respect human rights requires enterprises to avoid causing or contributing to adverse human rights impacts through their own activities, and to seek to prevent or mitigate such impacts when the enterprise is

[21] Universal Declaration of Human Rights (10 Dec. 1948), U.N.G.A. Res. 217 A (III) (1948) [hereinafter "UDHR"].

[22] Letter from the Legal Bureau, Jan. 9, 1979, reprinted in *Canadian Practice in International Law,* 1980 Can. Y.B. Int'L L. 326.

[23] International Covenant on Civil and Political Rights (New York, 16 Dec. 1966) 999 U.N.T.S. 171 and 1057 U.N.T.S. 407, entered into force 23 Mar. 1976, art. 2 [hereinafter "ICCPR"].

[24] International Covenant on Economic, Social and Cultural Rights (New York, 16 Dec. 1966) 993 U.N.T.S. 3, entered into force 3 Jan. 1976, art. 2(1) [hereinafter "ICESCR"].

[25] Human Rights Council Res. 17/4, Rep. of the Hum. Rts. Council, 17th Sess., June 16, 2011, U.N. Doc. A/HRC/RES/17/4 (July 6, 2011); Special Rep. of the Sec'y Gen., Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework, Hum. Rts. Council, U.N. Doc. A/HRC/17/31 (Mar. 21, 2011) [hereinafter "Guiding Principles"], Principle 5.

[26] Guiding Principles, Principle 11.

"directly linked" to them via a business relationship.[27] This, in turn, requires enterprises to engage in ongoing due diligence processes to identify, prevent, and mitigate salient human rights risks.[28] To the extent that adverse human rights impacts do occur, businesses should provide remediation for those impacts through legitimate mechanisms[29]—although it is emphatically the duty of the state to provide effective remedies through judicial and other mechanisms to those who have suffered business-related human rights abuses.[30]

Although the Guiding Principles do not themselves have the force of law, they clarify how pre-existing international human rights standards apply to business activities, and provide useful guidance on how businesses can operate in a rights-respecting manner.[31] In any event, since businesses are at the forefront of developing and deploying AI, the Guiding Principles are of immense importance to ensuring that the human rights impacts of these powerful new technologies are positive. Consequently, the Guiding Principles will feature prominently in the discussion that follows of the human rights impacts of AI systems that are currently in use, and in our suggestions regarding how they should be addressed.

[27] Ibid., Principle 13.

[28] Ibid., Principle 17.

[29] Ibid., Principle 22.

[30] Ibid., Principle 25.

[31] Justine Nolan, "The Corporate Responsibility to Respect Human Rights: Soft Law or Not Law?," in *Human Rights Obligations of Business*, ed. Surya Deva and David Bilchitz (Cambridge: Cambridge University Press, 2013), 138–61, https://doi.org/10.1017/CBO9781139568333.010.

4. Identifying the Human Rights Consequences of AI

AI is not being developed in a vacuum or deployed against a blank slate. Rather, specific actors in society are deploying AI to automate decision-making in particular fields of endeavor. They are doing so to achieve outcomes that they view as desirable, against the backdrop of social institutions that have their own, pre-existing human rights implications.

Consider, for example, the deployment of AI in the criminal justice system, which is discussed in more detail in the first case study below. Over the course of the last several hundred years, criminal defendants have been endowed with various rights to ensure the fairness of criminal proceedings. These include the presumption of innocence,[32] the principle of legality,[33] the right to a fair trial,[34] and many others. Even so, no existing criminal justice system comes close to perfectly respecting the rights of defendants and other relevant rights-holders: every such system has at least some negative impacts on rights-holders that predate the introduction of AI.[35]

It is only by embracing a comparative approach, that accounts for background conditions from the pre-AI world, that we can properly understand the human rights impacts of introducing AI into the criminal justice system or any other human institution. Unless the human rights implications, both positive and negative, of pre-existing institutional structures are identified and accounted for, the human rights impacts of introducing AI will be conflated with the ongoing impacts of whatever was there before. Below, we propose a two-step methodology for avoiding such difficulties.

Step 1: Establish the Baseline

As noted, the first step is to simply consider the existing human rights implications, both positive and negative, of whatever field of endeavor AI is being introduced into. This evaluation properly involves consideration of the availability and effectiveness of institutional mechanisms that are currently in place to regulate and redress the negative human rights implications arising from that field. When human decision-making in the field in question has already been supplanted by a first-generation automated decision-making technology, such as a closed-rule diagnostic algorithm, the first step consists of evaluating the human rights implications of the pre-AI status quo.

Step 2: Identify the Impacts of AI

The second step involves identifying how the introduction of AI changes the human rights impacts of the field into which the technology is introduced. If the introduction of AI improves the human rights performance of the field, AI can be said to have a positive impact on human rights. That is true even if the field of endeavor continues to produce adverse human rights impacts after the introduction

[32] UDHR art. 11.

[33] UDHR art. 11(2).

[34] ICCPR art. 14(1).

[35] For the last two years, the Berkman Klein Center has been conducting extensive research on the use of algorithms in the criminal justice system, in its capacity as one of the two anchor institutions for the Ethics and Governance of Artificial Intelligence initiative. The research outputs of this ongoing work can be found at https://cyber.harvard.edu/research/ai.

of AI. Conversely, if the human rights performance of the field of endeavor deteriorates with the introduction of AI, then it is clear that the technology has produced adverse human rights impacts. To a significant extent, the outcome of this evaluation will depend on whether the mechanisms currently in place to regulate and remedy the adverse human rights consequences of the field in question continue to be effective following the introduction of AI.

The human rights impacts of AI stem from at least three sources, two of which can be considered by conducting a human rights impact assessment before a particular system is deployed. The third source, meanwhile, can be hard to identify even after an AI system is in operation, due to the complexity of the technology:

1. *Quality of Training Data:* To the extent that the data used to "train" an AI system is biased, the resulting system will reflect, or perhaps even exacerbate, those biases.[36] This is a version of what is known as the "garbage in, garbage out" problem, and it can have profound consequences for a wide variety of human rights–depending on what the system is intended to do.

2. *System Design:* Decisions made by an AI system's human designers can have significant human rights consequences. Human designers can, for example, prioritize the variables they would like the AI system to optimize and decide what variables the AI should take into consideration as it operates. Such design decisions can have both positive and negative human rights impacts, which will be informed by the individual life experiences and biases of the designers. Some of these impacts will be foreseeable, while others will not be.

3. *Complex Interactions*: Once an AI system is introduced, it will interact with the environment in ways that produce outcomes that might not have been foreseen. These complex interactions can have significant human rights impacts. In some cases, the impacts of these interactions may be detectable through the use of certain analytical techniques, but the possibility exists that certain human rights impacts resulting from the deployment of an AI system will escape detection. This is not an issue that is unique to AI: pre-digital societies are staggeringly complex, and the human rights impacts of the actions of individuals and institutions are not always knowable at the time they are made or for some time thereafter.

Limitations of our approach

Our two-step methodology provides a useful, generalizable approach to identifying the positive and negative implications of introducing AI into an extant field of endeavor. This methodology, which we have validated in consultations with stakeholders from the technology and human rights communities, undergirds our assessment of the human rights impacts of AI across six different use cases below.

Our framework has its limitations, especially due to the scarcity of available information into the design and operation of any given AI system. This is due in part to the novelty of AI, but also because so much AI technology is proprietary, which results in information about the design, operation, and impact of

[36] Osonde Osoba & William Welser IV, *An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence*. (Santa Monica: Rand Corporation, 2017). https://www.rand.org/content/dam/rand/pubs/research_reports/RR1700/RR1744/RAND_RR1744.pdf.

the systems being treated by their creators as commercially sensitive information.[37]

Consequently, the analysis we undertake in our six case studies, below, is at the level of detail that one would find in a sectoral human rights impact assessment. Based on our desktop research, we have drawn reasonable inferences as to the likely human rights impacts of introducing particular AI systems into the prevailing social and institutional context.

[37] Rebecca Wexler, "Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System," *Stanford Law Review* 70, no. 5 (2017): 1343–1429, https://doi.org/10.2139/ssrn.2920883.

5. AI's Multifaceted Human Rights Impacts

Applying the two-step framework from the previous section, we now explore the wide-ranging human rights consequences of introducing AI decision-making into six fields:

- Criminal Justice (risk assessments)
- Finance (credit scores)
- Healthcare (diagnostics)
- Content Moderation (standards enforcement)
- Human Resources (recruitment and hiring)
- Education (essay scoring)

We chose these six fields out of many possibilities because they illustrate the promise and the perils of this technology across a range of human rights. What is more, AI decision-making technologies are already in use in all of these fields, which allows our analysis to be grounded in the here and now, rather than speculating about future developments.

In choosing these six use cases, we consciously decided not to include two AI applications that have generated a great deal of debate and controversy: namely, self-driving vehicles and autonomous weapons systems. We excluded these applications from our analysis because both are much better studied than the use of AI in the other six fields that we have chosen. Furthermore, the issues surrounding autonomous weapons systems are more appropriately answered with reference to international humanitarian law rather than international human rights law, since such systems are meant to be used in times of conflict.

In undertaking this analysis, it quickly became apparent to us that each AI deployment had the potential to impact a large number of rights via their first- and second-order effects. In the interest of clarity and analytical efficiency, however, we have focused our analysis on those rights that we believe to be most impacted by the deployment in question. This is an exercise in line-drawing that is subjective by its very nature, but is part and parcel of the approach embraced by the Guiding Principles to identifying human rights impacts so that they may be appropriately addressed.[38]

There are five main points that emerge from our analysis.

First and foremost, the six use cases we explore in detail reveal how AI-based decision-making technologies impact the full spectrum of political, civil, economic, social, and cultural rights secured by the UDHR and further expounded upon in the ICCPR and the ICESCR.

Second, the positive and negative human rights impacts caused by AI are not evenly distributed across society. Some individuals and groups experience positive impacts from the very same applications that adversely impact other rights-holders. In some

[38] Guiding Principles, Principles 17 and 24.

cases, a particular AI application can positively impact the enjoyment of a given human right for a particular class of individuals, while adversely affecting the enjoyment of the very same human right by others. For example, the use of automated risk scoring systems in the criminal justice system may reduce the number of individuals from the majority group who are needlessly incarcerated, at the very same time that flaws in the system serve to increase the rate of mistaken incarcerations for those belonging to marginalized groups.[39]

Third, AI carries the serious risk of perpetuating, amplifying, and ultimately ossifying existing social biases and prejudices, with attendant consequences for the right to equality. This problem, which has been termed by one analyst as "counter-serendipity," results from the fact that AI systems are trained to replicate patterns of decision-making they learn from training data that reflects the social status quo—existing human biases, entrenched power dynamics and all.[40] But therein lies the problem: to the extent that an AI accurately replicates past patterns of human decision-making, it will necessarily perpetuate existing social biases as well.[41] What is worse, unlike human decision-makers, who have the agency and the free will to change their moral perspective over time, for the foreseeable future AI systems will not have any such capabilities of their own. Instead, they require constant attention by those who are responsible for the design and operation of such systems to ensure that their outputs are consistent with evolving notions of fairness.

To be sure, the automation of decision-making through AI offers the possibility of righting significant social wrongs by designing the systems to have ameliorative effects. Such effects could be achieved by seeking to correct for biases in human decision-making, or more controversially, through "algorithmic affirmative action"—that is, by designing algorithms to counter the historical disadvantages that marginalized groups have faced.[42] The larger point, however, is that unless AI systems are consciously designed and consistently evaluated for their differential impacts on different populations, they have the very real potential to hinder rather than help progress towards greater equity.

Fourth, as is likely expected, most AI technologies have a deleterious impact on the right to privacy. AIs are data-hungry by their nature; they are fundamentally premised on algorithms automatically poring over vast datasets to generate answers, predictions, and insights. Accordingly, AI systems rely on the collection, storage, consolidation, and analysis of vast quantities of data. They also create powerful incentives to gather and store as much additional data as can be, in view of the possibility that new data streams will allow for AI systems to generate powerful new insights. Much of the data that fuels AI systems will either be personally identifiable, or rife with the possibility of being re-identified using an algorithm in the event that it was anonymized. Moreover, even if techniques such as differential privacy[43] are used to protect the privacy of particular individuals, AI technologies may gen-

[39] Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, "Machine Bias," *ProPublica*, May 23, 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[40] Edward Tenner, *The Efficiency Paradox: What Big Data Can't Do* (New York: Alfred A. Knopf, 2018); Berkman Klein Center for Internet and Society, "Artificial Intelligence and Inclusion," accessed June 22, 2018, https://aiandinclusion.org/.

[41] Anupam Chander, "The Racist Algorithm?," *Mich. Law Review* 115, no. 6 (2017): 1023, http://michiganlawreview.org/wp-content/uploads/2017/04/115MichLRev1023_Chander.pdf.

[42] Ibid.

[43] Cynthia Dwork et al., "Calibrating Noise to Sensitivity in Private Data Analysis," in Theory of Cryptography, ed. Shai Halevi and Tal Rabin

erate insights from such data that are then used to make predictions about, and act upon, the intimate characteristics of a particular person—all while refraining from identifying the natural person. For example, a retailer might train an AI-based marketing system using sales data that has been de-identified and subjected to differential privacy techniques. But even assuming that the training data is discarded once the system is in operation, the insights generated by the system from the data it is tasked with analyzing can nevertheless have a significant impact on an individual's privacy.[44] Given that most extant AI applications have very significant privacy implications, we focus our analysis in the case studies below on the other rights that are impacted by these systems. This is a pragmatic choice made in the interests advancing the AI and human rights conversation beyond privacy.

Fifth, the rise of artificial intelligence poses a challenge for many of the existing mechanisms that currently exist to right wrongs. In the United States, for example, individuals have a right to request a copy of their credit report and to require credit reporting agencies to investigate and correct any errors appearing on their report.[45] By contrast, there is currently no law in the United States that would provide an individual with recourse if a lender using an algorithm that crunches through thousands of variables from thousands of sources does so on the basis of erroneous data. Even in Canada[46] and the European Union,[47] where privacy laws currently in force allow individuals to demand the correction of errors in their data, the sheer volume of information that AI systems use as they make a decision makes it difficult to exercise this right effectively. Moreover, even if one aggrieved individual corrects errors in their own data, significant harms can occur due to the presence of systematic errors in a data set and ubiquitous data sharing, which can lead to unfair outcomes for potentially vast numbers of people.

(Springer Berlin Heidelberg, 2006), 265–84.

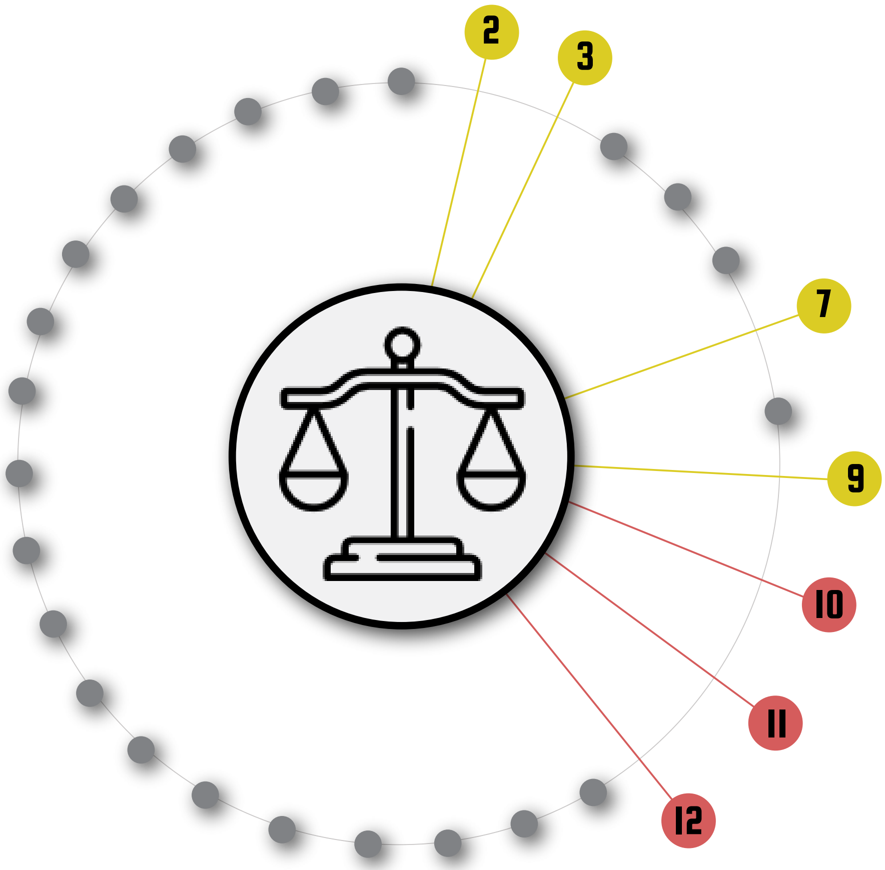
[44] Charles Duhigg, "How Companies Learn Your Secrets," *The New York Times*, February 16, 2012, sec. Magazine, https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html.

[45] *Fair Credit Reporting Act*, 15 U.S.C. § 1681i (2012) ("... 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 (2012) (free annual copy of one's credit report).

[46] *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.)

[47] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC, 2016 O.J. (L. 119) [henceforth "GDPR"], art. 19 ("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.").

5.1 Criminal Justice: Risk Assessments



2. Freedom from Discrimination and
7. Right to Equality Before the Law
These systems may reproduce and perpetuate biases in the training data, but other evidence suggests they may reduce racial and other disparities in bail and sentencing nonetheless.

3. Right to Life, Liberty, Personal Security
Low-risk individuals may benefit from greater pre-trial release and shorter sentences, and the community might benefit from a lower crime rate.

9. Freedom from Arbitrary Arrest, Detention, and Exile
Complex algorithms may erroneously classify certain individuals as "high-risk," raising the possibility of arbitrary pre-trial or post-conviction detention.

