

White Paper

How to Prevent Discriminatory Outcomes in Machine Learning

Global Future Council on Human Rights 2016-2018

March 2018



Contents

3	Foreword
4	Executive Summary
6	Introduction
7	Section 1: The Challenges
8	Issues Around Data
8	What data are used to train machine learning applications?
8	What are the sources of risk around training data for machine learning applications?
9	What-if use case: Unequal access to loans for rural farmers in Kenya
9	What-if use case: Unequal access to education in Indonesia
9	Concerns Around Algorithm Design
9	Where is the risk for discrimination in algorithm design and deployment?
10	What-if use case: Exclusionary health insurance systems in Mexico
10	What-if scenario: China and social credit scores
11	Section 2: The Responsibilities of Businesses
11	Principles for Combating Discrimination in Machine Learning
13	Bringing principles of non-discrimination to life: Human rights due diligence for machine learning
14	Making human rights due diligence in machine learning effective
15	Conclusion
16	Appendix 1: Glossary/Definitions
17	Appendix 2: The Challenges – What Can Companies Do?
23	Appendix 3: Principles on the Ethical Design and Use of AI and Autonomous Systems
22	Appendix 4: Areas of Action Matrix for Human Rights in Machine Learning
29	Acknowledgements

Foreword

The World Economic Forum Global Future Council on Human Rights works to promote practical industry-wide solutions to human rights challenges in context of the unfolding Fourth Industrial Revolution. This paper evolved from conversations among members of the Council and dozens of experts in the fields of human rights and machine learning.

Using machines to find patterns in large quantities of data, and make predictions from these patterns, is unlocking many new kinds of value – from better ways to diagnose cancer to enabling self-driving cars – and creating new opportunities for individuals: machine translation, for example, can break down linguistic barriers, and voice recognition can empower illiterate people. Our council has chosen to focus on how companies designing and implementing this technology can maximize its potential benefits. This work heeds the call of the Forum’s Founder and Executive Chairman, Klaus Schwab, for “ethical standards that should apply to emerging technologies,” which he rightly says are “urgently needed to establish common ethical guidelines and embed them in society and culture.”¹

This white paper offers a framework for understanding the potential risks for machine learning applications to have discriminatory outcomes, in order to arrive at a roadmap for preventing them. While different applications of ML will require different actions to combat discrimination and encourage dignity assurance, in this white paper we offer a set of transferable, guiding principles that are particularly relevant for the field of machine learning. We base our approach on the rights enshrined in the Universal Declaration of Human Rights and further elaborated in a dozen binding international treaties that provide substantive legal standards for the protection and respect of human rights and safeguarding against discrimination.²

Our emphasis on risks is not meant to undersell the promise of machine learning, nor to halt its use. The concern around discriminatory outcomes in machine learning is not just about upholding human rights, but also about maintaining trust and protecting the social contract founded on the idea that a person’s best interests are being served by the technology they are using or that is being used on them. Absent that trust, the opportunity to use machine learning to advance our humanity will be set back.

Many companies have begun to explore the ideas of fairness, inclusion, accountability, and transparency in machine learning, including Microsoft, Google, and Deepmind (Alphabet). Pervasive and justifiable concerns remain that efforts to promote transparency and accountability might undermine these companies’ IP rights and trade secrets, security and in some cases the right to privacy. However, with these systems continuing to influence more people in more socially sensitive spaces (housing, credit, employment, education, healthcare, etc.), and mostly in the absence of adequate government regulation – whether due to technology outpacing regulatory mechanisms, lack of government capacity, political turmoil, or other unfavorable conditions – we need more active self-governance by private companies.

Erica Kochi

Co-Chair of the Global Future Council on Human Rights
Co-Founder of UNICEF Innovation

¹ Schwab, “The Fourth Industrial Revolution,” Geneva: World Economic Forum, 2016, 90

² H.R.C. Res. 20/L.13, U.N.Doc.A/HRC/20/L.13 (June 29, 2012), accessed September 11, 2017, http://ap.ohchr.org/documents/dpage_e.aspx?si=A/HRC/RES/20/8

Executive Summary

Machine learning systems are already being used to make life-changing decisions: which job applicants are hired, which mortgage applicants are given a loan, which prisoners are released on parole. Such decisions affect human rights, often of the most vulnerable people in society.

Designed and used well, machine learning systems can help to eliminate the kind of human bias in decision-making that society has been working hard to stamp out. However, it is also possible for machine learning systems to reinforce systemic bias and discrimination and prevent dignity assurance. For example, historical data on employment may show women getting promoted less than men. If a machine learning system trained on such data concludes that women are worse hires, it will perpetuate discrimination.

Discriminatory outcomes not only violate human rights, they also undermine public trust in machine learning. If public opinion becomes negative, it is likely to lead to reactive regulations that thwart the development of machine learning and its positive social and economic potential.

The challenges

While algorithmic decision-making aids have been used for decades, machine learning is posing new challenges due to its greater complexity, opacity, ubiquity, and exclusiveness.

Some challenges are related to the data used by machine learning systems. The large datasets needed to train these systems are expensive either to collect or purchase, which effectively excludes many companies, public and civil society bodies from the machine learning market. Training data may exclude classes of individual who do not generate much data, such as those living in rural areas of low-income countries, or those who have opted out of sharing their data. Data may be biased or error-ridden.

Even if machine learning algorithms are trained on good data sets, their design or deployment could encode discrimination in other ways: choosing the wrong model (or the wrong data); building a model with inadvertently discriminatory features; absence of human oversight and involvement; unpredictable and inscrutable systems; or unchecked and intentional discrimination.

There are already examples of systems that disproportionately identify people of color as being at “higher risk” for committing a crime, or systematically exclude people with mental disabilities from being hired. Risks are especially high in low- and middle-income countries, where existing inequalities are often deeper, training data are less available, and government regulation and oversight are weaker.

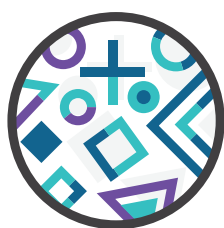
While ML has implications for many human rights, not least the right to privacy, we focus on discrimination because of the growing evidence of its salience to a wide range of private-sector entities globally, including those involved in data collection or algorithm design or who employ ML systems developed by a third party. The principle of non-discrimination is critical to all human rights, whether civil and political, like the rights to privacy and freedom of expression, or economic and social, like the rights to adequate health and housing.

The responsibilities of business

Governments and international organizations have a role to play, but regulations tend not to keep pace with technological development. This white paper makes the case that businesses need to integrate principles of non-discrimination and empathy into their human rights due diligence – a process by which businesses take ongoing, proactive, and reactive steps to ensure that they do not cause or contribute to human rights abuses.

Under international human rights law, all companies should respect human rights. According to the UN Guiding Principles on Business and Human Rights, the responsibility to respect human rights “exists over and above compliance with national laws and regulations protecting human rights.” That is, even if there is a lack of regulation specifically about machine learning, human rights principles and obligations still apply.