10. Right to Fair Public Hearing and
11. Right to be Considered Innocent until Proven Guilty
The proprietary nature of these tools, their inherent complexity, and the inscrutability of the results they produce makes it hard to challenge them in court.

12. Right to Privacy
Automated risk assessment systems are premised on the collection, storage, and analysis of vast amounts of personal data, which raises significant privacy concerns.

🔵 *Positive human rights impact* 🟡 *Human rights impact indeterminate* 🔴 *Negative human rights impact*

The criminal justice system is the most potent and fearsome institution through which democratic societies may restrict an individual's enjoyment of their fundamental human rights. In view of the severity of its impacts on human rights, society has evolved a system of procedural rights to protect criminal defendants and convicts from the vagaries of human decision-making, from intentional abuse of power to unconscious influences ranging from racism to fatigue.[48]

In search of both fairness and efficiency, justice systems are increasingly employing automated decision-making tools at every procedural stage. This is especially true of risk assessments, which are used to inform decisions about pretrial detention, sentencing, and parole. To the extent that they are fair and accurate, risk assessment tools can have a significant positive impact on the rights of individuals accused and convicted of crimes. The corollary, however, is that flaws or unknown limitations in the operation of such systems can have deleterious effects on a wide range of rights.

[48] Millicent H. Abel and Heather Watters, "Attributions of Guilt and Punishment as Functions of Physical Attractiveness and Smiling," *The Journal of Social Psychology* 145, no. 6 (December 2005): 687–702, https://doi.org/10.3200/SOCP.145.6.687-703; Shai Danziger, Jonathan Levav, and Liora Avnaim-Pesso, "Extraneous Factors in Judicial Decisions," *Proceedings of the National Academy of Sciences of the United States of America* 108, no. 17 (April 26, 2011): 6889–92, https://doi.org/10.1073/pnas.1018033108.

Traditional Approach to Risk Assessments

The first efforts to formalize the process of assessing an individual's risk of recidivism date back to the 1920s, when statisticians began to identify objective factors that are predictive of this risk for parolees.[49] As with AI now, the force driving the development of these earlier tools was the desire to avoid unnecessary deprivations of liberty and reduce the incidence of discrimination in the criminal justice system attributable to human bias. Statisticians developed these tools by collecting and analyzing information about defendants to identify factors that distinguish those that reoffend from those who do not.

As these assessments became more sophisticated, statisticians began to consider both static factors, such as a defendant's age and gender, as well as dynamic factors, such as a defendant's skill set or psychological profile.[50] Over time, these efforts led to the development of risk assessment inventories such as the Level of Service Inventory-Revised ("LSI-R")[51] that, while developed and validated by statisticians, are deployed in the field by individuals without much if any statistical expertise. Especially when such tools require their operators to make subjective determinations, such as whether an individual is engaging in antisocial behavior,[52] these tools may suffer from low inter-rater reliability ("IRR"), calling into question the validity of the predictions generated by such tools for any given individual.[53] Furthermore, the data available to actuarial risk assessment systems to identify who is truly at a high risk of re-offending is systematically skewed by the fact that the pre-existing system has sentenced those it believes to pose the highest risk to long prison sentences, during which time those inmates cannot reoffend.[54]

Risk assessment tools in the U.S. criminal justice system have been critiqued as inherently unfair due to the disproportionate targeting of minority individuals and communities by the police.[55] This, in turn, raises the risk that such tools will miscalculate the risk of recidivism for individuals from minority versus majority communities. Moreover, as the Supreme Court of Canada recently noted in *Ewert v. Canada*, risk assessment tools that are developed and validated based on data from majority groups

[49] Bernard E. Harcourt, *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age.* (Chicago: University of Chicago Press, 2006) 48-60; James Bonta, "Risk-Needs Assessment and Treatment," in *Choosing Correctional Options That Work: Defining the Demand and Evaluating the Supply*, ed. Alan T. Harland (Thousand Oaks, Calif: Sage Publications, 1996); Thomas Mathiesen, "Selective Incapacitation Revisited," *Law and Human Behavior* 22, no. 4 (1998): 455–69.

[50] James Bonta, "Risk-Needs Assessment and Treatment."

[51] Ibid.

[52] Thomas H. Cohen, "Automating Risk Assessment Instruments and Reliability: Examining an Important but Neglected Area in Risk Assessment Research," *Criminology & Public Policy* 16, no. 1 (February 2017): 271–79, https://doi.org/10.1111/1745-9133.12272.

[53] In the risk assessment context, inter-rater reliability refers to the degree of agreement between distinct raters applying an assessment tool. A high IRR means raters apply the tool in the same manner as others; in other words, a high IRR means a particular defendant would receive the same score regardless of who conducted the assessment. A low IRR, in turn, would indicates raters may score the same defendant differently. Grant Duwe and Michael Rocque, "Effects of Automating Recidivism Risk Assessment on Reliability, Predictive Validity, and Return on Investment (ROI): Recidivism Risk Assessment," *Criminology & Public Policy* 16, no. 1 (February 2017): 235–69, https://doi.org/10.1111/1745-9133.12270.

[54] Shawn Bushway and Jeffrey Smith, "Sentencing Using Statistical Treatment Rules: What We Don't Know Can Hurt Us," *Journal of Quantitative Criminology* 23, no. 4 (December 1, 2007): 377–87, https://doi.org/10.1007/s10940-007-9035-1.

[55] Solon Barocas and Andrew D. Selbst, "Big Data's Disparate Impact," *California Law Review* 104, no. 3 (2016): 671–732, https://doi.org/10.15779/z38bg31.

may lack validity in predicting the same traits in minority groups.[56] This may have deleterious effects on the rehabilitation of offenders from minority communities by impacting their access to cultural programming and their opportunities for parole, among other things.[57]

The answer to the question of whether earlier generations of risk assessment tools have a positive or negative impact on the rights of criminal defendants and convicts to life, liberty, and security of person[58] is unclear. On one hand, they may represent an improvement over the situation where judges had essentially unfettered discretion regarding bail and sentencing decisions. On the other hand, the possibility of negative impacts exists due to the potential for the misclassification of some number of defendants as "high risk," which results in their being sentenced more harshly than they otherwise would, or should, have been. Such tools also adversely impact criminal defendants' rights to a fair public trial, to a defense, and to an appeal,[59] because their predictions are not subject to meaningful review by courts. Not only do courts lack the institutional capacity to review the operation of such tools, but the objective veneer that coats the outputs of these tools obscures the subjective determinations that are baked into them.

Furthermore, these tools raise fundamental questions as to whether it is fair to treat a particular individual more harshly simply because they share characteristics with others who have reoffended. This is a particularly serious difficulty when it comes to individuals who are classified as "high risk" yet for whatever reason do not reoffend. While statistical techniques can determine with a high degree of accuracy the characteristics of individuals in a population who are likely to behave in a certain way, they cannot generate accurate predictions as to how any particular individual in that population will behave. This raises some truly vexing legal, moral, and philosophical questions that are common to all the case studies that follow.

AI-Generated Risk Assessments

In recent years, criminal justice systems in many different countries have begun to use algorithmic risk assessment tools. All such tools automate the analysis of whatever data has been inputted into the system. Most of these tools still rely on manually-inputted data from questionnaires similar to those that were part and parcel of the last generation of risk-assessment tools, while newer tools are fully automated and rely on information that already exists in various government databases.[60]

Full automation improves the predictive accuracy and validity of risk assessment tools because the software interprets every piece of data consistently.[61] Automation also obviates the need for manual data collection, entry, and scoring, which carries with it the possibility of improving the accuracy of these

[56] Ewert v. Canada, 2018 SCC 30. Note, however, that other tools—such as the Ontario Domestic Assault Risk Assessment Tool ("ODARA")—are in widespread use in Canada and have been adopted by courts in several provinces and territories. For examples of courts relying on these tools, see R v. Beharri, 2015 ONSC 5900; R v. Primmer 2017 ONSC 2953; R v. Sassie, 2015 NWTCA 7; R. v. Robertson, 2006 ABPC 88.

[57] Ewert v. Canada, 2018 SCC 30.

[58] UDHR art. 3.

[59] UDHR arts. 10 and 11(1); ICCPR art. 14(5).

[60] The Minnesota Screening Tool Assessing Recidivism Risk 2.0 ("MnSTARR 2.0") under development by the government of the U.S. State of Minnesota is a leading example of a fully-automated risk assessment tool in the criminal justice context. Kenneth C. Land, "Automating Recidivism Risk Assessment: Should We Stay or Should We Go?," *Criminology & Public Policy* 16, no. 1 (February 2017): 231–33, https://doi.org/10.1111/1745-9133.12271.

[61] Ibid.

systems by, for example, allowing additional variables to be considered.[62]

Beyond full automation, the latest generation of risk assessment tools leverages machine learning techniques to continually rebalance risk factors in response to new inputs. In theory, the predictive power and accuracy of such systems should improve over time. This was the finding of a proof-of-concept study in New York City, where researchers used machine learning techniques to determine which criminal defendants should receive bail.[63] The study's results suggest that New York could reduce the number of people held in pretrial detention by 40% without any corresponding increase in the crime rate. Alternately, the city could reduce its crime rate by 25% by incarcerating the same number of people, but changing the criteria for who gets bail. In so doing, the number of African-Americans and Hispanics housed in the city's jails would be significantly reduced, with concomitant positive effects on the right to equality and non-discrimination.[64]

For all of these potential positives, the single most widely-used algorithmic risk assessment system in the United States has been accused of perpetuating racial bias. An investigation by ProPublica found that COMPAS, a proprietary risk-assessment system that certain U.S. state courts use in making bail and sentencing decisions, misclassified African-American offenders as "high-risk" at twice the rate of Caucasians, even though the system had nearly the same accuracy rate (63% vs. 59%) in predicting when individuals from both racial groups would reoffend.[65] In other words, COMPAS classified 45% of those African-American convicts who ultimately did not reoffend as "high risk," as compared to just 23% for similarly-situated Caucasians. Questions have been raised about the accuracy and the methodological validity of the ProPublica report,[66] but more fundamentally, an important paper published in the aftermath of the COMPAS controversy suggests that it may be well-nigh impossible to design algorithms that treat individuals belonging to different groups equally fairly across multiple different dimensions of fairness.[67] Assuming, however, that the issues ProPublica identified with COMPAS are well-founded and are true of other risk assessment algorithms, then there is a substantial risk that the rights of minority groups to equality and non-discrimination will be adversely affected by such tools.[68]

Furthermore, there is a serious issue relating to the existence of systematic patterns of bias against

[62] According to Barocas and Selbst, one source of bias is inaccuracies in the selected features. Additional features should, in theory, allow for more accurate generalizations to be developed. Barocas and Selbst, "Big Data's Disparate Impact."

[63] Jon Kleinberg et al., "Human Decisions and Machine Predictions" (Cambridge, MA: National Bureau of Economic Research, February 2017), https://doi.org/10.3386/w23180.

[64] UDHR art. 2.

[65] Jeff Larson and Julia Angwin, "How We Analyzed the COMPAS Recidivism Algorithm," *ProPublica*, May 23, 2016, https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

[66] For example, Anthony W Flores, Kristin Bechtel, and Christopher T. Lowenkamp, "False Positives, False Negatives, and False Analyses: A Rejoinder to 'Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks.'," *Federal Probation* 80, no. 2 (2016): 9.

[67] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan, "Inherent Trade-Offs in the Fair Determination of Risk Scores," *ArXiv:1609.05807 [Cs, Stat]*, September 19, 2016, http://arxiv.org/abs/1609.05807.

[68] UDHR art. 2.

minorities in the data being used to train these algorithmic risk-assessment tools, arising from the disproportionate police scrutiny that minority community members receive. Consequently, minority communities are over-represented in the training data, which results in variables that are close proxies for race being over-weighted by these algorithms in assessing the risk that any particular individual poses.[69] This, too, raises concerns about algorithmic risk assessment tools having negative impacts on the rights of minority groups to equality and non-discrimination.[70]

There are also issues that arise from the development of these risk assessment tools by private companies who, for commercial reasons, guard their algorithms and the data that is used to train them as trade secrets.[71] The secrecy that often surrounds the operation of these risk assessment tools can have adverse impacts on the rights of criminal defendants to defend themselves against criminal charges[72] and to appeal a conviction.[73] The situation is further complicated when risk assessment algorithms rely upon machine learning techniques to adapt their performance over time, as the results generated by such techniques are oftentimes neither reproducible nor explainable in any meaningful way.

Summary of Impacts

The current generation of automated risk-assessment tools has the potential to positively impact the rights of "low-risk" criminal defendants and offenders to life, liberty, and security of the person.[74] If indeed such tools are more accurate than humans at predicting the risk of recidivism,[75] low-risk offenders will end up being incarcerated at a lower rate and for shorter periods of time than under the status quo. Members of society at large will also be more secure in the enjoyment of their right to security of the person should these tools result in a lower rate of crime.

It is hard to know, however, whether the current generation of automated risk assessment tools is having a negative or positive impact on the equality and non-discrimination rights of criminal defendants from groups that have historically been discriminated against, such as ethnic minorities and the mentally ill.[76] While the existence of systemic biases in the training data may result in the automation of existing social biases against individuals from these groups, the results of the New York City proof-of-concept study suggest that such systems may nevertheless ameliorate the over-representation of individuals from these groups in jail and prison populations.

[69] Barocas and Selbst, "Big Data's Disparate Impact."

[70] UDHR art. 2.

[71] Rebecca Wexler, "Life, Liberty, and Trade Secrets," *Stanford Law Review* 70, no. 5 (2018): 1343.

[72] UDHR art. 11(1).

[73] ICCPR art. 14(5).

[74] UDHR art. 2.

[75] This assumption has been questioned. Julia Dressel and Hany Farid, "The Accuracy, Fairness, and Limits of Predicting Recidivism," *Science Advances* 4, no. 1 (January 1, 2018), https://doi.org/10.1126/sciadv.aao5580.

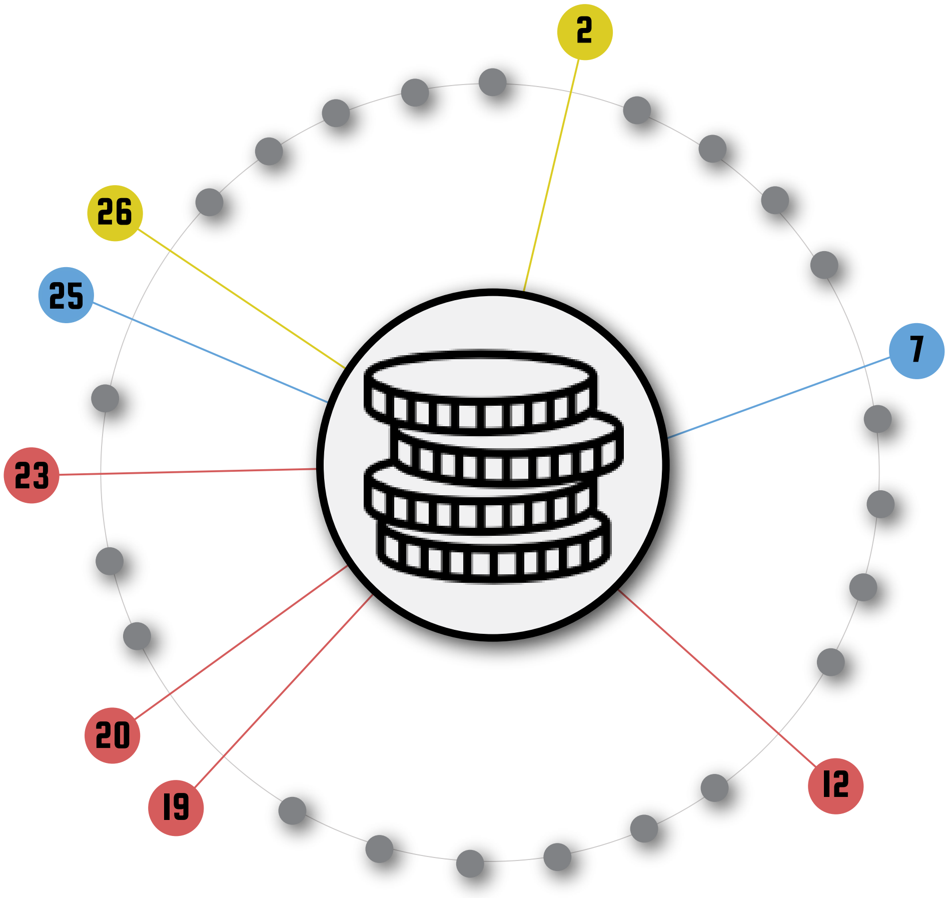
[76] UDHR art. 3.

Finally, in view of the inscrutability of the latest generation of automated risk assessment tools, and the secrecy surrounding these tools when they are developed by the private sector, we believe that these tools are likely to adversely impact the rights of criminal defendants to a fair and public hearing before an independent and impartial tribunal,[77] and to enjoy all of the guarantees needed for their defense.[78]

[77] UDHR art. 10.

[78] UDHR art. 11(1). Relatedly, the right of criminal convicts under ICCPR art. 14(5) is similarly impacted.

5.2 Access to the Financial System: Credit Scores



2. Freedom from Discrimination and
7. Right to Equality Before the Law
AI may reduce discrimination in lending by providing more accurate determinations of the creditworthiness of marginalized groups, yet it may also discriminate against them in novel ways.

12. Right to Privacy
AI-based credit scoring systems are premised on the collection, storage, and analysis of vast amounts of personal data, which raises significant privacy concerns.

19. Freedom of Opinion, Expression, and Information, and
20. Right of Peaceful Assembly and Association
Since "all data is credit data" for AI-generated credit scores, people may be chilled from expressing themselves or associating with certain others for fear of how this might impact their ability to borrow.

23. Right to Desirable Work,
25. Right to Adequate Standard of Living, and
26. Right to Education
AI will likely be used to extend credit to people who have been passed over by lenders using traditional credit scores, who can then use this money to improve their economic well-being.

🔵 *Positive human rights impact* 🟡 *Human rights impact indeterminate* 🔴 *Negative human rights impact*

Access to financial services such as banking and lending are an important means of promoting social and economic well-being. Access to credit in particular can help disadvantaged and marginalized individuals better enjoy their economic, social, and cultural rights by, for example, providing them with the means to pursue higher education,[79] access health care,[80] purchase property,[81] or start a business through which they can be gainfully employed.[82] In view of the role that credit can play in advancing the achievement of a wide range of human rights, the Nobel laureate Muhammad Yunus has suggested that access to credit itself ought to be considered a human right.[83]

[79] UDHR art. 26.

[80] UDHR art. 25.

[81] UDHR art. 17.

[82] UDHR art. 23.

[83] Matt Wade, "Access to Credit a 'Human Right', Says the Father of Microfinance," *Sydney Morning Herald,* October 9, 2014, https://www.smh.com.au/national/access-to-credit--a-human-right-says-the-father-of-microfinance-20141009-113j3x.html. For more on the debate, see the following two resources: Marek Hudon, "Should Access to Credit Be a Right?," *Journal of Business Ethics* 84, no. 1 (2009): 17–28 and John Gershman and Jonathan Morduch, "Credit Is Not a Right," in *Microfinance, Rights and Global Justice*, ed. Tom Sorell and Luis Cabrera (Cambridge: Cambridge University Press, 2015), 14–26, https://doi.org/10.1017/CBO9781316275634.002.

Traditional Approach to Credit Scoring

In deciding whether to extend someone credit, lenders have long sought to ascertain the prospective borrower's risk of defaulting on the debt. Such determinations have historically been of dubious accuracy and rife with the possibility of discrimination, as lenders based them on their personal impressions of the borrower coupled with references from the community.[84] Nor were these determinations improved much by the development of the first credit reports around the turn of the 20th century, which consisted of compilations of information about an individual's personal affairs that were subject to the discretionary review of the lender.[85]

In the United States, the legislative efforts of the 1970s to outlaw discrimination in lending based on race, religion, gender, age, and other similar traits roughly coincided with the development of the first credit scores, which attempted to reduce all of the information contained in an individual's credit score into a simple, numerical indication of that person's credit-worthiness.[86] Different companies use different approaches to calculate credit scores. FICO Scores, which are used by 90% of lenders in the United States, are generated based on a combination of an individual's payment history, the amount that they owe, the age of their accounts, their sources of credit, and how much additional credit they have sought recently.[87]

Despite their objective veneer, traditional credit scores suffer from several limitations that can adversely impact human rights. Since traditional credit scores rely on information gathered by credit bureaus about an individual's past financial history, oftentimes individuals with a "thin" credit file are given a credit score that is not indicative of their true risk of defaulting or are denied a credit score entirely.[88] Such "thin-file" borrowers tend to belong to marginalized groups such as minorities, young adults, immigrants, and recently-divorced women.[89] Since financial institutions are less likely to lend to individuals from these groups, even when in reality they are just as credit-worthy as "thick-file" applicants from other groups, the right to equality may be adversely impacted.[90]

There are also issues relating to the fairness and accuracy of the data being fed into credit-scoring algorithms. In the United States at least, credit bu-

[84] Matthew A. Bruckner, "The Promise and Perils of Algorithmic Lenders' Use of Big Data," *Chicago-Kent Law Review* 93, no. 1 (March 9, 2018): 2–60.

[85] Sean Trainor, "The Long, Twisted History of Your Credit Score," *Time*, accessed June 10, 2018, http://time.com/3961676/history-credit-scores/.

[86] Those laws are the Fair Credit Reporting Act ("FCRA") of 1970, the Equal Credit Opportunity Act ("ECOA") of 1974 and the Community Reinvestment Act of 1977. Willy E. Rice, "Race, Gender, Redlining, and the Discriminatory Access to Loans, Credit, and Insurance: An Historical and Empirical Analysis of Consumers Who Sued Lenders and Insurers in Federal and State Courts, 1950-1995," *San Diego Law Review* 33 (1996): 583–700.

[87] Rob Kaufman, "5 Factors That Determine a FICO Score," *myFICO* (blog), September 23, 2016, https://blog.myfico.com/5-factors-determine-fico-score/.

[88] Kenneth P. Brevoort, Philip Grimm, and Michelle Kambara, "Data Point: Credit Invisibles" (Consumer Financial Protection Bureau Office of Research, May 2015).

[89] Bruckner, "The Promise and Perils of Algorithmic Lenders' Use of Big Data." ("In other words, credit invisibles are generally either too young to have established a credit history, or have never been welcomed into the traditional banking system. As such, [algorithmic credit scores] would especially benefit the young, the low-income, and minorities." (internal citations omitted)).

[90] UDHR art. 2. Brevoort, Grimm, and Kambara, "Data Point: Credit Invisibles."

reaus rely on "furnishers"—banks, utilities, and other businesses—to voluntarily report relevant information, such as on-time payments, debt balances, and the like.[91] In view of the legal obligations that attach to furnishers when they provide information to a credit bureau, such businesses are more likely to report adverse events (such as a missed payment or foreclosure) that negatively impact its own bottom line, as opposed to routine, unremarkable positive events (such as timely payments).[92] Since individuals from minority communities suffer from adverse financial events (such as evictions) at a higher rate than would be predicted by their actual financial circumstances,[93] there is a significant risk that the information used to generate credit scores is systematically biased against minority communities.

Furthermore, even if all relevant information (both positive and negative) is reported to a credit agency, there is no guarantee that the credit scoring algorithm will consider it. For example, the FICO score in wide use in the United States considers only mortgage and credit-card payment history, but not rental or bill payment history.[94] In view of the long legacy of discriminatory lending policies in the U.S. and of housing policies that make it much more likely that individuals from minority groups will rent rather than own a home,[95] these practices can have significant discriminatory impacts.[96]

The growing use of credit scores beyond the lending context amplifies these effects. It is increasingly common for employers, landlords, and insurers to review an individual's credit score before offering them a job, renting them an apartment, or selling them insurance.[97] Employers may think that credit scores are a proxy for an applicant's integrity and responsibility, even though they have not been validated for that purpose.[98] Insurers may similarly view those with poor credit scores as posing a higher actuarial risk because "recklessness" in paying down one's debts shows the individual to be a reckless person in general.[99] Yet again, such practices pose a grave risk of perpetuating and amplifying age-old patterns of inequality and discrimination that bear little resemblance to reality.

[91] "Report to Congress Under Section 319 of the Fair and Accurate Credit Transactions Act of 2003" (Federal Trade Commission, January 2015).

[92] Jocelyn Baird, "What Gets Reported to Your Credit Reports (and What Doesn't)?,"*NextAdvisor* (blog), accessed June 20, 2018, https://www.nextadvisor.com/blog/what-gets-reported-to-your-credit-reports/.

[93] Deena Greenberg, Carl Gershenson, and Matthew Desmond, "Discrimination in Evictions: Empirical Evidence and Legal Challenges," *Harvard Civil Rights* 51, no. 1 (2016): 44.

[94] Preeti Vissa, "How Credit Scores Disproportionately Hurt Communities of Color," *Huffington Post* (blog), December 15, 2010, https://www.huffingtonpost.com/preeti-vissa/credit-scores-and-the-for_b_797148.html; "How Credit History Impacts Your FICO® Score," *myFICO* (blog), accessed June 20, 2018, http://www.myfico.com/credit-education/credit-payment-history.

[95] Christopher E. Hebert et al., "Homeownership Gaps Among Low-Income and Minority Borrowers and Neighborhoods" (U.S. Department of Housing and Urban Development, March 2005); Sarah Ludwig, "Credit Scores in America Perpetuate Racial Injustice. Here's How," *The Guardian*, October 13, 2015, sec. Opinion, http://www.theguardian.com/commentisfree/2015/oct/13/your-credit-score-is-racist-heres-why.

[96] UDHR art. 3.

[97] "Past Imperfect: How Credit Scores and Other Analytics 'Bake In' and Perpetuate Past Discrimination" (National Consumer Law Center, May 2016) ("Credit history is used as a gatekeeper for many important necessities – employment, housing (both rental and homeownership), insurance, and of course, affordable credit.").

[98] Gary Rivlin, "Employers Pull Applicants' Credit Reports," *The New York Times*, May 11, 2013, sec. Business Day, https://www.nytimes.com/2013/05/12/business/employers-pull-applicants-credit-reports.html.

[99] "Past Imperfect: How Credit Scores and Other Analytics 'Bake In' and Perpetuate Past Discrimination."

AI-Generated Credit Scores

In recent years, lenders have begun to use artificial intelligence to more accurately assess whether a potential borrower is a good credit risk. Unlike conventional credit scoring algorithms, the AI-based approach treats "all data as credit data" and analyzes vast amounts of data from many sources.[100] The resulting AI-generated credit scores are better than traditional scores at addressing some kinds of situations, at the same time as they create new challenges of their own.

The volume of data that AI-based credit scoring systems collect and analyze is so staggering as to be concerning. ZestFinance, one of the leading companies in this field in the US, considers over 3,000 variables in deciding whether to offer someone credit[101]—including whether the applicant tends to type in all-caps, which apparently is correlated with a higher risk of default.[102] Lenddo, another American company in this space, examines an applicant's entire digital footprint—including social media use, geolocation, website browsing habits, phone use history (including text and call logs), purchasing behavior, and more in deciding whether to extend them credit.[103]

AI-generated credit scores have particularly significant applications in emerging markets, where almost everyone is a "thin-file" borrower. For example, the MyBucks Haraka app in use in India, the Philippines, and several Sub-Saharan countries uses data gleaned from an applicant's mobile phone (call logs, geolocation information, and the like) and their social media accounts to generate an alternative credit score that partner banks can use to inform their lending decision.[104] This AI-based approach has the potential to help members of historically marginalized groups, such as women and ethnic minorities, gain access to credit in the developed and developing world alike,[105] thereby fostering financial inclusion and advancing the right to equality.[106]

Early results suggest that these technologies are succeeding in fostering financial inclusion. ZestFinance claims that its AI-based technology allowed it to reduce its default rate to less than half of the prevailing industry average,[107] while Lenddo claims to have increased its approval rate by 15% while slashing defaults by 12%.[108] If these early results are accurate and generalizable, the positive impact on all of the economic, cultural, and social rights that access to credit enables would be very significant indeed.

[100] James Rufus Koren, "Some Lenders Are Judging You on Much More than Finances," *Los Angeles Times*, December 19, 2015, http://www.latimes.com/business/la-fi-new-credit-score-20151220-story.html.

[101] "Zest Automated Machine Learning Data Sheet" (ZestFinance), accessed June 20, 2018, https://www.zestfinance.com/hubfs/Underwriting/Zest-Automated-Machine-Learning-Data-Sheet.pdf?hsLang=en.

[102] Koren, "Some Lenders Are Judging You on Much More than Finances."

[103] "Credit Scoring: The LenddoScore Fact Sheet" (Lenddo), accessed June 20, 2018, https://www.lenddo.com/pdfs/Lenddo_FS_CreditScoring_201705.pdf.

[104] Penny Crosman, "This Lender Is Using AI to Make Loans through Social Media," *American Banker*, December 8, 2017, https://www.americanbanker.com/news/this-lender-is-using-ai-to-make-loans-through-social-media.

[105] Brevoort, Grimm, and Kambara, "Data Point: Credit Invisibles."; Geri Stengel, "How One Woman Is Changing Business Lending In Africa," *Forbes*, January 14, 2015, https://www.forbes.com/sites/geristengel/2015/01/14/how-one-woman-is-changing-business-lending-in-africa.

[106] UDHR art. 2.

[107] John Lippert, "ZestFinance Issues Small, High-Rate Loans, Uses Big Data to Weed out Deadbeats," *The Washington Post*, October 11, 2014, sec. Business, https://www.washingtonpost.com/business/zestfinance-issues-small-high-rate-loans-uses-big-data-to-weed-out-deadbeats/2014/10/10/e34986b6-4d71-11e4-aa5e-7153e466a02d_story.html.

[108] "Credit Scoring: The LenddoScore Fact Sheet."

Yet there are considerable risks to this new approach as well. One arises from the quality and accuracy of the data used to train these systems, as well as the fairness and accuracy of the data these systems use to decide upon a particular individual's application for credit. The issues are similar in nature to those affecting traditional credit scoring algorithms, but they are different in degree due to the vast number of data sources that AI-based algorithms take into consideration.

Another arises from the subjective decisions that programmers make on how to code and categorize the data that they feed into their seemingly-objective algorithms.[109] For example, ZestFinance translates certain continuous variables (such as the length of time one spends reading their website's terms and conditions) into categorical values (like 0, 1, or 2).[110] This is an inherently subjective process that can introduce explicit or implicit bias into the data and consequently into the results generated by the algorithm.

Furthermore, AI-generated scores may perpetuate existing patterns of discrimination through "network discrimination,"[111] whereby individuals are penalized (or rewarded) based on the characteristics of others who are in their personal network. For example, if there are two individuals in an identical financial position, yet the first individual's friends live in "rich" neighborhoods while the second's friends live in "poor" neighborhoods, an algorithm may well determine the first to be a better credit risk than the second.[112] To the extent that such network factors correlate with invidious classifications such as those based on race and gender, the potential for discriminatory impacts is quite serious indeed.[113]

The use of AI in financial decision-making may even burden individuals' freedom of opinion, expression, and association by chilling individuals from engaging in activities that they believe will negatively affect their credit score. This is not a mere theoretical possibility. In 2009, American Express reduced the credit limit of an African-American businessman because "[o]ther customers who have used their card at establishments where [he] recently shopped . . . ha[d] a poor repayment history."[114] Another lender in the U.S. reduced the credit limit of its customers who had incurred expenses at "marriage counselors, tire retreading and repair shops, bars and nightclubs, pool halls, pawn shops, massage parlors, and others."[115]

An even more extreme example of this phenomenon is China's incipient "social credit score" system, which generates a numerical index of an individual's "trustworthiness" based on a vast array of data points, including social media data, arrest

[109] Mikella Hurley and Julius Adebayo, "Credit Scoring in the Era of Big Data," *Yale Journal of Law & Technology* 18, no. 1 (2016): 148–216.

[110] Ibid.

[111] danah boyd, Karen Levy, and Alice Marwick, "The Networked Nature of Algorithmic Discrimination," in *Data and Discrimination: Collected Essays*, ed. Seeta Peña Gangadharan (New America, 2014), 53–57.

[112] Kaveh Waddell, "How Algorithms Can Bring Down Minorities' Credit Scores," *The Atlantic*, December 2, 2016, https://www.theatlantic.com/technology/archive/2016/12/how-algorithms-can-bring-down-minorities-credit-scores/509333/.

[113] UDHR arts. 2 and 20.

[114] Ron Lieber, "American Express Watched Where You Shopped," *The New York Times*, January 30, 2009, sec. Your Money, https://www.nytimes.com/2009/01/31/your-money/credit-and-debit-cards/31money.html.

[115] Ibid.

and infraction records, volunteer activity, city and neighborhood records, and more.[116] Those with high social credit scores enjoy benefits such as lower utility rates and more favorable borrowing conditions, while those with unfavorable scores might be unable to purchase airline or high speed rail tickets.[117] These systems are being piloted in several communities, but a national roll-out of the "social credit score" system is expected by 2020. While anecdotal reports suggest that the "social credit scoring" system has curbed corruption and incentivized certain forms of good behavior, such as stopping for pedestrians at crosswalks,[118] it is not hard to imagine how this system could chill a great deal of expressive and associative activity.[119]

Summary of Impacts

Compared to the status quo credit scoring algorithms, the introduction of AI into the lending process is likely to have an overall positive impact on the ability of objectively low-risk borrowers to access credit. This is likely to have positive impacts on the enjoyment by these individuals of the right to an adequate standard of living,[120] the right to work,[121] and the right to education,[122] as access to credit is a powerful enabler of these economic and social rights.

The introduction of AI into the lending process is also likely to have a positive impact on the right to equality and non-discrimination for some individuals, while adversely affecting it for others. On the positive side, the fact that AI-based algorithms consider a wide variety of data sources may improve the ability of well-qualified individuals from marginalized communities to access credit by overcoming the "thin-file" problem. On the other hand, the specter of "network discrimination" having a negative impact on the ability of members of these very same communities to borrow money cannot be discounted.

Finally, it is likely that AI-based decision-making algorithms in the financial sector will adversely impact the freedoms of opinion, expression, and association. In an era where "all data is credit data," individuals may feel chilled from expressing certain points of view or associating with others, out of fear that an algorithm may use their behavior against them in the financial context.

[116] Mara Hvistendahl, "In China, a Three-Digit Score Could Dictate Your Place in Society," *WIRED*, Dec. 14, 2017, https://www.wired.com/story/age-of-social-credit/.

[117] Simina Mistreanu, "Life Inside China's Social Credit Laboratory," *Foreign Policy*, April 3, 2018, https://foreignpolicy.com/2018/04/03/life-inside-chinas-social-credit-laboratory/.

[118] Ibid.