Drawing on existing work, we propose four central principles to combat bias in machine learning and uphold human rights and dignity:



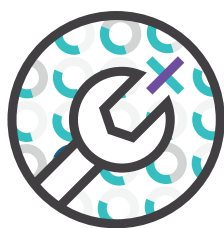
● **Active Inclusion**



● **Fairness**



● **Right to Understanding**



● **Access to Remedy**

- **Active Inclusion:** The development and design of ML applications must actively seek a diversity of input, especially of the norms and values of specific populations affected by the output of AI systems.
- **Fairness:** People involved in conceptualizing, developing, and implementing machine learning systems should consider which definition of fairness best applies to their context and application, and prioritize it in the architecture of the machine learning system and its evaluation metrics.
- **Right to Understanding:** Involvement of ML systems in decision-making that affects individual rights must be disclosed, and the systems must be able to provide an explanation of their decision-making that is understandable to end users and reviewable by a competent human authority. Where this is impossible and rights are at stake, leaders in the design, deployment and regulation of ML technology must question whether or not it should be used.

- **Access to Redress:** Leaders, designers and developers of ML systems are responsible for identifying the potential negative human rights impacts of their systems. They must make visible avenues for redress for those affected by disparate impacts, and establish processes for the timely redress of any discriminatory outputs.

We recommend three steps for companies:

1. **Identifying human rights risks linked to business operations.** We propose that common standards for assessing the adequacy of training data and its potential bias be established and adopted, through a multi-stakeholder approach.
2. **Taking effective action to prevent and mitigate risks.** We propose that companies work on concrete ways to enhance company governance, establishing or augmenting existing mechanisms and models for ethical compliance.
3. **Being transparent about efforts to identify, prevent, and mitigate human rights risks.** We propose that companies monitor their machine learning applications and report findings, working with certified third-party auditing bodies in ways analogous to industries such as rare mineral extraction. Large multinational companies should set an example by taking the lead. Results of audits should be made public, together with responses from the company.

We recognize that much of our work is still speculative, given the nascent nature of ML applications, particularly in the Global South, and the incredible rate of change, complexity, and scale of the issues. We hope this report will both advance internal corporate discussions of these topics and contribute to the larger public debate. Following the release of this white paper, our hope is to actively work with members of the Forum to see how these recommendations fit into the business practices of a variety of private-sector players working to build and engage machine learning applications. Compared with prior waves of technological change, we have an unprecedented opportunity to prevent negative implications of ML at an early stage, and maximize its benefits for millions.

Introduction

In contrast to traditional programming, in which people hand-code the solution to a problem step-by-step, a machine learning (ML) system sifts through data, recognizes patterns, and automates decision-making based on its discoveries. Machine learning is a kind of artificial intelligence (AI; for a glossary of terms, see Appendix 1). The nuances of how it works may be difficult for non-experts to digest, but its promise is plain: increased efficiency, accuracy, scale and speed in making decisions and finding the best answers to questions ranging from “What type of illness is this?” to “What should you do next?”

ML systems could potentially increase fairness in making decisions about which humans can be biased. A system for sifting job applications might, for example, ensure that women or ethnic minority candidates are fairly considered. However, ML systems can also do the opposite – reinforcing the kinds of systemic bias and discrimination that society has been working hard to stamp out. While ML systems are still nascent even in developed economies, there are already examples: in *Weapons of Math Destruction*, Cathy O’Neil cites systems that disproportionately identify people of color as being at “higher risk” for committing a crime, or systematically exclude people with mental disabilities from being hired.

ML applications are already being used to make many life-changing decisions – such as who qualifies for a loan, whether someone should be given parole, or what type of care a child should receive from social service programs. These decisions affect human rights, especially of society’s most vulnerable: as framed by the Universal Declaration of Human Rights, a pillar of the international legal system since 1948, “the idea of human rights is as simple as it is powerful: that all people are free and equal, and have a right to be treated with dignity.” Machine learning can be disproportionately harmful in low- and middle-income countries, where existing inequalities are often deeper, training data are less available, and government regulation and oversight are weaker.

Many current ML applications might not seem relevant to human rights, such as the image recognition systems used to tag photos on social media. However, it is easy to conceive of scenarios in which they become so: image recognition systems can, for example, identify a person’s sexual orientation with reasonable accuracy – consider how they might be used by governments in countries

where homosexuality is illegal. The potential for bias and discrimination goes well beyond sectors such as lending, insurance, hiring, employment, and education. As Cathy O’Neil says, “Predictive models are, increasingly, the tools we will be relying on to run our institutions, deploy our resources, and manage our lives.”

Discriminatory outcomes not only violate human rights, they also undermine public trust in machine learning. If public opinion about machine learning becomes negative, it is likely to lead to reactive regulations that are poorly informed, unimplementable, and costly – and that thwarts the development of machine learning and close off myriad opportunities to use it for good by augmenting the capabilities of individuals and opening up new ways to apply their talents. A new model is needed for how machine learning developers and deployers address the human rights implications of their products.

³ “Accelerating innovation through responsible AI,” PWC

⁴ Derek Hawkins, “Researchers use facial recognition tools to predict sexual orientation. LGBT groups aren’t happy,” *The Washington Post*, <https://www.washingtonpost.com/news/morning-mix/wp/2017/09/12/researchers-use-facial-recognition-tools-to-predict-sexuality-lgbt-groups-arent-happy/>

Section 1: The Challenges

Algorithmic decision-making aids have been used for decades – banks, for instance, automating mathematical functions to score the eligibility of credit applicants. Such experiences show that algorithms can discriminate in unexpected ways. For example, race is not included in US data sets on credit applicants, but the use of proxy indicators such as zip-codes can still result in racial minorities' access to credit being unfairly limited. Regulations have been adopted to prevent such accidental discrimination. But machine learning is posing new challenges, due to its greater complexity, opacity, ubiquity, and exclusiveness.



Complexity

Past algorithmic decision-making systems relied on rules-based, “if/then” reasoning. ML systems create more complex models in which it is difficult to trace decisions back to ask questions about why and how they were made. They

offer no “logical flow that a human can understand, which is very different from traditional software,” explains [Guy Katz](#), a postdoctoral research fellow in computer science at Stanford.⁵



Opacity

In past systems, one could easily determine the source of a discriminatory decision and put in place ways to prevent it. Machine learning systems are more opaque, due not only to

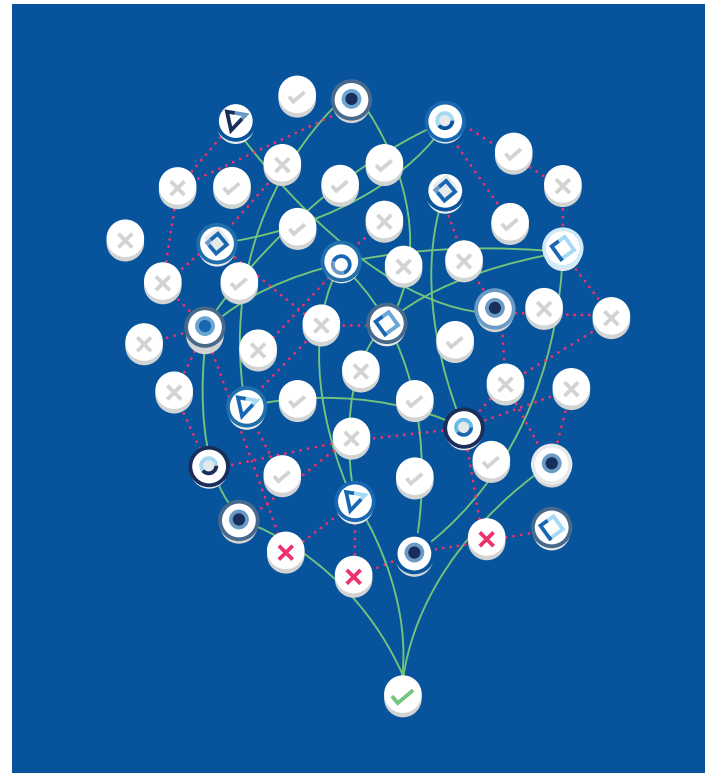
their complexity but also to the proprietary nature of their algorithms. Lack of transparency and auditability contributes to the popular understanding of ML systems as “black boxes.”



Ubiquity

Many people, particularly in the US and Europe, already “interact with machine learning-driven systems on a daily basis.”⁶ Examples come from the New York Times: “Algorithms can decide

where kids go to school, how often garbage is picked up, which police precincts get the most officers, where building code inspections should be targeted, and even what metrics are used to rate a teacher.”⁷



Exclusiveness

ML systems require massive data sets to learn from, and programmers with technical education. Both exclude huge subsets of people. ML systems today are almost entirely

developed by small, homogenous teams, most often of men.⁸ Data sets are often proprietary and require large-scale resources to collect or purchase. While great strides are being made in open-source sharing of datasets and transfer learning technology that minimizes the data needed to develop ML systems, companies who own massive proprietary datasets still have definite advantages.

These challenges manifest in two categories: related to the data itself, and related to the way algorithms are designed, developed, and deployed.⁹ Appendix 2 summarizes what companies can do to tackle each of the issues explored in this section.

⁵ Marina Krakovsky, “Finally a Peek Inside the ‘Black Box’ of Machine Learning Systems,” <https://engineering.stanford.edu/magazine/article/finally-peek-inside-black-box-machine-learning-systems>

⁶ “Machine learning: The Power and Promise of Computers,” Royal Society, <https://royalsociety.org/-/media/policy/projects/machine-learning/publications/machine-learning-report.pdf> p. 16

⁷ Jim Dwyer, “Showing the Algorithms Behind New York City Services,” *The New York Times* https://www.nytimes.com/2017/08/24/nyregion/showing-the-algorithms-behind-new-york-city-services.html?_r=0

⁸ Kate Crawford, “Artificial Intelligence’s White Guy Problem,” *The New York Times*, https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html?_r=0

⁹ While there are important new challenges around machine learning technologies, it is worth bearing in mind that past recommendations to eliminate bias in computer systems hold value and relevance. In 1996, Friedman and Nissenbaum identified three categories of “bias in computer systems,” including: preexisting (roots in social institutions, practices, and attitudes); technical (arises from technical constraints or considerations); and emergent (arises in a context of use). Read more, “[Bias in Computer Systems](#)”

Issues Around Data

Training data, the foundation of machine learning

What data are used to train machine learning applications?

Machine learning requires data on which to train. For example, ML applications to determine eligibility for credit/lending, housing, or insurance traditionally draw on factors such as historical salary ranges, payment and debt histories, family situations and residence status. For education and employment opportunities, the historical data used can include grades, time spent in prior jobs, and number of promotions received.

In some countries, regulations prevent the use of factors such as gender, race, religion or marital status to determine access to credit/lending, housing, insurance, education, and employment. In others, such regulations do not exist or there are not enough resources or political will to enforce them.

Increasingly, social media and mobile usage data inform decisions such as who is likely to be a credible loan recipient or a good hire.¹⁰ They enable lenders or employers to assess an applicant's spending habits, risky behavior, work and education histories, and professional networks. Very little regulation exists in this area, and the implications are less well understood. For example, if an application's training data demonstrates that people who have influential social networks or who are active in their social networks are "good" employees, that application might filter out people from lower-income backgrounds, those who attended less prestigious schools, or those who are more cautious about posting on social media.

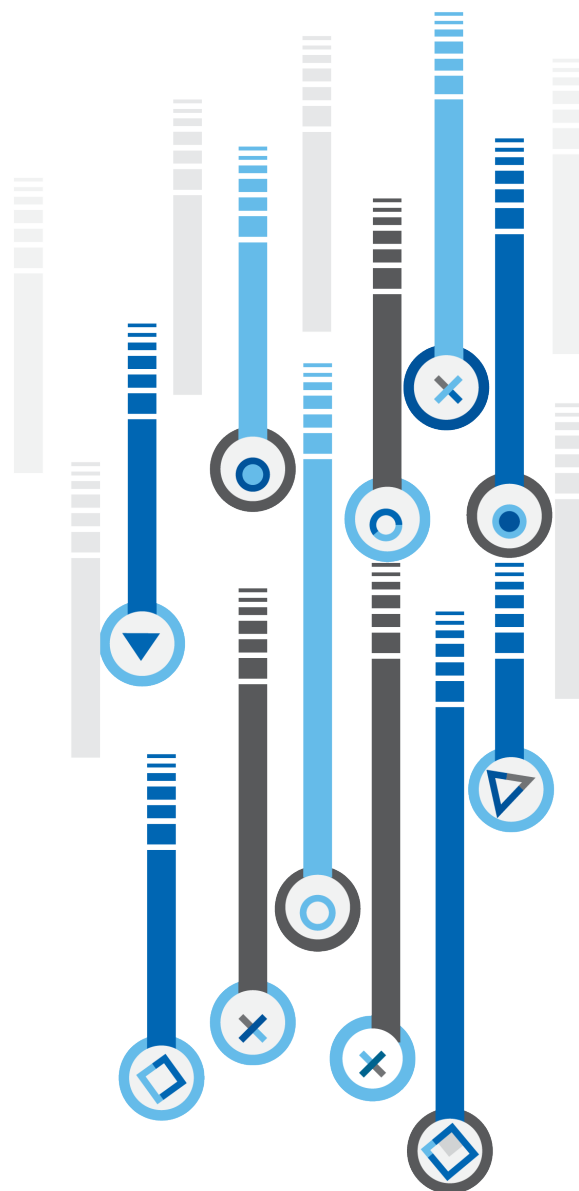
What are the sources of risk around training data for machine learning applications?

Data availability

In many cases, data belong to the mobile network operator, internet platform, or other service provider collecting them. Often the companies that generate or purchase data choose to keep them private. Companies, public and civil society bodies that lack the resources to purchase or collect data are effectively excluded from participating in the machine learning market.