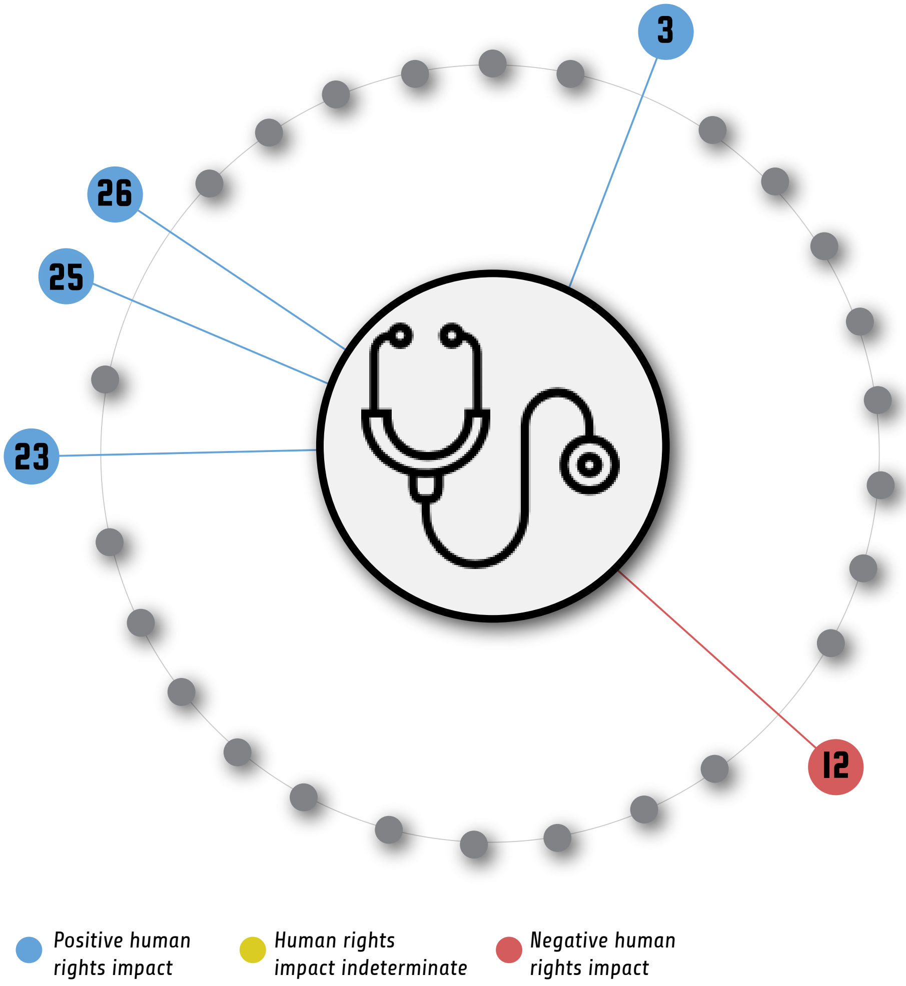
[119] UDHR arts. 19 and 20. In this context, the U.N. Human Rights Committee has noted that it is "impermissible" for states to engage in conduct that create "chilling effects that may unduly restrict the exercise of freedom of expression...." United Nations Human Rights Committee, General Comment 34 (ICCPR Art. 19: Freedoms of opinion and expression) (2011), U.N. Doc. CCPR/C/GC/34, p. 12.

[120] UDHR art. 17.

[121] UDHR art. 23.

[122] UDHR art. 26.

5.3 Healthcare: Diagnostics


3
26
25
23
12
● Positive human rights impact
● Human rights impact indeterminate
● Negative human rights impact

3. Right to Life, Liberty, and Security of Person
AI-based diagnostic systems enhance the enjoyment of the right to life by making accurate, high-quality diagnostic services more widely available.

12. Right to Privacy
AI-based diagnostic systems require the collection of vast quantities of sensitive data relating to an individual's often-immutable health characteristics, raising serious privacy concerns.

23. Right to Desirable Work,
The improved health outcomes that AI-based diagnostic systems are likely to produce will reduce the number of people who are excluded from the dignity of work for medical reasons.

25. Right to Adequate Standard of Living, and
By detecting diseases earlier and more accurately, AI-based diagnostic systems will improve living standards and quality of life.

26. Right to Education
Should AI-based diagnostic systems deliver on their promise, fewer people will be excluded from the enjoyment of the right to the education for reasons of ill-health.

In the course of the last century, modern medicine has produced astonishing improvements in the length and the quality of the lives of all those who can access it. Not only does the ICESCR recognize "the right of everyone to the enjoyment of the highest attainable standard of physical and mental health,"[123] but good health is arguably a necessary condition for each and every one of us to enjoy the full range of human rights that we are guaranteed by law.

Recent advances in health outcomes are attributable to improvements in the three pillars of healthcare: prevention, diagnosis, and treatment. AI has applications across all three pillars, but its greatest impact to date has been on improving the accuracy of medical diagnosis.

[123] ICESCR art. 12.

Traditional Approach to Diagnostics

Physicians use a wide range of approaches to diagnose disease. Perhaps the simplest and most widespread is to identify the patient's symptoms and correlate them to conditions or diseases that are characterized by the same pattern of symptoms.[124] The same basic approach can be applied to interpreting the results of diagnostic tests: a radiologist reviewing MRI imagery or a pathologist analyzing a biopsy sample compare what they are seeing to what they have learned in order to make a diagnosis.

Needless to say, it takes years of training and years more of experience to develop the knowledge and mastery required to accurately diagnose the wide range of maladies that afflict our species. To simplify matters, physicians often rely on "diagnostic criteria" in determining what ails someone. These are essentially statistically-validated rules of thumb that can be used to rule in or rule out a particular condition.[125] By contrast, experts in particular diseases engage in *gestalt* pattern recognition to recognize the characteristic indicators of a particular disease in a sea of information.[126]

Unfortunately, errors in diagnostics are extremely common, and they can have life-and-death consequences. One recent study found that 5% of patients in the U.S. are misdiagnosed every year,[127] while another found that misdiagnosis is the cause of 10% of patient deaths.[128] The challenge for physicians is growing as the number of diagnostic tests and procedures multiplies with the advance of medical science. Since each of these procedures has its own unique operating parameters and error rates,[129] it is becoming increasingly difficult for the average medical practitioner to choose the right test for their patient—or to even refer their patients to the right sub-specialists in view of their symptoms.[130]

AI-Assisted Diagnostics

Medical diagnostics is one of the fields in which AI-based technologies went into widespread use. Efforts began in the 1970s to start codifying the knowledge of human diagnostic experts into automated "expert systems."[131] These systems, which are known as "diagnostic decision support systems" and are used in many healthcare settings today, require the human clinician to answer a series of questions

[124] Committee on Diagnostic Error in Health Care et al., *Improving Diagnosis in Health Care*, ed. Erin P. Balogh, Bryan T. Miller, and John R. Ball (Washington, D.C.: National Academies Press, 2015), https://doi.org/10.17226/21794.

[125] The "Centor Criteria" for diagnosing strep throat in adults is an example of this. Since roughly 50% of patients who have all five of the criteria (cough, tonsillar exudates, swollen lymphatic nodes, fever, neither young nor old) will turn out to have strep throat, the Centor criteria presents a quick and easy way to rule in or rule out strep throat as a possibility in a patient presenting with these symptoms. Robert M. Centor et al., "The Diagnosis of Strep Throat in Adults in the Emergency Room," *Medical Decision Making* 1, no. 3 (August 1981): 239–46, https://doi.org/10.1177/0272989X8100100304.

[126] John P. Langlois, "Making a Diagnosis," in *Fundamentals of Clinical Practice*, ed. Mark B Mengel, Warren Lee Holleman, and Scott A Fields, 2nd ed. (New York: Kluwer Academic/Plenum Publishers, 2005).

[127] Hardeep Singh, Ashley N D Meyer, and Eric J Thomas, "The Frequency of Diagnostic Errors in Outpatient Care: Estimations from Three Large Observational Studies Involving US Adult Populations," *BMJ Quality & Safety* 23, no. 9 (September 2014): 727–31, https://doi.org/10.1136/bmjqs-2013-002627.

[128] Ibid.

[129] Committee on Diagnostic Error in Health Care et al., *Improving Diagnosis in Health Care*.

[130] J. Hickner et al., "Primary Care Physicians' Challenges in Ordering Clinical Laboratory Tests and Interpreting Results," *The Journal of the American Board of Family Medicine* 27, no. 2 (March 1, 2014): 268–74, https://doi.org/10.3122/jabfm.2014.02.130104.

[131] MYCIN was one of the pioneer expert systems developed at Stanford starting in 1972. Nancy McCauley and Mohammad Ala, "The Use of Expert Systems in the Healthcare Industry," *Information & Management* 22, no. 4 (April 1, 1992): 227–35, https://doi.org/10.1016/0378-7206(92)90025-B.

about the patient's condition that help to rule in or rule out certain specific diagnoses.

In the last several years, systems based on machine learning or deep learning have begun to be developed to facilitate and automate the diagnosis of illness across a range of medical specialties. Few of these technologies are currently in use, though early results suggest that they have great promise in improving the accuracy of medical diagnosis.

For example, an AI-powered image recognition system was able to detect cancerous skin lesions correctly 72% of the time, whereas human dermatologists correctly diagnosed the cancers 66% of the time.[132] There are also anecdotes about AI-powered diagnostic systems quickly solving intractable mysteries. For example, a diagnostic system powered by IBM's Watson was able to diagnose a patient as possessing a rare form of leukemia within 10 minutes, even though her symptoms had stumped experts for several months.[133] The system did so by comparing information in the patient's medical records with over 20 million oncology records held by the University of Tokyo. To be sure, physicians currently outperform AI systems on a wide variety of diagnostic tasks—from microscopy to general diagnosis. Yet it is impressive that AI systems rival or outperform human experts in diagnosing conditions ranging from brain cancer to autism and Alzheimer's disease.[134]

AI-based diagnostic systems also have the potential to provide greater access to specialist-level treatment than is currently possible. One of the few AI-based diagnostic systems to be approved for clinical use by the U.S. Food and Drug Administration is able to detect and diagnose diabetic retinopathy (a disorder affecting the vision of individuals suffering from diabetes) autonomously.[135] Whereas this condition was previously one that could only be diagnosed by a specialist, AI makes it possible for anyone trained in using the machinery to do so.

Summary of Impacts

AI-based diagnostic systems, especially the latest generation of systems that leverage artificial intelligence, are very likely to positively impact the right each of us enjoys to the highest attainable standard of health.[136] Not only do AI-based diagnostic systems appear to meet or exceed the performance of human experts in diagnosing disease, they have the potential to be much more accessible than specialized human experts, who require years of training and experience to rival the accuracy of an AI.

It is significant that in recognizing the right that each of us possesses "to a standard of living adequate for the health and well-being of himself and of his family," Article 25 of the UDHR links access to medical care to the basic requisites of life, such as food, clothing, and housing. In view of this link

[132] Siddhartha Mukherjee, "A.I. Versus M.D.," *The New Yorker*, March 27, 2017, https://www.newyorker.com/magazine/2017/04/03/ai-versus-md.

[133] James Billington, "IBM's Watson Cracks Medical Mystery with Life-Saving Diagnosis for Patient Who Baffled Doctors," International Business Times UK, August 8, 2016, https://www.ibtimes.co.uk/ibms-watson-cracks-medical-mystery-life-saving-diagnosis-patient-who-baffled-doctors-1574963.

[134] "AI vs Doctors," IEEE Spectrum: Technology, Engineering, and Science News, September 26, 2017, https://spectrum.ieee.org/static/ai-vs-doctors.

[135] Angela Chen, "AI Software That Helps Doctors Diagnose like Specialists Is Approved by FDA," *The Verge*, April 11, 2018, https://www.theverge.com/2018/4/11/17224984/artificial-intelligence-idxdr-fda-eye-disease-diabetic-rethinopathy.

[136] ICESCR art. 12.

between good health, access to health care, and the full range of economic, social, and cultural rights that each of us enjoys, the use of AI in medical diagnostics is likely to have positive impacts on the right of each of us to work and in so doing, ensure ourselves an existence worthy of human dignity.[137] Likewise, the better health outcomes that AI-based diagnostics are likely to produce will positively impact the enjoyment of the right to education by those who would otherwise be excluded by reasons of illness.[138]

As with many other automated technologies, there is the possibility that AI-based diagnostic technologies will cause employment losses in the medical field. While the right to work does not entail the right to work in any particular position, occupation, or field, the state obligation to protect the right to work and progressively adopt measures to realize full employment could be burdened by the widespread adoption of AI-based technologies that displace workers.[139] Indeed, there is already evidence that the impressive performance of AI-based diagnostic systems is leading medical students to shy away from entering certain specialty fields, such as radiology, where AI systems routinely outperform humans.[140]

Furthermore, the gathering of personal data necessary to create AI-powered tools creates particularly acute privacy risks in the healthcare context. In order to train the algorithms, healthcare providers must collect a vast range of intensely personal health and genetic data. The scope for the misuse of this data is vast—especially since an individual's genetic and health characteristics are often immutable—with potential implications for privacy,[141] dignitary rights,[142] freedom from discrimination,[143] and fair criminal procedure.[144] For example, such data could be used to deny a person health coverage on the basis of genetic factors that are beyond their control.[145] Or such data might be appropriated by the government for law enforcement purposes, as in the recent case from California of a 1970s-era serial killer who was identified based on the statistical analysis of DNA samples that his distant relatives submitted to a family ancestry website.[146]

Going further, one can argue that the fundamental right to life may be positively impacted by the introduction of AI diagnostic systems, which hold the promise of not only reducing the rate of diagnostic errors, but making high quality diagnostic services cheaper or more widely available. Although the right to life is generally viewed as a protection against the arbitrary deprivation of life by the

[137] UDHR art. 23.

[138] UDHR art. 26.

[139] UN Committee on Economic, Social and Cultural Rights ("CESCR"), General Comment No. 18: The Right to Work (Art. 6 of the Covenant), 6 February 2006, UN Doc. E/C.12/GC/18.

[140] Thomas H. Davenport and D. O. Keith J. Dreyer, "AI Will Change Radiology, but It Won't Replace Radiologists," Harvard Business Review, March 27, 2018, https://hbr.org/2018/03/ai-will-change-radiology-but-it-wont-replace-radiologists.

[141] UDHR art. 12.

[142] UDHR art. 1.

[143] UDHR art. 7.

[144] UDHR art. 10. In particular, it raises questions of self-incrimination, as protected by ICCPR art. 14(3)(g).

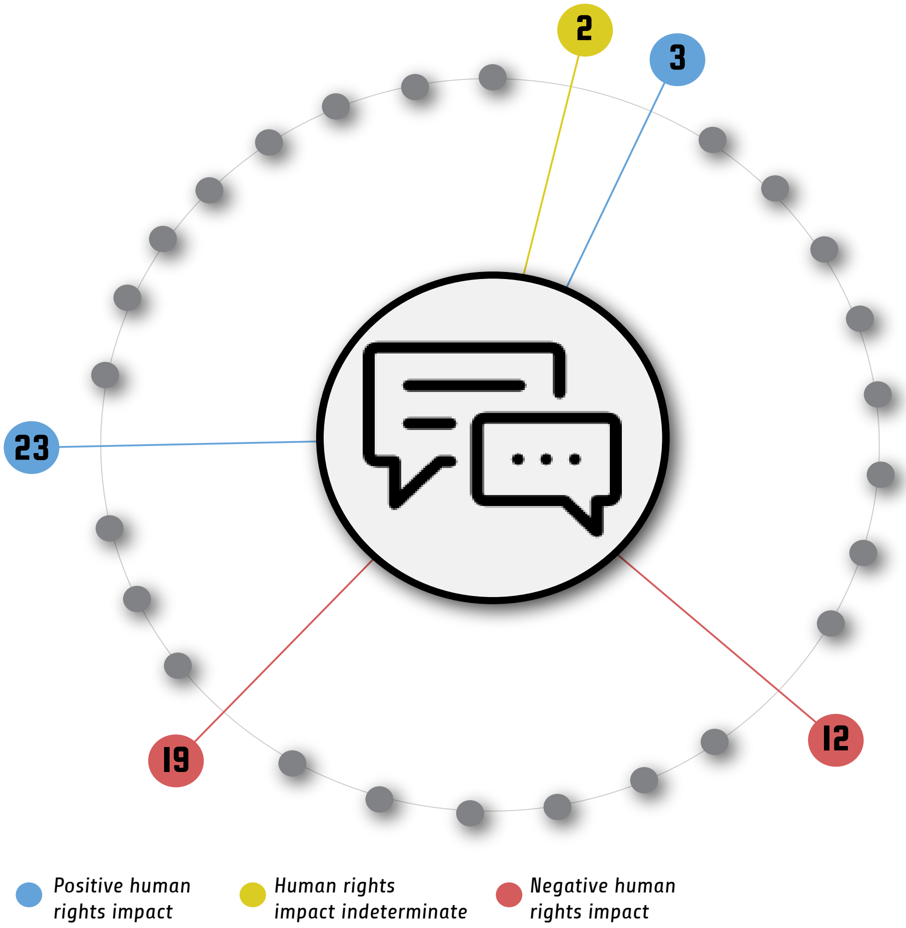
[145] For discussion on attempts to regulate genetic discrimination in the United States, see Louise Slaughter, "Genetic Information Non-Discrimination Act", *Harvard J. on Legislation* 50, no. 1 (2013): 41.

[146] Thomas Fuller, "How a Genealogy Site Led to the Front Door of the Golden State Killer Suspect," *The New York Times*, April 26, 2018, sec. United States, https://www.nytimes.com/2018/04/26/us/golden-state-killer.html.

state, the Supreme Court of Canada has ruled that inadequate access to medical care can result in deprivations of the right to life.[147] Correspondingly, improvements in the availability of high-quality medical services can be viewed as enhancing the right to life.

[147] Chaoulli v. Quebec (Attorney General), 2005 SCC 35.

5.4 Online Content Moderation: Standards Enforcement


2
3
23
12
19
● Positive human rights impact
● Human rights impact indeterminate
● Negative human rights impact

2. Freedom from Discrimination
AI systems may replicate the biases that human reviewers appear to show in reviewing content posted by marginalized groups, though they could be trained to avoid doing so.

3. Right to Life, Liberty, and Security of Person
AI-based content moderation systems are better than humans at finding content that is per se unlawful, such as child pornography, thereby enhancing community safety.

12. Right to Privacy
Privacy may be impacted by AI content moderation systems that automatically scan non-public communications for material that violates the law or the policies of a platform.

19. Freedom of Opinion, Expression, and Information
Current AI-based content moderation systems have higher error rates than humans, which may lead to large volumes of lawful content being erroneously removed.