Often, groups that generate a smaller digital footprint include those who have traditionally been discriminated against and people in low-income countries. For example, a household

¹⁰ "Is It Time for Consumer Lending to Go Social? How to Strengthen Underwriting and Grow Your Customer Base with Social Media Data," PWC, <https://www.pwc.com/us/en/consumer-finance/publications/social-media-in-credit-underwriting-process.html>



in the US with just one home automation product can generate a data point every six seconds;¹¹ in Mozambique, where about 90% of the population lack internet access, the average household generates zero digital data points. In South Asia, 38% fewer women than men own a phone; women in low- and middle-income countries report using phones less frequently and intensively than men do, especially for mobile internet.¹² Those who live in rural areas have a thinner digital footprint: 10% of the global population lack access to basic voice and text services, and 30% to 3G/4G mobile broadband internet, mostly in rural Asia and sub-Saharan Africa.¹³

¹¹ "Internet of Things, Privacy and Security in a Connected World," FTC, <https://www.ftc.gov/system/files/documents/reports/federal-trade-commission-staff-report-november-2013-workshop-entitled-internet-things-privacy/150127iotrpt.pdf>

¹² "Bridging the Gender Gap: Mobile Access and Usage in Low- and Middle-Income Countries," GSMA, https://www.gsma.com/mobilefordevelopment/wp-content/uploads/2016/02/GSM0001_03232015_GSMARepor_NEWGRAYS-Web.pdf

¹³ "Rural Coverage: Strategies for Sustainability," GSMA, <https://www.gsmainelligence.com/research/?file=53525bcdac7cd801eccef740e001fd92&download>

What-if use case: Unequal access to loans for rural farmers in Kenya



In Kenya, microloan company Tala's smartphone app collects data on loan applicants including the number of people they contact daily, their movements and routine habits, like whether they call their mother every day or pay their bills on time. Tala suggests that using these inputs to gauge credit risk offers an alternative pathway to credit

for those who lack a credit history.¹⁴ However, there are risks: as rural Kenyans have less digital infrastructure and fewer opportunities to develop a digital footprint, might they be unfairly excluded by algorithms trained on data points captured from more urban populations?

Biased or error-ridden data



The computing law of “garbage in, garbage out” dictates that training ML systems on limited, biased or error-strewn data will lead to biased models and discriminatory outcomes. For example, historical data on employment will often show women getting promoted less than men – not because women are worse at their jobs, but because

workplaces have historically been biased. ML systems trained to understand women as worse hires than men will continue to favor men, and continue to generate discriminatory baseline data.

Data mining, “the automated process of extracting useful patterns from large data sets, and in particular, patterns that can serve as a basis for subsequent decision-making,”¹⁵ is especially sensitive to statistical bias because it helps to discover patterns that organizations tend to treat as generalizable even though the analyzed data includes only a partial sample from a circumscribed period. To ensure that data mining reveals patterns that hold true more widely, the sample must be proportionally representative of the population.

¹⁴ Kathleen Siminyu, “Artificial Intelligence in Low/Middle Income Countries; The East African Experience,” <http://kathleensiminyu.com/2017/09/14/artificial-intelligence-in-lowmiddle-income-countries-the-east-african-experience/>

¹⁵ Solon Barocas, “Big Data's Disparate Impact”

What-if use case: Unequal access to education in Indonesia



In Indonesia, economic development has unfolded disparately across geographical (and, subsequently, ethnic) lines. For example, while access to higher education is relatively uniform across the country,

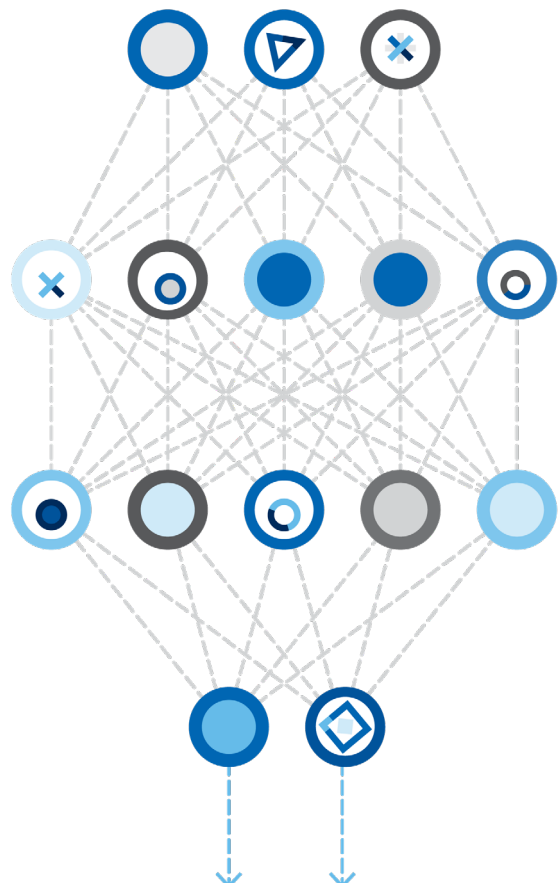
the top 10 universities are all on the island of Java, and a large majority of the students who attend those universities are from Java. As firms hiring in white-collar sectors train ML systems to screen applicants based on factors like educational attainment status, they may systematically exclude those from poorer islands such as Papua.

Concerns Around Algorithm Design

Modeling for fairness

Where is the risk for discrimination in algorithm design and deployment?

Even if machine learning algorithms are trained on good data sets, their design or deployment could encode discrimination in five main ways.



1. Choosing the wrong model

Algorithms are often designed based on other algorithms that have proven successful in ostensibly similar contexts – but algorithms that work well in one context may discriminate in another.¹⁶ ML systems used for predictive policing, for example, are based on earthquake modeling; but as earthquakes are recorded more consistently than crimes, predictive policing models may skew towards overpredicting crime in areas where reporting rates are higher. Similarly, an ML algorithm which successfully assesses relative risk of applicants for loans in the US may overlook relevant data points if deployed in other countries.

2. Building a model with inadvertently discriminatory features

Humans have to define for algorithms what “success” looks like – and it usually means maximizing profits or accuracy or efficiency, rather than maximizing fairness.¹⁷ For example, one ML model tasked with predicting likelihood to re-offend had a similar error rate for black and white defendants, but was more likely to err by wrongly predicting that black defendants would re-offend and that white defendants would not.¹⁸ When humans specify what weight ML algorithms should give to variables, this can create bias: for example, an algorithm to assess loan applicants may consider both income levels and reliability of past repayments; a human decision to give more weight to the former may unfairly discriminate against members of groups which tend to be lower-income, such as women. AI teams have a tendency to develop conforming, self-perpetuating approaches, which hinder their ability to innovate and spot incorrect outputs.¹⁹

3. Absence of human oversight and involvement

As machine learning becomes more sophisticated, it includes less human supervision. However, having a human in the loop is necessary to notice where important factors are being unexpectedly overlooked. For example, the University of Pittsburgh Medical Center used ML to predict which pneumonia patients were at low risk of developing complications and could be sent home. The ML model recommended that doctors send home patients who have asthma, having seen in the data that very few developed complications; doctors, however, knew this was only because they routinely placed such patients in intensive care as a precaution. Because it is impossible to define in advance when discrimination may happen in any given context, humans need to be kept involved and systems made interpretable for them.^{20,21}

¹⁶ Calders, T., Zliobaite, I., “Why unbiased computational processes can lead to discriminative decision procedures,” in B. Custers, T. Calders, B. Schermer, T. Zarsky (eds.), *Discrimination and Privacy in the Information Society*, pp. 43–57 (2013), as cited in “[Algorithmic Accountability](#)”

¹⁷ Interview with Cathy O’Neil

¹⁸ <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

¹⁹ Input from Svetlana Sicular, Gartner

²⁰ For a useful overview of what interpretability means in machine learning, please read Lipton 2016, “The Mythos of Model Interpretability,” <https://arxiv.org/abs/1606.03490>

²¹ Aaron Bornstein, “Is Artificial Intelligence Permanently Inscrutable?” *Nautilus*, September 1, 2016, <http://nautil.us/issue/40/learning/is-artificial-intelligence-permanently-inscrutable>

4. Unpredictable and inscrutable systems

When a human makes a decision, such as whether or not to hire someone, we can inquire as to why he or she decided one way or the other. ML systems lack this transparency and traceability. Sometimes this does not matter – we may not need to understand how the ML behind Google Maps determines our suggested route. When decisions impact on rights, however, it is imperative: for instance, when an ML system makes decisions on parole, identifying and remedying possible discrimination depends on being able to understand the steps taken to reach the decisions. Criminal justice, public housing, welfare and health provision are examples of areas where “black box” systems should not be developed or used.