23. Right to Desirable Work,
AI-based content moderation systems may free humans from the psychological toll that comes from policing online platforms for graphic, disturbing content.

The sheer amount of information that is available online has mostly been a blessing for humanity, though sometimes it can be a curse. On the one hand, never has so much information about so many different topics been available to most anyone, anywhere, who is fortunate enough to have an Internet connection. On the other hand, the dark side of humanity is also plain to view on the Internet. By virtue of the volume of what is available online, there is a substantial amount of content that is racist, sexist, gruesome, or harmful in other ways—such as by fomenting violence against identifiable groups or targeting individuals for bullying or harassment.

Some of the objectionable content online is subject to regulation by governments in conformity with international human rights law. Article 19(3) of the ICCPR recognizes that the right to free expression may be subject to certain exceptions provided by law that are necessary to protect the rights and reputations of others, or to protect national security, public order, public health, and morals. Moreover, Article 20 of the ICCPR expressly requires states to prohibit "propaganda for war" and the advocacy of "national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence."

Beyond expression that is unlawful and therefore subject to bona fide government regulation, there is also the problem of content that may be lawful but is nonetheless undesirable—either because it is not being posted in an appropriate place online, or because it violates the standards of an established online community. This is precisely the context in which the private companies that operate the Internet platforms that house so much of the world's online content promulgate standards as to what is acceptable and what is not, and engage in the "virtue of moderating"[148] materials that are inconsistent with those standards.

Traditional Approach to Content Standards Enforcement

The setting and enforcement of content standards by private companies is a controversial topic. Some liken it to a form of censorship due to the burdens it places on the rights to free expression, thought, and association.[149] Indeed, some commentators have noted that the largest online platforms, such as Facebook and Google, exercise more power over our right to free expression than any court, king, or president ever has[150]—in view of the very significant percentage of human discourse that occurs within the boundaries of these "walled gardens."[151]

By the same token, however, the failure of companies to adequately deal with online content that is harmful to others places its own burden on human rights. Companies therefore face the unenviable challenge of balancing between different rights belonging to different rights-holders, all the while remaining mindful of their responsibility to respect human rights in so doing.[152]

To discharge this difficult task with dispatch while avoiding discriminatory or speech-chilling outcomes, companies communicate their content guidelines to the public on their websites, at the same time as they have developed detailed internal guidance documents for their employees on when particular forms of content are subject to removal.[153] Ideally, both the external and internal guidelines will be informed by the principles of international human rights law, both with regards to their substantive content and the process that they outline.[154]

Until recently, the primary means by which questionable content was brought to the attention of a company was through the efforts of individual users of the platform, who flagged content as unlawful or inappropriate. Human reviewers working for the company then assess the content against the guidelines and determine whether it should stay up

[148] James Grimmelmann, "The Virtues of Moderation," *Yale Journal of Law & Technology* 17, no. 1 (2015): 68.

[149] UDHR arts. 18, 19 and 20.

[150] Jeffrey Rosen, "The Deciders: Facebook, Google, and the Future of Privacy and Free Speech," in *Constitution 3.0: Freedom and Technological Change.*, ed. Jeffrey Rosen and Benjamin Wittes (Brookings Institution Press, 2013).

[151] Jonathan Zittrain, *The Future of the Internet and How to Stop It* (New Haven, [Conn.]: Yale University Press, 2008).

[152] Perhaps counter-intuitively, content moderation may be among the AI applications where the distributive effects of these technologies are most apparent. Absent government-established policies, companies will be tasked with choosing which rights and rights-holders are prioritized over others.

[153] Alexis C. Madrigal, "Inside Facebook's Fast-Growing Content-Moderation Effort," *The Atlantic*, February 7, 2018, https://www.theatlantic.com/technology/archive/2018/02/what-facebook-told-insiders-about-how-it-moderates-posts/552632/.

[154] Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Human Rights Council, U.N. Doc. A/HRC/38/35 (Apr. 6, 2018) (by David Kaye), available at http://ap.ohchr.org/documents/dpage_e.aspx?si=A/HRC/38/35.

or come down.[155] In view of the massive volume of content that the leading Internet platforms host, these companies now each employ thousands if not tens of thousands of individuals whose sole job is to determine the fate of content that has been flagged.

In recent years, companies have been coming under increased scrutiny both as to the substance of the standards they employ in adjudging content, and also for their decisions in particular cases. For example, Facebook's broader policy against the display of nudity on its platform drew controversy when it removed images of breast-feeding women and the infamous "napalm girl" photograph from the Vietnam War from its platform.[156] Facebook ultimately relented in the face of public pressure in both incidents, but that too raises further questions about the consistency of its application of policies that burden the right to free expression.

There are also growing concerns that company policies on acceptable content may discriminate against certain viewpoints or perspectives, usually in a manner that favors the powerful over the marginalized.[157] For example, Facebook permitted a U.S. Congressman to state his view that all radicalized Muslims should be "hunted" or "killed," whereas it banned activists associated with the Black Lives Matter movement from stating that "all white people are racist."[158] While these anecdotes are not necessarily indicative of a larger pattern of bias or discrimination, they do raise troubling questions about how well these companies are meeting their responsibility to respect the full range of human rights in this important operational area.

Finally, there is also the issue of how companies are coming under growing pressure from governments to comply with their local laws on a global basis,[159] even when these laws are inconsistent with international guarantees of free expression, the right to association, and other human rights.[160] To date, companies have attempted to blunt these efforts by complying with local laws on a local basis, such as by prophylactically blocking content that is unlawful in a particular jurisdiction while leaving it available everywhere else. A string of recent court decisions has called into question the continuing viability of this technique, however.[161]

Meanwhile, other governments are even beginning to use company terms of service as a way to act against content that is lawful under domestic and international law, yet undesirable in the eyes of public policymakers. Some in the human rights community have expressed concerns that this activity of "referring" content for removal constitutes an end run around well-established judicial procedures for removing unlawful content.[162]

[155] Brittan Heller, "What Mark Zuckerberg Gets Wrong—and Right—About Hate Speech," *WIRED*, May 2, 2018, https://www.wired.com/story/what-mark-zuckerberg-gets-wrongand-rightabout-hate-speech/.

[156] Kate Klonik, "Facebook Erred by Taking down the 'Napalm Girl' Photo. What Happens Next?," *Slate*, September 12, 2016, http://www.slate.com/articles/technology/future_tense/2016/09/facebook_erred_by_taking_down_the_napalm_girl_photo_what_happens_next.html.

[157] Julia Angwin and Hannes Grassegger, "Facebook's Secret Censorship Rules Protect White Men From Hate Speech But Not Black Children," *ProPublica*, June 28, 2017, https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms.

[158] Ibid.

[159] Alicia Solow-Niederman et al., "Here, There, or Everywhere?," Berkman Klein Center Working Paper, March 2017, http://blogs.harvard.edu/cyberlawclinic/files/2017/03/Here-There-or-Everywhere-2017-03-27.pdf.

[160] For example, see Jens-Henrik Jeppesen & Laura Blanco, "European Policymakers Continue Problematic Crackdown on Undesirable Online Speech," Center for Democracy and Technology (blog), Jan. 18, 2018, https://cdt.org/blog/european-policymakers-continue-problematic-crackdown-on-undesirable-online-speech/.

[161] For example, Google Inc. v. Equustek Solutions Inc. 2017 SCC 34.

[162] Jens-Henrik Jeppesen, "First Report on the EU Hate Speech Code of Conduct shows need for transparency, judicial oversight, and appeals," Center for Democracy and Technology (blog), Dec. 12, 2016, https://cdt.org/blog/first-report-eu-hate-speech-code-of-conduct-shows-need-trans-

AI-Assisted Content Standards Enforcement

As the volume of online content in need of moderation grows inexorably and exponentially, the major online platforms are making significant investments in developing AI systems to automate this task. One major impetus for so doing are recently-enacted laws that require companies to promptly remove content that violates national laws, at the risk of facing substantial penalties for noncompliance.[163] These technologies are still in their infancy, and most simply work to identify potentially problematic content for a human reviewer to evaluate. That being said, fully automated content removal systems have been used against content that is suspected of violating copyright for a number of years,[164] and there are indications that some Internet platforms are employing fully automated content review and removal systems for at least some purposes.[165]

The current generation of AI-based content review and removal systems is built on natural language processing ("NLP") technology. As things stand right now, NLP technologies are domain-specific: that is to say, they were only built to identify the particular types of content on which they were trained and nothing else.[166] Hence an NLP system that is trained to detect say, racist speech, is incapable of detecting violent content. What is more, even within a particular domain, NLP technologies are not sophisticated enough to understand all of the nuances of human speech. A system that is able to detect racist content in a blog post might not accurately identify such content in a tweet, which results in these technologies having a very substantial error rate. This has led skeptics, including Facebook CEO Mark Zuckerberg, to conclude that AI systems are not yet sophisticated enough to replace human reviewers.[167]

This, however, does not render AI technologies useless. The speed at which they can sift through content makes them a powerful tool to assist, rather than to replace, human reviewers by identifying content that appears to be suspect.[168] AI systems can also be used to study the evolution of hate speech to spot emerging trends, as the Anti-Defamation League is currently doing with its Online Hate Index.[169]

Summary of Impacts

Due to the higher error rates of existing AI-based content flagging systems as compared to human reviewers,[170] the use of these systems to automatically remove content that is suspected to violate the law or an online platform's community standards is likely to have a negative impact on the rights to free

parency-judicial-oversight-appeals/.

[163] Yascha Mounk, "Verboten," *The New Republic*, April 3, 2018, https://newrepublic.com/article/147364/verboten-germany-law-stopping-hate-speech-facebook-twitter.

[164] "How Content ID Works - YouTube Help," accessed June 21, 2018, https://support.google.com/youtube/answer/2797370?hl=en.

[165] Lizzie Dearden, "New Technology Can Detect ISIS Videos before They Are Uploaded," *The Independent*, February 12, 2018, http://www.independent.co.uk/news/uk/home-news/isis-videos-artificial-intelligence-propaganda-ai-home-office-islamic-state-radicalisation-asi-data-a8207246.html.

[166] Natasha Duarte, Emma Llanso, and Anna Loup, "Mixed Messages? The Limits of Automated Social Media Content Analysis," Center for Democracy & Technology (blog), November 2017, https://cdt.org/insight/mixed-messages-the-limits-of-automated-social-media-content-analysis/.

[167] Heller, "What Mark Zuckerberg Gets Wrong—and Right—About Hate Speech."

[168] Susan Wojcicki, "Expanding our work against abuse of our platform," YouTube Official Blog, Dec. 4, 2017, https://youtube.googleblog.com/2017/12/expanding-our-work-against-abuse-of-our.html.

[169] "The Online Hate Index," Anti-Defamation League, accessed June 18, 2018, https://www.adl.org/resources/reports/the-online-hate-index.

[170] Duarte, Llanso, and Loup, "Mixed Messages?"

expression, opinion, and information.[171] This is because such systems are likely to remove a significant volume of content that is lawful and consistent with the platform's own community standards. Such errors would deprive individuals of the opportunity to express themselves, and their audience from viewing the opinions those individuals have expressed, with few opportunities for recourse.[172] That said, future developments in AI could well have a positive impact on these rights if they result in a lower error rate than the systems and procedures that are currently in use.

In their current state, these systems have the potential to positively impact the rights to life, liberty, and security of person[173] by improving the detection and removal of content that incites terrorism or hatred or violence against vulnerable populations. For example, automated techniques have been quite effective at detecting child pornography with a low error rate, and some forms of terrorist content exhibit consistent patterns that facilitate their detection by training a machine learning algorithm, by contrast to the fast-evolving and always-subjective nature of hate speech.[174]

The net impact of these systems on the right to be free from discrimination[175] is indeterminate. As with free expression, AI systems could be trained to avoid the biases exhibited by human reviewers, but on the other hand there is a considerable risk that machine learning techniques will result in the replication and scaling of existing human patterns of bias into new automated content review systems.[176]

One additional right that merits discussion in this context is that of the individuals employed in content review and moderation positions to just and favorable conditions of work.[177] The psychological toll that the frontline work of content review and moderation takes is considerable, as these individuals are exposed to the very worst of humanity day in and day out—from child pornography to gruesome acts of violence. Content reviewers are disproportionately female, but reviewers of all genders suffer from depression, burnout, anxiety, sleep difficulties, and even from post-traumatic stress disorder at extraordinary rates.[178] Using AI to lessen the psychological burden associated with this work could well have positive human rights impacts on a group of individuals who are often forgotten in conversations about how best to respond to problematic content online.

[171] UDHR art. 19.

[172] For example, it is only in April of this year that Facebook announced that it was creating a system by which its users could appeal content removal decisions that they believe to be in error. Monika Bickert, "Publishing Our Internal Enforcement Guidelines and Expanding Our Appeals Process," Facebook (official blog), April 24, 2018, https://newsroom.fb.com/news/2018/04/comprehensive-community-standards/.

[173] UDHR art. 3.

[174] Hanna Kozlowska, "Facebook Is Revealing Data on How Good It Is at Moderating Content, but the Numbers Have Holes," *Quartz*, May 18, 2018, https://qz.com/1277729/facebook-is-revealing-data-on-how-good-it-is-at-moderating-content-but-the-numbers-dont-say-much/.

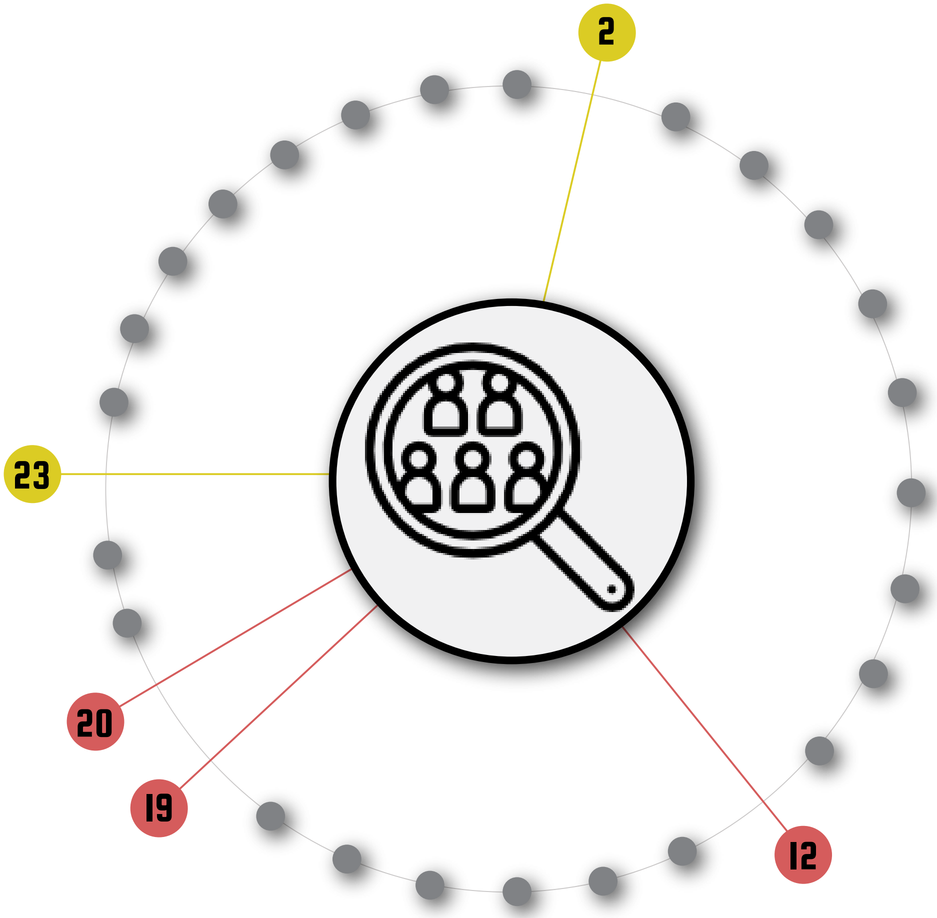
[175] UDHR art. 2.

[176] Reuben Binns et al., "Like Trainer, like Bot? Inheritance of Bias in Algorithmic Content Moderation," *ArXiv:1707.01477 [Cs]* 10540 (2017): 405–15, https://doi.org/10.1007/978-3-319-67256-4.

[177] UDHR art. 23.

[178] Andrew Arsht and Daniel Etcovitch, "The Human Cost of Online Content Moderation," *Harvard Journal of Law & Technology Digest*, March 2, 2018, https://jolt.law.harvard.edu/digest/the-human-cost-of-online-content-moderation.

5.5 Human Resources: Recruitment and Hiring



2. Freedom from Discrimination
AI-based hiring systems may perpetuate patterns of discrimination in employment due to biases in the training data, though they could be programmed to avoid doing so.

12. Right to Privacy
Especially if they share their data with employers, AI-based hiring systems that canvass a wide variety of data sources can impact the privacy of job-seekers and of employees too.

19. Freedom of Opinion, Expression, and Information
20. Right of Peaceful Assembly and Association
Since "all data" could be "hiring data" for AI, people may be chilled from saying certain things or from associating with certain others for fear of the impact on their employability.

23. Right to Desirable Work
AI-based hiring systems could help improve pay equity if designed and implemented correctly, otherwise it might exacerbate current pay inequalities.

● *Positive human rights impact* ● *Human rights impact indeterminate* ● *Negative human rights impact*

Organizations cannot operate without employees, but the process of identifying, recruiting, and hiring new employees is hard work. Increasingly, public- and private-sector employers are turning to AI to help with the hiring process for at least two reasons.[179] The first is capacity: the number of applicants per position has multiplied in the last several years, while staffing levels at human resources ("HR") departments remain flat. The second is fairness: there is a growing awareness that hiring processes are rife with implicit bias and discrimination, and that hiring decisions often boil down to "is this person like me?" Many organizations believe that AI may offer at least a partial solution to this challenge.