5. Unchecked and intentional discrimination

In some cases, bias is intentionally built into algorithms. For instance, if employers want to avoid hiring women who are likely to become pregnant, they might employ ML systems to identify and filter out this subset of women. In the absence of adequate regulation, the burden lies with the company leadership, designers, data scientists, engineers, and others involved in creating ML systems to build them in ways that predict, prevent, and monitor bias.

What-if use case: Exclusionary health insurance systems in Mexico



Mexico is among countries where, for most, quality healthcare is available only through private insurance. At least two private multinational insurance companies operating in Mexico are now using ML to maximize their efficiency and profitability, with potential implications for the human right to fair access to adequate healthcare. Imagine a scenario in which insurance companies use ML to mine data such as shopping history to recognize patterns associated with high-risk customers, and charge them more: the poorest and sickest would be least able to afford access to health services.

What-if scenario: China and social credit scores



While few details are publicly available, reports suggest that China is creating a model to score its citizens by analyzing a wide range of data from banking, tax, professional, and performance records, to smartphones, e-commerce, and social media.²² The aim is speculated to be “to use the data to enforce a moral authority as designed by the Communist Party.”²³ One open question is what it will mean if governments act on scores computed using data that is incomplete, historically biased, and using models not built for “fairness”.

²² Julie Makinen, “China Prepares to Rank its Citizens on Social Credit.” *The Los Angeles Times*. November 2015. <http://www.latimes.com/world/asia/la-fg-china-credit-system-20151122-story.html>

²³ Julie Makinen, “China Prepares to Rank its Citizens on Social Credit.” *The Los Angeles Times*. November 2015. <http://www.latimes.com/world/asia/la-fg-china-credit-system-20151122-story.html>

Section 2: The Responsibilities of Businesses

Under international human rights law, while states have the primary obligation to uphold human rights, all companies should respect human rights. According to the UN Guiding Principles on Business and Human Rights,²⁴ this responsibility is “a global standard of expected conduct for all business enterprises wherever they operate.”²⁵ The responsibility to respect human rights “exists independently of States’ abilities and/or willingness to fulfil their own human rights obligations”, and “exists over and above compliance with national laws and regulations protecting human rights.”²⁶

Given the complex nature of ML and rapid pace of technical development, most governments are unlikely to be able to develop legal and regulatory frameworks to protect human rights in the deployment and use of ML in a timely and effective manner. Occasionally, regulators get ahead of widespread deployment of new technologies – for example, Germany has introduced laws on self-driving vehicles.²⁷ But many governments and regulators are still struggling today with questions that first arose with the rise of the internet in the mid-1990s, such as the role of intermediaries in relation to content, and the limits of privacy.

Even if there is a lack of regulation specifically about machine learning, human rights principles and obligations still apply. In the context of ML, these include:

- Ensuring respect for the principle of non-discrimination, including by using representative training data for the particular use case, addressing bias in training data, and designing algorithms in a way that does not favor particular groups over others;
- Ensuring that ML applications that could manifestly violate human rights are not developed or used (for example, systems that could be used to predict a person’s sexual orientation and thus be used to persecute LGBTI people);
- Ensuring that ML applications that prevent people from enjoying their human rights or actively put them at risk of human rights violations are not used (such as black box systems in the provision of public services that deny people access to effective redress).

²⁴ The UNGPs were endorsed by the UN Human Rights Council in 2011. UN Office of the High Commissioner for Human Rights, *Guiding Principles on Business and Human Rights: Implementing the United Nations “Protect, Respect and Remedy” Framework (2011)*, online at http://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf

²⁵ UN Guiding Principles on Business and Human Rights (UNGPs), commentary to principle 11.

²⁶ UNGPs, commentary to principle 11

²⁷ Germany Adopts Self-Driving Vehicles Law,” Reuters, online at: <https://www.reuters.com/article/us-germany-autos-self-driving/germany-adopts-self-driving-vehicles-law-idUSKBN1881HY>, 12 May 2017.



Principles for Combating Discrimination in Machine Learning

Emerging communities of researchers, businesses, and developers are thinking about machine learning’s social, economic, and ethical implications in our everyday lives, and how we might design systems that maximize human benefit. Notable recent initiatives to define principles for the ethical and accountable use of AI, summarized in Appendix 3, include:

- **The Asilomar Principles (2017)** – on the safe, ethical, and beneficial use of AI; developed by the Future of Life Institute and endorsed by leading figures including Elon Musk and Stephen Hawking.
- **The FATML (Fairness, Accountability and Transparency in Machine Learning) Principles (2016)** – on accountable algorithms; developed by a large network of scientists, researchers, and industry professionals.
- **The Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems (2017)** – developed by the Institute of Electrical and Electronics Engineers (IEEE), the world’s largest technical professional organization for the advancement of technology.

In focusing here on the human right to non-discrimination, we synthesize the existing body of work into four critical principles to combat the reinforcement of bias in machine learning.



● Active Inclusion

The development and design of ML applications must involve a diversity of input, especially of the norms and values of specific populations affected by the output of AI

systems. Individuals must give explicit consent before the system can use protected or sensitive variables²⁸ (such as race, religion, gender) or his or her personal data to make decisions.

Guiding Questions:

- How diverse is the pool of designers involved in the creation of the system?
- Have we evaluated the veracity of the data and considered alternative sources?
- Have we mapped and understood if any particular groups may be at an advantage or disadvantage in the context in which the system is being deployed?
- Have we sufficiently researched and taken into account the norms of the context in which the system is being deployed?
- Have we calculated the error rates and types for different sub-populations and assessed the potential differential impacts?



● Fairness

There are many different ways of defining fairness; people involved in conceptualizing, developing, and implementing machine learning systems should consider which

definition best applies to their context and application. In every case, fairness and the dignity of affected people should be prioritized in the architecture of the machine learning system and its evaluation metrics, as issues with bias are long-term and structural.²⁹

Guiding Questions:

- Have we identified a definition of fairness that suits the context and application for our product and aligns with the International Declaration of Human Rights?
- Have we included all the relevant domain experts whose interdisciplinary insights allow us to understand potential sources of bias or unfairness and design ways to counteract them?
- Have we mapped and understood if any particular groups may be at an advantage or disadvantage in the

²⁸ Simply removing sensitive categories from datasets is not sufficient; undesirable biases are often preserved (unobserved) in the remaining attributes. One solution is to use the sensitive attributes to test the system behavior and to find and correct the discrimination/ bias.