The responsibility of business to respect human rights applies not just to the services they provide and the products they sell, but also to their internal operations. Flawed hiring processes may have significant implications for the right to freedom from

[179] "A New Age of Opportunities: What Does Artificial Intelligence Mean for HR Professionals?" (Ontario: Human Resources Professionals Association, 2017).

discrimination,[180] the right to equal pay for equal work,[181] and the rights to freedom of expression and association.[182] Governments have recognized the need for mechanisms to provide remedy for individuals subjected to discriminatory hiring practices and have created institutions such as the U.S. Equal Employment Opportunity Commission ("EEOC") and the Canadian Human Rights Commission. As AI-based hiring systems become commonplace, it will be important to evaluate whether these existing mechanisms are up to the task of ensuring that these new technologies are free from bias.

Traditional Approach to Recruitment and Hiring

Recruiters have long relied on technology to streamline the hiring process. Today, HR departments commonly use applicant tracking systems ("ATS") to aggregate applicant information and filter them based on certain criteria, such as years of experience, education, or other keywords. Shortlisted candidates are then interviewed and a decision is made on who to hire. Few quantitative data points are used in this process; instead, it relies on an often-flawed combination of pedigree and gut instinct.

Problems abound in this human-based system. There has been much debate about why we continue to see men from the majority group hired and promoted over women and minorities. Some claim that it may reflect systemic inequities in society leading to men being more highly educated and better prepared for a particular job.[183] But research shows that much of the problem is discrimination resulting from implicit biases that manifest themselves in individual decision-making. In fact, our very definitions of success can exhibit bias. One experimental study shows that individuals are prone to shifting their definition of merit when evaluating applicants to advantage certain groups, which plays a role in gender and racial discrimination.[184] This distinction is important because, while systemic inequity is a serious problem, decisions marred by individual bias more directly implicate the right to be free from discrimination.[185]

Research in the United States has shown, repeatedly, that "white-sounding" names receive 50% more callbacks than "African-American-sounding" names—despite otherwise identical resumes.[186] Moreover, males often receive more callbacks for traditionally "male" jobs. In one experiment designed to explore the dearth of women in STEM professions, researchers found that women were half as likely as men to be hired for a job, based on a flawed assumption that the female candidates performed worse on an arithmetic task. This was the case even when the women actually performed better than their male counterparts. The reason

[180] UDHR art. 7.

[181] UDHR art. 23(2).

[182] UDHR art. 20.

[183] Matthew Scherer, "AI in HR: Civil Rights Implications of Employers' Use of Artificial Intelligence and Big Data," *SciTech Lawyer* 13 (2017 2016): 12-16.

[184] Eric Luis Uhlmann and Geoffrey L. Cohen, "Constructed Criteria: Redefining Merit to Justify Discrimination," *Psychological Science* 16, no. 6 (June 1, 2005): 474–80, https://www.ncbi.nlm.nih.gov/pubmed/15943674.

[185] UDHR art. 2.

[186] Marianne Bertrand and Sendhil Mullainathan, "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination" (Cambridge, MA: National Bureau of Economic Research, July 2003).

for this was that men were more likely to inflate their performance on the task in an interview and women were more likely to underestimate their performance.[187] Furthermore, the right to equal pay for equal work[188] is implicated by the fact that men receive higher starting salary offers than women for the same job.[189]

Finally, the right to freedom of association[190] is implicated by human bias in the hiring process. Involvement in certain organizations, such as ethnic affinity groups or LGBTQ networks, can negatively impact job prospects. A résumé audit study found that women with a leadership role in a student LGBTQ organization received 30% fewer callbacks for a job posting than applicants with an identical resume but without the LGBTQ association.[191] Individuals may not know that this information is limiting their job prospects, but if they did, they might feel pressure to disassociate themselves from controversial groups.

Ultimately, the impact of AI hiring systems on human rights will depend, in part, on whether the controls meant to mitigate or to remedy rights-related harms in the existing human-based system can be applied to the new technology. It is to the technology now that we must turn in order to evaluate this question.

AI Assisted Hiring and Recruitment

Artificial intelligence is now being used to augment much of the hiring process. Job descriptions can be run through text analysis software to flag gendered language that might discourage highly qualified women from applying.[192] Companies can enlist algorithms to advertise openings to eligible candidates through LinkedIn or Google's ad network.[193] AI is also being used to screen applicants using natural language processing to parse resumes.[194] Some technologies even draw on social media and other public data to supplement their analyses. After AI is used to narrow down the applicant pool, companies might invite candidates to conduct recorded interviews, where algorithms evaluate word choice, vocal inflection, and even emotions (using facial recognition).[195] These technologies purport to identify candidate "personalities" and help establish "fit" within the company. AI is clearly streamlining the hiring process, but the verdict on AI's ability to mitigate negative human rights impacts is unclear.

[187] Ernesto Reuben, Paola Sapienza, and Luigi Zingales, "How Stereotypes Impair Women's Careers in Science," *Proceedings of the National Academy of Sciences*, March 5, 2014, http://www.pnas.org/content/111/12/4403.

[188] UDHR art. 23(2)

[189] "2018 State of Wage Inequality in the Workplace Report," Hired, accessed June 21, 2018, https://hired.com/wage-inequality-report.

[190] UDHR art. 20.

[191] Emma Mishel, "Discrimination against Queer Women in the U.S. Workforce: A Résumé Audit Study," *Socius* 2 (January 1, 2016). https://doi.org/10.1177/2378023115621316.

[192] Software employed by Textio and Gender Decoder use NLP paired with research on language pattern and implicit word associations to make job descriptions more gender neutral. See Textio, https://textio.com/, and http://gender-decoder.katmatfield.com/.

[193] John Jersin, "How LinkedIn Uses Automation and AI to Power Recruiting Tools," *LinkedIn Talent Blog* (blog), October 10, 2017, https://business.linkedin.com/talent-solutions/blog/product-updates/2017/how-linkedin-uses-automation-and-ai-to-power-recruiting-tools.

[194] See examples "Sovren," accessed June 20, 2018, https://www.sovren.com/, and "Textkernel Launches the First Fully Deep Learning Powered CV Parser," Textkernel (blog), February 8, 2018, https://www.textkernel.com/extract-4-0-textkernel-launches-the-first-fully-deep-learning-powered-cv-parsing-solution/.

[195] HireVue.com, "HireVue: Video Interview Software for Recruiting & Hiring," accessed June 20, 2018, https://www.hirevue.com.

Previous sections have noted how the veneer of objectivity that technology provides can be dangerous, because it obscures how AI often replicates human biases at scale. This is particularly worrying when AI is used to devise predictors of success that will determine hiring and advancement opportunities for future applicants and employees. There is already evidence that gender stereotypes have seeped into the "word embedding frameworks"[196] used in many machine learning and natural language processing technologies. One of the more egregious cases revealed in a 2016 study found that an algorithm trained on Google News articles to understand word meanings would respond to the query "man is to computer programmer as woman is to x" with x = homemaker.[197] In view of this example, there is a very real danger that an AI-based hiring algorithm trained on performance reviews,[198] employee surveys, and other data points meant to uncover the attributes of successful employees will reproduce existing patterns of bias in future hiring decisions. The system may produce more consistent results across candidates than human hiring managers, but the outputs of such a system can hardly be described as fair.[199]

Summary of Human Rights Impacts

With the foregoing in mind, the question of how machine learning technologies used in hiring will impact the right to be free from discrimination becomes more complicated.[200] Without intentional intervention in the programming, it seems likely that AI will reproduce the existing systemic patterns of bias and prejudice exhibited in the training data. This may lead AI-based hiring systems to identify metrics for assessing candidates that reflect structural biases rather than the objective determinants of real-world employment performance.

Some technologists and researchers have identified this as a concern and are devising technical solutions. One solution proposes a decoupling technique[201] that, in the resume screening context, would allow an algorithm to identify top candidates using variables optimized based on other applicants of a certain category (e.g. race or gender) rather than against the entire applicant pool. In practice, this means the traits to select for a female or minority applicant would be identified based on the trends of other female or minority applicants, and these could differ from the identified successful traits of

[196] Tolga Bolukbasi et al., "Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings," ArXiv:1607.06520 [Cs, Stat], July 21, 2016, http://arxiv.org/abs/1607.06520.

[197] Word embedding is a set of language and feature modeling techniques used in NLP to map words or phrases onto vectors of real numbers. This allows computer programs to understand and use word meaning, see Tolga Bolukbasi et al., "Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings," ArXiv:1607.06520 [Cs, Stat], July 21, 2016, http://arxiv.org/abs/1607.06520.

[198] Performance reviews and promotion structures often exhibit gender and racial bias, which contributes to the paucity of women and minorities as you move up the corporate ladder. See Kieran Snyder, "The Abrasiveness Trap: High-Achieving Men and Women Are Described Differently in Reviews," *Fortune*, August 26, 2014, http://fortune.com/2014/08/26/performance-review-gender-bias/, and Buck Gee and Janet Wong, "Lost in Aggregation: The Asian Reflection in the Glass Ceiling" (The Ascend Foundation, September 2016), and Hannah Riley Bowles, Linda Babcock, and Lei Lai, "Social Incentives for Gender Differences in the Propensity to Initiate Negotiations: Sometimes It Does Hurt to Ask," *Organizational Behavior and Human Decision Processes* 103, no. 1 (May 2007): 84–103, https://doi.org/10.1016/j.obhdp.2006.09.001.

[199] Companies like Koru are using this method to establish their client organizations' "predictive hiring fingerprint" and are claiming that it reduces bias and the proclivity to hire based on pedigree. https://www.joinkoru.com/.

[200] UDHR art. 2.

[201] Cynthia Dwork et al., "Decoupled Classifiers for Fair and Efficient Machine Learning," *ArXiv:1707.06613 [Cs]*, July 20, 2017, http://arxiv.org/abs/1707.06613.

a male or majority applicant. The feasibility—and legality—of implementing a technical solution that optimizes fairness by distinguishing between individuals on sensitive personal characteristics attributes is highly dependent on jurisdiction.

More hope lies in companies intentionally designing algorithms to control for human biases and implementing auditing systems that can regularly test for bias and errors. Applied and Pymetrics, for example, have worked with academics to devise AI-based hiring systems that use anonymization, skills testing, work product analysis, neurological brain games, and other methods to remove bias based on gender, race, and pedigree.[202] While these efforts are promising, their outcomes still depend on the reliability of their algorithms and the level of bias in their data. Where AI is being used in the hiring process, it will be important to implement robust auditing structures to regularly examine biases in the data and how these are influencing the outputs.

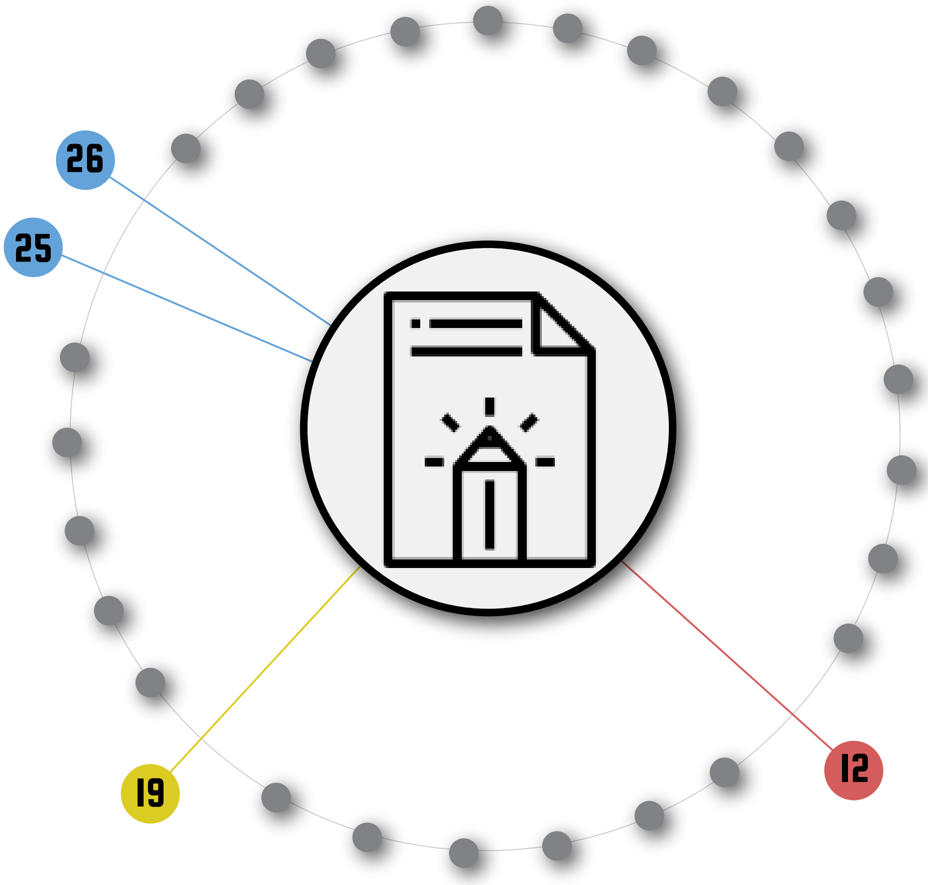
AI-based hiring systems may have a greater negative impact on the freedoms of association[203] and expression[204] than the current human-based system. Similar to the concern that people will carefully curate their associations to maximize their credit scores, there is a risk that applicants will feel compelled to disassociate themselves from organizations that might hurt their chances of securing employment, because AI-based systems are more likely to detect such associations than human recruiters. Likewise, AI-based hiring systems may chill individuals from engaging in certain forms of expressive activity out of fear that their words will be used against them in the employment context.

[202] Applied, "Can We Predict Applicant Performance without Requiring CVs? Putting Applied to the Test — Part 1," *Medium* (blog), September 21, 2016, https://medium.com/finding-needles-in-haystacks/putting-applied-to-the-test-part-1-9f1ad6379e9e; Josh Constine, "Pymetrics Attacks Discrimination in Hiring with AI and Recruiting Games," *TechCrunch* (blog), accessed February 22, 2018, http://social.techcrunch.com/2017/09/20/unbiased-hiring/.

[203] UDHR art. 20.

[204] UDHR art. 19.

5.6 Education: Essay Scoring



12. Right to Privacy
Like nearly all other AI, automated grading systems require the store and analysis of large volumes of writing, which may impact privacy.

19, Freedom of Opinion, Expression, and Information
The improvements in writing ability that automated grading systems promise may enhance enjoyment of free expression rights, but the capture of such writing in databases may chill it.

25. Right to Adequate Standard of Living
Automated grading systems may positively impact the enjoyment of economic rights, as writing is a key skill for success in the labor market.

26. Right to Education
Automated grading systems may improve educational outcomes by increasing the availability of feedback on students' writing abilities, thereby spurring performance improvements.

🔵 *Positive human rights impact* 🟡 *Human rights impact indeterminate* 🔴 *Negative human rights impact*

Access to education is a right in and of itself and a key enabler of a panoply of other human rights. Educational attainment is the primary engine of social mobility;[205] those who are more educated are better able to participate in the economy, engage in civic and public life, and improve their personal and local circumstances.

One of the key skills education systems around the world seek to develop in their students is the ability to write. Not only is the quality of one's writing an important factor in the job market, but it is also an important ability for achieving the "full development of the human personality" in the words of the Universal Declaration of Human Rights,[206] as well as one of the primary means through which the right to free expression is exercised.

The importance of writing is perhaps why this skill is so frequently tested at every level of the education system. Most countries periodically administer standardized tests of their students' written abilities

[205] Michael Greenstone et al., "Thirteen Economic Facts about Social Mobility and the Role of Education" (The Hamilton Project, June 26, 2013).

[206] UDHR art. 26(2).

in the local *lingua franca* as they progress through the educational system. In many countries, students must pass writing tests to graduate from a particular level of schooling, or for admission to the next level. In North America, for example, both the Graduate Management Administrative Test ("GMAT") and the Graduate Record Examinations ("GRE") evaluate the ability of prospective management and law students to produce an analytical writing sample under time constraints.

How writing is marked matters. In the day-to-day school context, the quality of the feedback students receive on their writing will impact the prospects of student improvement over time. And in the context of gatekeeping exams such as the GMAT and the GRE, how test-takers' writing is evaluated may have life-long impacts on their future opportunities. Consequently, as automated techniques are increasingly being used to evaluate student writing, the question arises of what their impacts will be.

Traditional Approach to Essay Grading

The traditional approach to grading writing, whether in the ordinary schooling or the standardized testing context, is for trained individuals to perform this task. When a large volume of writing needs to be evaluated against an established performance standard, a common approach is for the individuals responsible for the grading to each evaluate a representative sample of what has been submitted, and to compare results in order to calibrate their approach for the rest of the grading.

In most contexts, evaluating the quality of writing requires not just considering the mechanics of grammar and syntax, but also evaluating the writing for the accuracy of the substance and on considerations such as style and rhetorical impact. Even so, in the context of one common standardized test, a study found that the score assigned to the writing component could be predicted accurately 90% of the time by considering just one variable: length.[207]

Although educators may have more time to engage with the style of the substance of a student's writing than the evaluator of a standardized test, the resource constraints under which most school systems operate may lead teachers to turn, intentionally or not, to more mechanical assessments of their students' writing. Such assessments may fail to provide the feedback students need for growth and improvement.[208] Indeed, studies have found that educators often emphasize form over substance in teaching writing, which may well result in students prioritizing form over substance in their own writing.[209]

[207] Arthur C. Graesser and Danielle S. McNamara, "Automated Analysis of Essays and Open-Ended Verbal Responses.," in *APA Handbook of Research Methods in Psychology, Vol 1: Foundations, Planning, Measures, and Psychometrics.*, ed. Harris Cooper et al. (Washington: American Psychological Association, 2012), 307–25, https://doi.org/10.1037/13619-017; Michael Winerip, "SAT Essay Test Rewards Length and Ignores Errors," *The New York Times,* May 4, 2005, sec. Education, https://www.nytimes.com/2005/05/04/education/sat-essay-test-rewards-length-and-ignores-errors.html.

[208] Deborah Reck and Deb Sabin, "A High-Tech Solution to the Writing Crisis," *The Atlantic*, October 16, 2012, https://www.theatlantic.com/national/archive/2012/10/a-high-tech-solution-to-the-writing-crisis/263675/.

[209] Gary A. Troia and Steve Graham, "Effective Writing Instruction Across the Grades: What Every Educational Consultant Should Know," *Journal of Education and Psychological Consultation* 14, no. 1 (2003): 75–89.

AI-Graded Essays

Following decades of research and many fruitless attempts, automated essay scoring has recently become a reality. Indeed, these technologies are among the more mature current applications of artificial intelligence. Using machine learning, these systems are trained to grade written materials by ingesting a set of exemplars that have been graded by human experts. These systems identify features in the set of training materials that correlate with success, and they then assess any written materials that are fed into the system according to what they have learned.[210]

Such automated systems are already in use in high-stakes standardized testing environments. For example, the written component of the GMAT exam is currently scored by an automated system and an expert human greater working separately. Should the AI and the human differ in the grade they assign by more than one point, a second human grader acts as a tiebreaker.[211]

The use of automated grading systems has the potential to positively impact the right to education in a number of different ways. Automated systems permit students to engage in more deliberate practice of their writing, receiving more feedback on at least some elements of their writing than they would otherwise. In the context of communities without reliable access to quality writing instruction, whether due to poverty or other forms of marginalization, automated systems may be one of the only reasonably available means to obtain the feedback required to develop one’s writing skill. Some studies have demonstrated that the instantaneous feedback of rudimentary grammar checkers have powerful effects on the quality of written expression, so it is reasonable to assume that the same is true of more nuanced AI-based approaches to the same basic endeavor.[212]

Relatedly, automating certain aspects of the grading of writing might free educators to spend more time focusing on higher-order teaching tasks, such as engaging with students’ ideas and arguments. To many, writing is more than sentence structure and word-choice: it is the expression of ideas and emotions, drawing on cognitive skills distinct from those underpinning grammar and syntax.[213] Moreover, automated systems have the potential to eliminate bias in grading by removing the opinions of the author and the grader and any relationship between them from the equation.[214]

On the flipside, there are serious concerns relating to the fact that these systems cannot understand

[210] Corey Palmero, “A Gentle Introduction to Automated Scoring.Pdf” (Measurement Incorporated, October 2017), http://www.measurementinc.com/sites/default/files/2017-10/A%20Gentle%20Introduction%20to%20Automated%20Scoring.pdf.

[211] “How the Analytical Writing Assessment Is Scored,” Economist GMAT Tutor, April 28, 2017, https://gmat.economist.com/gmat-advice/gmat-overview/gmat-scoring/how-analytical-writing-assessment-scored/.

[212] Paul Morphy and Steve Graham, “Word Processing Programs and Weaker Writers/Readers: A Meta-Analysis of Research Findings,” Reading and Writing 25, no. 3 (March 2012): 641–78, https://doi.org/10.1007/s11145-010-9292-5.

[213] Troia and Graham, “Effective Writing Instruction Across the Grades: What Every Educational Consultant Should Know.”

[214] For examples of bias in writing grading, see John Malouff, “Bias in Grading,” *College Teaching* 56, no. 3 (July 2008): 191–92, https://doi.org/10.3200/CTCH.56.3.191-192. Bias may also come in the form of pre-existing stereotypes influencing how critically one reviews the essay. For an example in the legal profession, see Arin N. Reeves, “Written in Black & White: Exploring Confirmation Bias in Racialized Perceptions of Writing Skills,” Yellow Paper (Nextions, 2014).

what is written in the same way as human readers.[215] Some systems might well be able to detect offensive content, but at least for the foreseeable future, artificial intelligence systems will not realistically possess the general intelligence that humans do which enables them to evaluate the validity of written material.[216] Especially in our era where the truth and accuracy of written materials has become an issue of the utmost public importance, the likely inability of automated grading systems to assess factual validity is of concern.

Furthermore, there are also significant concerns about what incentives these systems create for those who are subject to being evaluated by them.[217] Consider, for example, that a famous essay by the renowned MIT linguist Noam Chomsky received a grade of "fair" when it was fed into an automated grading system.[218] If students respond to the growing prevalence of automated grading systems by focusing on form and length to the detriment of style and substance, these technologies may be doing them a disservice.

Finally, automated grading systems depend on the collection, storage, and analysis of vast quantities of written material. This raises not only the standard privacy-related concerns that accompany most AI systems,[219] but there is an additional risk that these systems might chill the full enjoyment of the right to free expression. Even in the educational context, a student might be more willing to share writing about something that is deeply personal or controversial with a trusted teacher as opposed to an AI system, if it is going to end up being catalogued in a vast database and subject to being used as training data for some other purpose.

Summary of Human Rights Impacts

On balance, the rise of automated grading systems is likely to have a positive impact on the right to education, as these systems can potentially increase global access to at least some feedback on people's writing. In much of the world today, educational systems are simply too overburdened to provide the kind of individualized evaluation and feedback on writing that is desirable, so the potential of these technologies to improve the situation over the sta-

[215] Stephen P. Balfour, "Assessing Writing in MOOCs: Automated Essay Scoring and Calibrated Peer Review," *Research & Practice in Assessment* 8, no. 1 (2013): 40–48. For an example of the concerns, see Kai Riemer, "On Rewarding 'Bullshit': Algorithms Should Not Be Grading Essays," Undark, accessed April 21, 2018, https://undark.org/article/rewarding-bullshit-algorithms-classroom/.

[216] For example, Dr. Les Perelman of MIT and his team developed a piece of software they call the "Basic Automatic B.S. Essay Language Generator" (BABEL) which, upon the input by the user of three words, auto-generates essays that routinely receive near-perfect scores from various automated grading systems that are currently in use, although their content is nothing but machine-generated nonsense. See Steve Kolowich, "Writing Instructor, Skeptical of Automated Grading, Pits Machine vs. Machine," Chronicle of Higher Education, April 28. 2014, https://www.chronicle.com/article/Writing-Instructor-Skeptical/146211. Dr. Perelman's BABEL system is available to try out at http://lesperelman.com/. To be sure, humans are not immune to being fooled by machine-generated gibberish. See Danielle Wiener-Bronner, "More Computer-Generated Nonsense Papers Pulled From Science Journals," *The Atlantic,* Mar. 3, 2014, https://www.theatlantic.com/technology/archive/2014/03/more-computer-generated-nonsense-papers-pulled-science-journals/358735/.

[217] Jinhao Wang and Michelle Stallone Brown, "Automated Essay Scoring Versus Human Scoring: A Correlational Study," *Contemporary Issues in Technology and Teacher Education* 8, no. 4 (2008): 16.

[218] "Parts of Noam Chomsky's Essay 'The Responsibility of Intellectuals' Grammar Checked by ETS's Criterion and WhiteSmoke | Les Perelman, Ph.D.," accessed June 21, 2018, http://lesperelman.com/writing-assessment-robo-grading/parts-noam-chomskys-essay-grammar-checked/.

[219] Related privacy concerns have been raised about automated plagiarism-detection software, which also require the storage and retention of vast quantities of student-written material to be effective. See Bo Brinkman, "An Analysis of Student Privacy Rights in the Use of Plagiarism Detection Systems," *Science and Engineering Ethics* 19, no. 3 (2013): 1255–1266.

tus quo is considerable. The feedback these systems provide might fall short of the Platonic ideal of individualized attention from expert human instructors, but they are a step in the right direction for the vast majority of students worldwide who lack affordable access to high-quality instruction.

Considering that the ability to write is a key enabler of a panoply of civil and political rights, from that of free expression to the right to take part in cultural and scientific life, the improvements that automated grading systems will make in individuals' ability to write when judged on a global basis is likely to improve their ability to enjoy these rights. This is especially true of the crucial right that everyone possesses to an adequate standard of living, as the ability to write is so central to one's employment prospects and ability to participate effectively in various aspects of society.

The impact of these automated systems on the right to free expression, however, is more complex. On one hand, anything that improves the ability of people to express themselves in an effective manner would seem to positively impact the enjoyment of this right. On the other hand, as noted above, large-scale collection of written materials that automated systems necessarily entail may chill some individuals from setting down in writing things that they might have said otherwise.

6. Addressing the Human Rights Impacts of AI: The Strengths and Limits of a Due Diligence-Based Approach

The six use cases explored in the previous section demonstrate how artificial intelligence has the potential to positively impact the full range of human rights. Automating complex tasks that currently require the labor of highly trained professionals could well usher in greater access to specialized healthcare, education, and financial services. Such technologies also have the considerable potential to reduce and correct for various biases that plague human decision-making, from outright discrimination to our reliance on heuristics that sometimes lead us astray. Yet this promise comes at an almost inevitable cost to our privacy due to the data-intensive nature of these technologies which, in turn, may chill the exercise by many of their civil and political rights. Likewise, the possibility that these technologies will reproduce and ossify existing patterns of discrimination and bias, while also producing troubling distributive consequences, must be contended with.

This conundrum gives rise to the question of how we can enjoy the benefits of artificial intelligence—especially the vast potential for positive impacts on human rights— while minimizing its real negative risks.

This is not a new question, but rather one that has arisen with every major technological innovation throughout history. From the development of industrial machinery to the invention of the automobile, transformative technological changes have posed a profound challenge to the existing social order. These technologies utterly transformed society in their time—oftentimes for the better, yet they were frequently accompanied by bad consequences, too. Industrialization, for example, democratized the availability of goods that were once luxuries, though at the cost of widespread economic displacement, Dickensian working conditions, suffocating air pollution, and colonial patterns of natural resource exploitation. Likewise, the automobile revolutionized human mobility and fundamentally transformed the economy, though with the negative consequences of air pollution, urban sprawl, and millions of traffic casualties every year.

The negative impacts of these and other transformative technologies were felt most acutely as they were first coming into widespread use. Over time, however, society responded by developing control mechanisms to attempt to enjoy the good while minimizing the bad. Some controls are regulatory in nature, such as laws and norms that specify how and when a technology might be used, while others are technological, such as design features that channel a technology towards certain uses and away from others.[220] Oftentimes, controls are a mix of the two, such as regulatory standards that mandate specific design characteristics.

History shows that it can take quite some time to develop effective mechanisms to control new technologies. The Industrial Revolution began to transform Britain in the 18th century, but it was only in the mid-19th century that Parliament started enact-

[220] In the U.S. context, Lawrence Lessig famously noted that the "East Coast Code" of laws and regulations promulgated in Washington D.C., and the "West Coast Code" produced by software engineers in Silicon Valley, are both fundamentally forms of regulation. Lawrence Lessig, *Code*, Version 2.0 (New York: Basic Books, 2006), http://codev2.cc/download+remix/Lessig-Codev2.pdf.

ing legislation to address its consequences.[221] Likewise, although automobiles became commonplace in the first decades of the 20th century, the first systematic studies of vehicular safety took place in the 1940s,[222] and the first comprehensive automobile safety laws weren't enacted until the 1960s.[223]

Of course, not all of these controls are effective or ideal. Most, if not all of them, are at least partially flawed. Some are too permissive to adequately address the negative consequences of the regulated technology, while others are too restrictive to permit the realization of its benefits. All the same, we must think about which controls are necessary, sufficient, and appropriate to reduce and redress the human rights impacts of artificial intelligence.

We are fortunate to be in a position to design the regulatory and technological controls required to maximize the human rights and other benefits of AI concurrently with the technology itself. AI is the first truly transformative technology to come of age following the articulation of the United Nations Guiding Principles on Business and Human Rights. It is emerging at a time that it is widely understood that businesses have a responsibility to respect human rights, and that due diligence is the key to doing so. Whereas in the past private sector innovators could be ignorant or willfully blind to the human rights consequences of the technologies they are developing, that is no longer the case.

Due diligence, as the term is used in the Guiding Principles, is the essential first step toward identifying, mitigating, and redressing the adverse human rights impacts of AI. Therefore, as a minimum, public policy efforts should be directed toward ensuring that all who are involved in building these systems engage in the kinds of due diligence that will ensure that they respect human rights by design. Such efforts may be enhanced by mandating or incentivizing the developers and operators of AI systems to make available the training data and the outputs of their systems to external reviewers.[224]

It is heartening that many of the biggest players in developing AI have risk management systems in place that trigger human rights due diligence processes at all appropriate stages in the lifecycle of a technology.[225] That being said, there are at least three challenges endemic to the AI space that may prevent human rights due diligence from being as effective as it might otherwise be.

[221] B.L. Hutchins and A. Harrison, *A History of Factory Legislation*, 2nd ed. (London: P. S. King & Son, 1911), https://archive.org/details/historyoffactory014402mbp.

[222] For a popular and accessible history of automobile safety, listen to 99% Invisible, *The Nut Behind the Wheel*, accessed June 21, 2018, https://99percentinvisible.org/episode/nut-behind-wheel/.

[223] "National Traffic and Motor Vehicle Safety Act of 1966," Pub. L. No. 89–563 (1966).

[224] In this vein, New York University's AI Now Institute has developed a framework for public-sector entities in the United States to use in carrying out "algorithmic impact assessments" prior to purchasing or deploying automated decision systems. Dillon Reisman et al., "Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability" (New York University AI Now Institute, April 2018), https://ainowinstitute.org/aiareport2018.pdf.

[225] For example, the eleven member-companies of the Global Network Initiative ("GNI"), which include some of the biggest players in the AI space, commit to "carry out human rights due diligence to identify, prevent, evaluate, mitigate and account for risks to the freedom of expression and privacy rights that are implicated by the company's products, services, activities and operations." GNI member-companies are independently assessed every two years to evaluate their compliance with this and other commitments. "Implementation Guidelines" (Global Network Initiative), accessed June 21, 2018, https://globalnetworkinitiative.org/implementation-guidelines/.

The first arises from the relatively low awareness among small, early-stage companies of the corporate responsibility to carry out human rights due diligence. This is by no means a challenge that is unique to AI, but given the potential of these technologies to scale up rapidly, it might be more problematic in this space than in other industry verticals.[226] Moreover, given that certain AI systems are often empty vessels into which the end-user can feed whatever training data it wants to automate a formerly manual process, technology developers can be too remote from on-the-ground realities to assess the human rights impacts of the uses of their products.[227] Correspondingly, there is an opportunity to significantly advance human rights by adopting measures to incentivize much wider due diligence efforts throughout the entire AI ecosystem.

The second arises from the difficulty in ascertaining the real-world impacts of any given AI application prospectively. That difficulty arises from the inscrutability of so many AI systems and from the complex interactions that these systems have once they begin to operate in the real world. It is hard enough to predict what human rights impacts a relatively anodyne product will have when it is released into the marketplace, hence the challenge of assessing the human rights impacts of AI systems before they are deployed is all the more considerable. The problem is particularly acute in AI systems which utilize machine or deep learning, such that the AI developer herself may not be able to predict or understand the system's output.[228]

To be sure, the Guiding Principles make clear that human rights due diligence is an ongoing responsibility for precisely this reason: not all impacts can be predicted, even with reasonable diligence. Correspondingly, all entities that are involved in the development or use of these technologies must have measures in place to ensure that human rights due diligence is not a matter of "once and done." Especially given the complexity of AI systems and the fact that the results are often not explainable by conventional means, new analytical techniques and performance metrics may need to be developed to determine whether AI systems are helping or harming human rights. Developing these techniques and metrics is a challenge that the computer science community is working to tackle with alacrity. In Europe, this challenge has been framed in part by the provisions of the General Data Protection Regulation, which requires some human involvement in automated decision-making[229] and encourages the development of "a right to an explanation."[230]

[226] Dalia Ritvo, Vivek Krishnamurthy, and Sarah Altschuller, "Managing Users' Rights Responsibly—A Guide for Early-Stage Companies," 2016, http://www.csrandthelaw.com/wp-content/uploads/sites/2/2016/03/Managing-Users-Rights-Responsibly_A-Guide-For-Early-Stage-Companies-no-logos.pdf.

[227] Consider, for example, the controversy that emerged around the time that this report was being finalized regarding the U.S. government's use of facial recognition technology supplied by Microsoft in implementing its (now-rescinded) policy of separating the children of unlawful migrants from their parents. One can speculate that Microsoft could not have foreseen how the U.S. government would use this technology at the time it contracted to provide this system, which highlights the need for companies in this space to conduct *ongoing* due diligence. Catherine Shu, "Microsoft Says It Is 'Dismayed' by the Forced Separation of Migrant Families at the Border," *TechCrunch*, June 19, 2018, http://social.techcrunch.com/2018/06/18/microsoft-says-it-is-dismayed-by-the-forced-separation-of-migrant-families-at-the-border/.

[228] To be clear, many AI applications reason in ways beyond human comprehension. This is particularly true for applications based on machine learning and deep learning. Yet, that difference may be insufficient to justify holding AI back. In fact, it may even be a reason to delegate decisions to AI. David Weinberger, "Our Machines Now Have Knowledge We'll Never Understand," *WIRED*, April 18, 2017, https://www.wired.com/story/our-machines-now-have-knowledge-well-never-understand/.

[229] GDPR, art. 22.