²⁹ In "Fairness in Criminal Justice Risk Assessments: The State of the Art" Berk et al, 2017 provide a through review of the technical pathways towards promoting fairness in machine learning. Berk et al, 2017, Fairness in Criminal Justice Risk Assessments: The State of the Art <https://arxiv.org/abs/1703.09207>

context in which the system is being deployed?

- Have we applied "rigorous pre-release trials to ensure that [the ML system] will not amplify biases and error due to any issues with the training data, algorithms, or other elements of system design?"³⁰
- Have we outlined an ongoing system for evaluating fairness throughout the life cycle of our product? Do we have an escalation/emergency procedure to correct unforeseen cases of unfairness when we uncover them?



● Right to Understanding

If ML systems are involved in decision-making that affects individual rights, this must be disclosed. The systems must be able to provide an explanation of their decision-

making that is understandable to end users and reviewable by a competent human authority.

Guiding Questions:

- Have we logged all sources of potential AI error and uncertainty?
- Have we openly disclosed what aspects of the decision-making are algorithmic?
- How much of our data sources have we made transparent?
- How much of our system can we explain to end users?
- How much of our ML-related code and procedures have we made open-source?
- Have we provided detailed documentation, technically suitable APIs, and permissive use of terms to allow third parties to provide and review the behavior of our system?



● Access to Redress

The designers and developers of ML systems are responsible for the use and actions of their systems. They must make visible avenues for redress for those affected by disparate

impacts, and establish processes for the timely redress of any discriminatory outputs.

Guiding Questions

- How confident are we in the decision-making output of the algorithm?
- Do intended decision-makers understand the probabilistic nature of algorithms, recognizing that outputs will not be 100% correct and amending errors?
- Do we have a method for checking if the output from an algorithm is decorrelated from protected or sensitive features?
- Have we tested a series of counterfactual propositions to track if the results of the algorithm would be different if the end user was a different race or age or lived elsewhere?
- What reporting processes and recourse do we have in place?
- Do we have a process in place to make necessary fixes to the design of the system based on reported issues?

³⁰ Ai Now Institute 2017 Report

Bringing principles of non-discrimination to life: Human rights due diligence for machine learning

Companies developing and using ML systems must integrate these principles of non-discrimination into their human rights due diligence – a process by which businesses take ongoing, proactive, and reactive steps to ensure that they uphold people’s dignity and do not cause or contribute to human rights abuses. This responsibility lies not only with the engineers building ML models: the goal of leadership should be to steer ML technology to uphold human rights. Where organizations have existing discrimination policies, and industries have standards to protect human rights, these must be updated to reflect new considerations pertaining to ML; where such policies and industry standards are absent, they must be developed and implemented. The nature of human rights due diligence will differ depending on whether a company is a developer or user of ML, the particular use cases, and the potential impact on human rights is. In all circumstances, businesses must take three core steps.

Step 1: Identifying human rights risks linked to business operations

Companies engineering and/or implementing ML systems have a responsibility to map human rights risks before and during the life cycle of the product – from development to deployment and use. Developers and leadership should take into account risks inherent in ML, as defined in Section 1. In deployment, the nature of the use, identity of the end user, and developer’s human rights record can lead to different assessments: for example, a system might have no human rights risks when used for an airline’s customer service but could impact the human right to housing if used by a mortgage company.

Step 2: Taking effective action to prevent and mitigate risks

For leadership, this step requires establishing a framework and incentives for ML development teams to prioritize positive human rights outcomes. For developers, this step requires detecting and correcting for data bias and ensuring that data sets (including training data) represent the populations an ML application will affect. For example, software for sifting through job applications should not use training data that embeds existing discriminatory practices against women or minorities. Often, this requires developers to consult with external domain experts as well as end users and clients. For instance, for a developer of an ML application to determine eligibility for mortgages, it would be important to consult with public and nonprofit bodies that work on housing issues. Where the use of machine learning systems can potentially have a significant impact on human rights, companies should seek independent auditing of algorithms based on agreed-upon industry standards and the human rights framework. Businesses using ML should have ongoing human-in-the-loop checks to identify and amend any bias in the system.

Step 3: Being transparent about efforts to identify, prevent, and mitigate human rights risks

For leadership, this step involves explicitly encouraging transparency. For developers, transparency would include explaining the process of identifying human rights risks, the risks that have been identified, and the steps taken to prevent and mitigate them. When possible, use of open-source software can improve transparency; when not possible, companies can publish technical papers to explain the design and workings of their ML applications. Transparency also requires that people know when ML has been used to make a decision that impacts them.

Making human rights due diligence in machine learning effective

The promise of the human rights due diligence process has often not been realized due to weaknesses in its implementation. Current due diligence models have resulted in inconsistencies in identifying and mitigating human rights risks between companies in the same industry and similar operating environments. For example, in response to risks to the right to privacy from mass surveillance programs and cybercrime, Apple applied default end-to-end encryption to its iMessage and Facetime platforms; Google applied it to its Duo service only as an option; and Blackberry did not apply it to its regular BBM service at all.³¹

Human rights due diligence involves self-assessment and self-reporting. There is no independent assessment of how well a company is doing, except where independent research is undertaken by media or civil society. These shortcomings must be addressed and a more robust, independent, and trusted process adopted.

In relation to identifying human rights risks linked to business operations, we propose that common standards for assessing the adequacy of training data and its potential bias be established and adopted, along with common minimum standards and procedures for identifying human rights risks in ML system design. Where industry-specific standards already exist, these should be strengthened and adapted to account for new challenges related to ML. A multi-stakeholder approach is key. Initiatives like the AI for Good Summit or the Partnership on AI could be focal points if they include of a diverse range of companies, civil society, and academics. Existing initiatives by the IEEE, FATML, and others can provide a solid basis. The participation of agencies such as the Human Rights Council and the Office of the High Commissioner for Human Rights is critical for legitimacy and international applicability.

In relation to taking effective action to prevent and mitigate risks, we propose that company leadership work on concrete ways to enhance company governance over ML activities. This will require augmenting existing mechanisms and models for ethical compliance where such tools are already established. For instance, in the credit industry, existing standards for evaluating and enforcing fair lending should be expanded to address ML. Where there are no existing internal codes of conduct or accountability schemes, they should be developed, taking an inclusive approach.

In relation to being transparent about efforts to identify, prevent, and mitigate human rights risks, we propose that companies monitor their ML applications and report findings. We suggest working with certified third-party auditing bodies to evaluate the effects of policies and practices on human rights, analogous to industries such as rare mineral extraction. Large multinational companies should set an example by taking the lead in submitting to such independent audits; in the future, there may be an opportunity to establish an international independent auditing body to carry out evaluations on a global scale. Results of audits should be made public, together with responses from the company.

Appendix 4 contains matrices which go into further detail on areas for companies regarding these steps.