[230] Ibid., art 22 and recital 71. For more information, see Bryce Goodman and Seth Flaxman, "European Union Regulations on Algorithmic Decision-Making and a 'Right to Explanation,'" *AI Magazine* 38, no. 3 (October 2, 2017): 50, https://doi.org/10.1609/aimag.v38i3.2741 and Sandra

A third complication arises from the uncertainty as to what constitutes an effective remedy to an adverse human rights impact generated by an AI system. Since a right without a remedy is no right at all, the right to remedy is the "third pillar" of the "protect, respect, remedy" framework on which the Guiding Principles rest. Specifically, Guiding Principle 25 recognizes the duty of the state to provide access to effective judicial and other remedies to all those who have been affected by business-related human rights abuses. Judicial remedies in particular may be better suited to addressing the adverse consequences of AI on some human rights over others. For example, judicial remedies may well be more effective in detecting and redressing adverse human rights impacts caused by the use of AI in the criminal justice system, as compared to some other fields of endeavor. Especially since criminal procedural rights are articulated in much more detail than most other human rights in domestic and international law, there is simply more material to work with in terms of identifying when an AI system is adversely impacting rights in this category.[231]

Another major challenge facing both judicial and non-judicial remedies is the nature of the harm that AI makes possible. That is, all remedial systems, whether public or private, are much better at remediating substantial harms suffered by the few, as opposed to less significant harms suffered by the many. Consider, for example, some of the most widely-known operational level grievance mechanisms that have been established in the last decade to remedy the adverse human rights impacts of businesses. These include mechanisms established by mining companies to compensate the victims of sexual violence,[232] or by global technology companies to vindicate the right that some courts have recognized of individuals to be "forgotten" online.[233] These systems provide remedies to individuals or to small groups of people that have suffered a particularized human rights harm, but they are simply not designed to cope with much more diffuse and oftentimes covert harms that might be every bit as pernicious.[234]

These difficulties are magnified in the AI space by the challenge of detecting the harm and determining and proving causation. Consider, for example, the difficulties that a loan applicant would face in proving that a lending algorithm has discriminated against them, in a situation where seven prospective lenders turned them down, but three others offered them credit. Assuming that the objective truth of

Wachter, Brent Mittelstadt, and Luciano Floridi, "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation," *International Data Privacy Law* 7, no. 2 (May 2017): 76–99, https://doi.org/10.1093/idpl/ipx005.

[231] The fact that the judiciary might be better able to grapple with the adverse impact of AI in its own backyard does not mean that it will reach the right answer in every case, or that the remedies it provides when a violation is found will be sufficient. For example, the U.S. Supreme Court has been roundly criticized for denying review of *Wisconsin v. Loomis,* 881 N.W.2d 749 (2016). See Ellora Israni, "Algorithmic Due Process: Mistaken Accountability and Attribution in State v. Loomis," *Harvard Journal of Law & Technology Digest*, 2017, https://jolt.law.harvard.edu/digest/algorithmic-due-process-mistaken-accountability-and-attribution-in-state-v-loomis-1 and Taylor R Moore, "Trade Secrets and Algorithms as Barriers to Social Justice" (Center for Democracy & Technology, August 2017). The point is simply that some alleged rights violations attributed to AI systems are easier to redress judicially in some fields as opposed to others.

[232] Yousuf Aftab, "Pillar III on the Ground: An Independent Assessment of the Porgera Remedy Framework" (Enodo Rights, January 2016), http://www.enodorights.com/assets/pdf/pillar-III-on-the-ground-assessment.pdf.

[233] Jacques de Werra, "ADR in Cyberspace: The Need to Adopt Global Alternative Dispute Resolution Mechanisms for Addressing the Challenges of Massive Online Micro-Justice," *Swiss Review of International & European Law* 26, no. 2 (2016): 289, https://doi.org/10.2139/ssrn.2783213.

[234] To be sure, the same criticism can be leveled against traditional courts which, despite such innovations as class-action lawsuits and contingency fee arrangements, remain an unaffordable and inaccessible option for many victims of rights violations.

the matter is that some of the seven decliners engaged in discrimination, would this person even suspect that they have been the victim of discrimination? What might be the required elements of proof to establish a discrimination claim? How costly would it be to bring such a claim, as against the anticipated value of the remedies available? What expert evidence and analysis would be required to open the "black box" of the algorithm, especially when it is protected by trade secrets and intellectual property law? Now assume that the stakes of what the algorithm is deciding are much lower than a loan, and yet there are adverse consequences distributed over a large population. Who would go through the trouble of seeking a remedy for that harm, and how and where might they do so?

Then there are the additional complications around remedying AI's impacts on economic, social, and cultural rights. International and domestic legal systems alike have much more developed doctrines and procedures with regard to civil and political rights than they do for economic, social, and cultural rights. This is due in part to the fact that the international human rights community has prioritized civil and political rights over economic, social, and cultural rights for the last 70 years.[235] But it is also because the state duty in relation to economic, social, and cultural rights is to progressively realize them over time, in view of the available resources, which makes it much harder to identify when these rights are adversely impacted—especially by businesses.

The recent General Comment on "State obligations under the International Covenant on Economic, Social and Cultural Rights in the context of business activities" makes clear the difficulties.[236] The General Comment notes that the state obligation to protect human rights requires them to "prevent effectively infringements of economic, social and cultural rights in the context of business activities" by adopting "legislative, administrative, educational and other appropriate measures, to ensure effective protection against Covenant rights violations linked to business activities."[237] Yet all of the examples the General Comment provides of "rights violations linked to business activities" are attributable to state failures to regulate the marketplace.[238]

The underdevelopment of the regime of economic, social, and cultural rights makes it difficult for businesses engaged in human rights due diligence to know what they should do when their systems adversely impact one of these rights. Consider, for example, the impact that at least some AI systems are sure to have on employment. While it is the duty of the state to protect the right to work, large-scale workforce displacement caused by the deployment of AI obviously burdens this right. As things stand right now, however, it is difficult for a business to determine what if anything it should do to mitigate this impact on the right to work.

The example of the right to work points to the profound challenges related to addressing the distributive consequences of AI—especially with regard to

[235] Samuel Moyn, *Not Enough: Human Rights in an Unequal World* (Cambridge, Massachusetts: The Belknap Press of Harvard University Press, 2018); Samuel Moyn, "How the Human Rights Movement Failed," *The New York Times*, April 24, 2018, sec. Opinion, https://www.nytimes.com/2018/04/23/opinion/human-rights-movement-failed.html.

[236] United Nations Economic and Social Council, Committee on Economic, Social and Cultural Rights, General Comment 24 on "State obligations under the International Covenant on Economic, Social and Cultural Rights in the context of business activities" (2007), U.N. Doc. E/C.12/GC/24.

[237] Ibid., para. 14.

[238] Ibid., paras. 18-22.

economic rights, in view of the fears that AI might trigger widespread unemployment. As several of the case studies find, a particular AI system being used in a particular field of endeavor can positively impact the enjoyment of a particular human right by some individuals, at the same time as it adversely affects that right for others.

It may well be that over the course of time, the international human rights system can develop guidance that is more responsive to such distributive issues in the context of AI. Until that happens, however, there is a strong case to be made that national governments should weigh in on these questions, whether by soft suasion or hard regulation, in a manner that is rights-respecting, yet also reflects the country's particular values and public policy priorities. In other words, when due diligence reveals human rights impacts that have complex distributive consequences, the need for government public policy leadership is at its greatest.[239]

There is also an important role for governments to play in creating conditions that encourage businesses to take their human rights responsibilities seriously. In his influential study of private initiatives to improve labor rights in global supply chains, Richard Locke found that "the effectiveness of private regulatory programs is very much tied to the strength of public authoritative rule-making institutions."[240] In the AI space, governments could consider creating incentives to ensure that effective due diligence is undertaken, and to build capacity among earlier-stage companies to develop their technologies in a rights-respecting manner.

Finally, there is a crucial role for governments to play in creating accountability and redress systems for their own use of algorithmic tools, and for those adverse impacts that cannot easily be addressed by private grievance mechanisms. The General Data Protection Regulation ("GDPR"), which recently came into force in the European Union, is noteworthy in this regard for its provisions requiring "data subjects" to be provided with "meaningful information about the logic involved, as well as the significance and the envisaged consequences" of the automated processing of their personal data.[241]

It is too soon to tell whether the GDPR's embrace of algorithmic "explainability" will prove to be successful in creating greater accountability for such systems, or whether this approach will instead chill promising developments in AI that produce useful results even if their logic defies human comprehension. What is certain, however, is that collaboration between technology companies, governments, and representatives of the diverse community of stakeholders that AI will impact is required to develop new ways of ensuring that this technology delivers on its promise in a rights-respecting manner.

[239] As Richard Locke notes in a related context, "[t]he inherent problem with private voluntary initiatives... is their inability to reconcile diverse and conflicting interests and thus promote solutions that require collective action among [] myriad actors...." Richard M. Locke, *The Promise and Limits of Private Power: Promoting Labor Standards in a Global Economy*, Cambridge Studies in Comparative Politics (Cambridge; New York: Cambridge University Press, 2013): 178.

[240] Ibid., 68.

[241] GDPR art. 14(2)(g).

7. Conclusion

As should now be clear, the relationship between artificial intelligence and human rights is complex. A single AI application can impact a panoply of civil, political, economic, social, and cultural rights, with simultaneous positive and negative impacts on the same right for different people. Multiply these impacts across the full range of cases where AI is already in use or will soon become commonplace, and the magnitude of this technology's impact on society begin to become clear.

Society has dealt with revolutionary technological change in the past, and we have always arrived at a new equilibrium. But we today are better placed than our forebears for the change that is upon us because of the adoption, 70 years ago this December, of the Universal Declaration of Human Rights.

The UDHR gives us a powerful and universally-accepted framework not just for identifying and overcoming past and present wrongs, but also for building a future that respects and honors the rights of every person. This depends, however, on our remaining vigilant to the impacts of our actions on the rights of others. Hence the importance the Guiding Principles place on due diligence both before we deploy these powerful new technologies, and throughout their lifecycle, too.

We are heartened by the growing attention that human rights-based approaches to assessing and addressing the social impacts of AI have begun to receive. We view it as a promising sign that so many of the private enterprises at the forefront of the AI revolution are recognizing their responsibility to act in a rights-respecting manner. But the private sector cannot do it alone, nor should it: governments have a crucial role to play, both in their capacities as developers and deployers of this technology, but also as the guarantors of human rights under international law.

The fundamental role of government in defining and making available remedies for human rights violations cannot be overstated. Of equal or greater importance, however, is the governmental responsibility to evaluate and address the distributive consequences of AI. The institutions and processes of democratic government are the only ones with the legitimacy to determine what distribution of benefits and burdens across society is fair, so now is the time for them to embrace their role in shepherding society through the changes that lie ahead.

8. Further Reading

Understanding AI

David Weinberger, “Our Machines Now Have Knowledge We’ll Never Understand,” *Wired*, April 18, 2017, https://www.wired.com/story/our-machines-now-have-knowledge-well-never-understand/.

> This article explores how artificial intelligence is changing the way we think about knowledge by delving into the “explainability” debate. It explains machine learning, artificial neural networks and their implications in an accessible manner. Many machine learning models, like Google’s AlphaGo algorithm, are “ineffably complex and conditional”: they make decisions based on opaque and unexplainable patterns, not transparent principles. These systems are becoming more accurate as data becomes more abundant, yet there is a tradeoff between being able to understand why an algorithm makes a decision (a function of its complexity) and its accuracy. Because reality is complex, artificial intelligence may be able to account for an abundance of factors that are beyond human comprehension. Yet the lack of explainability can make it difficult to identify bias in algorithms. *(Category: Concise overview)*

Jenna Burrell, “How the Machine ‘Thinks’: Understanding Opacity in Machine Learning Algorithms,” *Big Data & Society* 3, no. 1 (January 5, 2016), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2660674.

> This article describes three different kinds of “opacity” in algorithms: (1) opacity as intentional corporate or state secrecy, (2) opacity as technical illiteracy, and (3) opacity arising from the characteristics of machine learning algorithms that make them useful. Burrell argues that recognizing these distinct forms of opacity is important to determining what technical and non-technical solutions can prevent algorithms from causing harm. On (1), secrecy may be essential to the proper function of an algorithm (such as to prevent it from being gamed), but such algorithms are easily reviewable by trusted and independent auditors. Regarding (2), the solution to technical illiteracy is simply greater public education. Finally, (3) is difficult because there may be a trade-off between fairness, accuracy, and interpretability. Certain AI techniques could be avoided in fields where transparency is crucial, or new benchmarks could be developed to assess such algorithms for discrimination and other issues. *(Category: In-depth)*

Defining the Problem: What's at Stake?

Navneet Alang, "Turns Out Algorithms are Racist," *New Republic*, August 31, 2017. https://newrepublic.com/article/144644/turns-algorithms-racist/.

> This article explains that AI is only as good as the data it is fed. Computers and software, even at their most sophisticated, are essentially input-output systems that are "taught" by feeding them enormous amounts of data. Hence if the input data reflect gender, racial, or other biases, so too will the output. (*Category: Concise overview)*

Robert Hart, "If you're not a white male, artificial intelligence's use in healthcare could be dangerous," Quartz. July 10, 2017, https://qz.com/1023448/if-youre-not-a-white-male-artificial-intelligences-use-in-healthcare-could-be-dangerous/.

> While acknowledging the potential of AI to revolutionize healthcare, this article points to the danger that this technology will perpetuate healthcare inequalities due to its reliance on existing stores of medical data. Groups including women (especially pregnant women), the young, and the elderly are excluded from many medical research studies, which may result in errors when individuals from these groups are treated by AI systems trained on this data. (*Category: Case study – healthcare*)

J. Kleinberg et al., "Human Decisions and Machine Predictions," National Bureau of Economic Research, February 2017, https://www.cs.cornell.edu/home/kleinber/w23180.pdf.

> This article highlights how machine learning can help humans make better decisions, using the example of judges making bail decisions. While acknowledging the difficulties associated in fully automating bail determinations—both due to biases in the training data that would be fed into such a system and the complex mix of factors that judges weigh—AI-based analyses show that the number of people jailed before trial can be substantially reduced without impacting the crime rate. Such insights can then be used by judges to improve their own decision-making. (*Category: Case study – criminal justice)*

danah boyd, Karen Levy, and Alice Marwick, *The Networked Nature of Algorithmic Discrimination,* Washington, DC: New America Foundation, 2014, https://www.danah.org/papers/2014/DataDiscrimination.pdf.

> This study points to the dangers surrounding algorithms making predictions about you based on those you associate with—such as your friends and neighbors. While existing laws prohibit racial and gender discrimination, among other things, an individual's position in a social network is deeply affected by these and other variables—which algorithms might then use to make predictions about us, leading to unfair results. Consequently, the authors argue that we need to rethink our models of discrimination to consider not just an individual's immutable characteristics, but also the impacts of how algorithms position us within a network and society, too. (*Category: Novel issues – networked discrimination)*

Approaches to Regulating AI

S. Wachter et al., "Transparent, explainable, and accountable AI for robotics," *Science Robotics* 2, no. 6 (May 31, 2017), http://robotics.sciencemag.org/content/2/6/eaan6080.full.

> This article provides a brief overview of the challenges facing governments seeking to regulate AI. After briefly grounding the regulation of automated systems in its historical context, it raises the critical questions facing regulators seeking to enact optimal laws in the space. (*Category: Concise overview*)

IEEE Global Initiative on Ethics of and Autonomous and Intelligent Systems, "Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems" (IEEE, 2017), https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf

> This document, prepared by the world's largest professional organization in the technology space, calls for an ethics- and values-based approach to dealing with the impacts of intelligent and autonomous systems that prioritizes human well-being within a given cultural context. (*Category: In-depth analysis*)

Corrine Cath et al., "AI and the 'Good Society': the US, EU, and UK approach," *SSRN Electronic Journal* (2016). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2906249

> In October 2016, the White House, the European Parliament, and the UK House of Commons each issued reports outlining their vision on how to prepare society for the widespread use of AI. This article provides a comparative assessment of these three reports to facilitate the design of policies favorable to the development of a 'good AI society'. (*Category: Comparative overview)*

National Science and Technology Council: Committee on Technology, *Preparing for the Future of Artificial Intelligence*. Washington, D.C.: Executive Office of the President, October 2016. https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf.

> This report surveys the current state of AI research, its existing and potential applications, and the questions that AI raises for society as a whole and for public policy in particular. The report recommends certain specific governmental and non-governmental actions within the U.S. context and sets forth a strategic plan for U.S. government funding of AI. (*Category: Regulatory template)*

Dillon Reisman et al., *Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability*. New York University AI Now Institute, April 2018. https://ainowinstitute.org/aiareport2018.pdf.

> This report proposes an "Algorithmic Impact Assessment" framework to be used by public entities in the U.S. in procuring and deploying AI systems. The framework is designed to support affected communities and stakeholders as they seek to assess the claims made about automated decision systems, and to determine where their use is acceptable. It offers key elements for the construction of a public agency algorithmic impact assessment, and a practical accountability framework combining agency review and public input. (*Category: Regulatory template)*

Business and AI

Sherif Elsayed-Ali, “Why Embracing Human Rights Will Ensure AI Works for All,” April 2018, https://www.weforum.org/agenda/2018/04/why-embracing-human-rights-will-ensure-AI-works-for-all/.

> This article recommends four actions to prevent discriminatory outcomes in machine learning based on the UN Guiding Principles on Business and Human Rights. These are: (1) active inclusion of underrepresented populations in datasets and AI development, (2) fairness in the interpretation of biased data, (3) a right to understand how algorithmic decisions are made, and (4) access to redress when wrong decisions are made. (*Category: Concise overview*)

R. Jorgenson, “What Platforms Mean When They Talk About Human Rights,” Policy & Internet 9, no. 3 (May 29, 2017) https://doi.org/10.1002/poi3.152.

> This article examines how two major Internet platforms—Google and Facebook—make sense of human rights. Based on primary research, the authors find that the companies frame themselves as strongly committed to human rights. Yet this framing focuses primarily on government rights violations, rather than the companies’ own adverse impacts on its users’ rights. (*Category: Case study—Internet platforms)*

Shift, Oxfam and Global Compact Network Netherlands, Doing Business with Respect for Human Rights: A Guidance Tool for Companies, 2016, https://www.businessrespecthumanrights.org/image/2016/10/24/business_respect_human_rights_full.pdf.

> This report provides comprehensive guidance to businesses on what they should do to operationalize their responsibility to respect human rights, as recognized in the UN Guiding Principles on Business and Human Rights. (*Category: In-depth)*