³¹ See Amnesty International, *How Private are you favourite messaging apps*, 21 October 2016, online at: <https://www.amnesty.org/en/latest/campaigns/2016/10/which-messaging-apps-best-protect-your-privacy/>

Conclusion

The application of human rights standards to machine learning is a very recent topic of inquiry, and the recommendations in this paper are among the first to be developed and published in this area. We expect that they will be further developed and elaborated by others. These recommendations are meant to function not as a universal manual, but as a useful starting point for companies (from leadership through to development teams), building on any existing mechanisms in their sector. We encourage readers to choose the elements from these recommendations that are relevant to them, and integrate them as best fits their individual needs and context.³²

This white paper has sought to move non-discrimination as a human rights issue to the center of the discussion about the potential social impacts of machine learning, and to expand the focus of these concerns to include parts of the world that are currently absent from the conversation. In our exploration of this emerging and complex subject, we have sought to identify areas (geographic, industry-specific, technical) where discrimination in machine learning is most likely to impact human rights, evaluate where businesses' responsibilities lie in addressing algorithmic discrimination, and present the realistic ways forward in overcoming these challenges.

Identifying and eliminating bias or discrimination that can result from machine learning applications is not an easy task. Following our recommendations, companies can work together with domain experts, stakeholders, and relevant partners from both the public and private sectors to leverage machine learning in a way that includes and benefits people, and prevents discrimination. In doing so, they will not only cultivate huge value for society, but also build public trust and reduce the risks of reactive, poorly informed regulation that can be confusing, unimplementable, and at times very economically costly.

There is no one-size-fits-all solution to eliminate the risk of discrimination in machine learning systems, and we recognize that many of our recommendations will require context-specific tailoring. We further recognize that much of our work is still speculative, given the nascent nature of ML applications, particularly in the Global South. Having undertaken this research over the last eight months, we are mindful of the incredible rate of change, complexity, and scale of the issues that companies face when integrating machine learning into their business models.

We hope this report will both advance internal corporate discussions of these topics and contribute to the larger public debate.

Following the release of this white paper, our hope is to actively work with members of the Forum to see how these recommendations fit into the business practices of a variety of private-sector players working to build and engage machine learning applications.

³² For a more technically rigorous review of available tools for creating accountable algorithms, we recommend Kroll et al.'s impressive article "Accountable Algorithms". University of Pennsylvania Law Review, October 2016

Appendix 1: Glossary/Definitions

Algorithm

An algorithm is a formally specified sequence of logical operations that provides step-by-step instructions for computers to act on data and thus automate decisions. Algorithms play a role in both automating the discovery of useful patterns in data sets and automating decision-making that relies on these discoveries.³³ In simpler terms, it is “a set of rules a computer follows to solve a problem.”³⁴

Algorithmic Accountability

“The responsibility of algorithm designers to provide evidence of potential or realised harms.”³⁵

Artificial Intelligence (AI)

The science of making machines smart.³⁶

Auditability

The ability for “third parties to probe, understand, and review the behavior of the algorithm through disclosure of information that enables monitoring, checking, or criticism, including through provision of detailed documentation, technically suitable APIs, and permissive terms of use.”³⁷

Big Data

“Large and heterogeneous forms of data that have been collected without strict experimental design. Big data is becoming more common due to the proliferation of digital storage, the greater ease of acquisition of data (e.g. through mobile phones) and the higher degree of interconnection between our devices (i.e. the internet).”³⁸

Discrimination

“Any distinction, exclusion or preference made on the basis of race, colour, sex, religion, political opinion, national extraction or social origin, which has the effect of nullifying or impairing equality of opportunity or treatment.”³⁹

Human Rights

Human rights are rights inherent to all human beings, whatever our nationality, place of residence, sex, national or ethnic origin, color, religion, language, or any other status. We are all equally entitled to our human rights without discrimination. These rights are all interrelated, interdependent, and indivisible. Universal human rights are often expressed and guaranteed by law, in the forms of treaties, customary international law, general principles, and other sources of international law. International human rights law lays down obligations of governments to act in certain ways or to refrain from certain acts in order to promote and protect human rights and fundamental freedoms of individuals or groups.⁴⁰ Businesses have a responsibility to respect human rights. “This means that they should avoid infringing on the human rights of others and should address adverse human rights impacts with which they are involved.”⁴¹

Machine Learning

A machine learning (ML) model is one that leverages computer programs that automatically improve with experience and more data.⁴² There are three main subsets of machine learning: supervised learning, unsupervised learning, and reinforcement learning. Deep learning is a powerful class of learning techniques and models that can be used in all of these settings, and the mechanics and implications of deep learning are outside of the scope of this paper. More in-depth definitions of these subsets can be found here: <https://medium.com/towards-data-science/types-of-machine-learning-algorithms-you-should-know-953a08248861>

Training Data

Data that is used as an input to a machine learning algorithm in the process of populating (a.k.a., training) the machine learning model such that the trained model represents the patterns contained in the training data.

Transparency

The ability to “know when and why a machine learning system performs well or badly.”⁴³

³³ “Big Data’s Disparate Impact”

^{34,36} “Machine Learning: The Power and Promise of Computers That Learn by Example,” Royal Society, <https://royalsociety.org/~media/policy/projects/machine-learning/publications/machine-learning-report.pdf>

³⁵ “Algorithmic Accountability”

³⁷ (fatml.org)

³⁸ “Machine Learning: The Power and Promise of Computers That Learn by Example,” Royal Society, <https://royalsociety.org/~media/policy/projects/machine-learning/publications/machine-learning-report.pdf>

³⁹ Human Rights Committee in its General Comment 18 on Non-Discrimination (The ICCPR itself does not provide a definition of discrimination)

⁴⁰ United Nations Human Rights, <http://www.ohchr.org/EN/Issues/Pages/WhatAreHumanRights.aspx>

⁴¹ “Guiding Principles on Business and Human Rights: Implementing the United Nations ‘Protect, Respect and Remedy’ Framework (2011),” UN Office of the High Commissioner for Human Rights, online at http://www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf, principle 11.

⁴² Tom Mitchell, *Machine Learning*, 1997.

⁴³ (Royal Society)

Appendix 2: The Challenges – What Can Companies Do?

What can companies do to avoid issues with incomplete data?

Action	Impact
Organize research	Determine whether certain data sets fit internally agreed upon standards of “adequate” and “representative” data (looking to both quantitative and qualitative metrics); identify opportunities to expand data collection efforts where contextually appropriate, viable, and possible to do so without violating privacy
Ensure diversity in ML development teams	Bring different perspectives together; afford insights into whether certain populations are adequately included and represented in training data
Participate in Open Source data and algorithm sharing	Collect data from diverse sources in a format that is free and accessible to all in order to get a more representative and far reaching spread of data.
Build out harmonized standards for Data labelling	All companies will benefit from greater transparency requirements around licensed datasets. This will be particularly important for startups/ smaller companies who are not resourced to undergo extensive testing prior to release
Develop standards to track the provenance, development, and use of training data sets throughout their life cycle ⁴⁴	Better understand and monitor issues of potential bias and biases that may already be at work ⁴⁵
Map out risks	Have a sense of what could go wrong in order to be able to course correct (either by expanding the data set or changing the way the machine learning system is designed and deployed) if necessary
Engage stakeholders and domain experts in participatory manner	Better identify the entire range of data types necessary to adequately train an ML system for a given context; better understand how to appropriately source the data needed
Train ML designers/developers and AI leaders on human rights responsibilities	Equip technical teams and leadership with the knowledge and ability to translate human rights responsibilities into code, making it easier to avoid discriminatory outcomes
Promote transparency and understandability in ML systems/ applications	Allow domain experts to have a say in what data sets might be inadequate or invasive to use. Provide a mechanism for a safe feedback from the audience to which AI is delivered.

What can companies do to avoid issues with biased data

Action	Impact
Organize research	Uncover whether certain data sets fit widely accepted standards of “fair” and “representative” data (looking to both quantitative and qualitative metrics)
Ensure diversity in ML development teams	Bring different perspectives together; afford insights into whether certain populations are adequately included and represented in training data
Map out risks	Have a sense of what could go wrong in order to be able to course correct (either by expanding the data set or changing the way the machine learning system is designed and deployed) if necessary
Develop standards to track the provenance, development, and use of training data sets throughout their life cycle ⁴⁶	Better understand and monitor issues of potential bias and biases that may already be at work ⁴⁷
Engage stakeholders and domain experts in participatory manner	Better identify the entire range of data types necessary to adequately train an ML system for a given context; better understand how to appropriately source the data needed
Train ML designers/developers and AI leaders on human rights responsibilities	Equip technical teams with the knowledge and ability to translate human rights responsibilities into code, making it easier to avoid discriminatory outcomes
Promote transparency and understandability in ML systems/ applications	Allow domain experts to have a say in what data sets might be inadequate or invasive to pull from. Provide a mechanism for a safe feedback from the audience to which AI is delivered.

^{44,45, 46, 47} Language and concept from the [AI Now Institute 2017 Report](#)

What can companies do to avoid choosing the wrong model or data?

Action	Impact
Organize research	Have an up-to-date understanding of how certain models have performed in similar contexts to guide model and data selection
Ensure diversity in ML development teams	Bring different perspectives together; afford insights into which model types and data types may need to be considered in order to design an ML system that is both accurate and non-discriminating
Monitor ML model's use across different contexts and communities ⁴⁸	Ensure that the ML applications don't introduce errors or bias as cultural assumptions and domains shift ⁴⁹
Keep models up to date and contextually relevant	Reduce chances of bias and error that can result from static ML applications that no longer reflect the real-time realities and needs of a given context
Map out risks	Have a sense of what could go wrong in order to identify what stages will require human-in-the-loop checks and how to leverage dynamic testing
Include dynamic testing	Determine how algorithms are performing according to a chosen set of indicators that reflect non-discrimination in order to course correct (either by changing the training data, target variables, parameters, cost functions, or other elements of the ML application) if necessary
Engage stakeholders and domain experts in participatory manner	Best identify what types of considerations should be made for an ML model being applied in a particular domain (industry, geography, population, etc.) to design a fair and contextually appropriate ML model
Train ML designers/developers and AI leaders on human rights responsibilities	Have leaders and data scientists who are able to translate ethics into code that runs the ML systems; minimize risk of inadvertent or blatant discrimination

What can companies do to avoid building a model with discriminatory features?⁵⁰

Action	Impact
Organize research	Have an up-to-date understanding of how certain models have performed in similar contexts to guide model and data selection
Ensure diversity in ML development teams	Bring different perspectives together; afford insights into which model types and data types may need to be considered in order to design an ML system that is both accurate and non-discriminating
Keep models up to date and contextually relevant	Reduce chances of bias and error that can result from static ML applications that no longer reflect the real-time realities and needs of a given context
Map out risks	Have a sense of what could go wrong in order to identify what stages will require human-in-the-loop checks and how to leverage dynamic testing; determine what set of indicators could be used to detect discrimination and might be helpful for dynamic testing
Include dynamic testing	Determine how algorithms are performing according to a chosen set of indicators that reflect non-discrimination in order to course correct (either by changing the training data, target variables, parameters, cost functions, or other elements of the ML application) if necessary
Calibrate models to include fairness criteria where appropriate	Balance a model's success according not only to accuracy but also to fairness and non-discrimination
Engage stakeholders and domain experts in participatory manner	Best identify what types of considerations should be made for an ML model being applied in a particular domain (industry, geography, population, etc.) to design a fair and contextually appropriate ML model
Train ML designers/developers and AI leaders on human rights responsibilities	Have leaders and data scientists who are able to translate ethics into code that runs the ML systems; minimize risk of inadvertent or blatant discrimination

⁴⁸ "Top 10 Recommendations for the AI Field in 2017," Medium, <https://medium.com/@AINowInstitute/the-10-top-recommendations-for-the-ai-field-in-2017-b3253624a7>

⁴⁹ From the *AI Now Institute 2017 Report*

⁵⁰ Here we recommend actions that most companies working on designing and implementing ML systems can take to minimize risks for discrimination resulting from the ML algorithms themselves. For a more technically rigorous and specific set of guidelines, we highly recommend Kroll et al.'s report on "[Accountable Algorithms](#)," (University of Pennsylvania Law Review, 2016).

What can companies do to ensure human involvement and oversight?⁵¹

Action	Impact
Organize research	Have an up-to-date understanding of how certain models have performed in similar contexts to guide model and data selection; contribute to a shared body of knowledge to inform standards for auditing and understanding ML systems ⁵²
Ensure diversity in ML development teams	Bring different perspectives together; afford insights into which model types and data types may need to be considered in order to design an ML system that is both accurate and non-discriminating
Map out risks	Have a sense of what could go wrong in order to identify what stages will require human-in-the-loop checks and how to leverage dynamic testing; determine what set of indicators could be used to detect discrimination and might be helpful for dynamic testing
Engage stakeholders and domain experts in participatory manner	Best identify what types of considerations should be made for an ML model being applied in a particular domain (industry, geography, population, etc.) to design a fair and contextually appropriate ML model
Train ML designers/developers and AI leaders on human rights responsibilities	Have leaders and data scientists who are able to translate ethics into code that runs the ML systems; minimize risk of inadvertent or blatant discrimination

What can companies do to avoid issues with unpredictable and inscrutable systems?

Action	Impact
Promote transparency and understandability in ML systems/applications	Build in ability to ask how a decision is made, thereby promoting accountability and ability to act accordingly to put limits on the sources of risk for discrimination ⁵³ Provide a mechanism for a safe feedback from the audience to which AI is delivered.
Map out risks	Provide a sense of what could go wrong in order to identify what stages will require human-in-the-loop checks and how to leverage dynamic testing; determine what set of indicators could be used to detect discrimination and might be helpful for dynamic testing
Include dynamic testing	Provide accountability when ML applications are inscrutable to better understand how ML systems are treating certain subgroups within a population and identify discrimination.
Engage stakeholders and domain experts in participatory manner	Best identify what types of transparency and scrutability will be particularly critical for a certain domain (for instance, in ML systems used to score applicants for hireability, it is important to be able to trace and identify what variables are taken into account and to understand how they are weighted in order to be sure they are calculated in non-discriminating ways)

⁵¹ One of the biggest advantages of ML applications is the ability to compute at a pace that no human could ever hope to keep up with. Not only would it be unreasonable to have a human checking every single computation an ML application executes, but it would also be a hindrance to the technology's benefits. Here, we suggest instead that companies need to keep human oversight at a few vital moments in the ML design, monitoring, and deployment stages. These moments will be different across different applications, and some will be more rigorous in their human-resource needs than others.

⁵² [AI Now 2017 Report](#)

⁵³ For a good overview of the different definitions of interpretable / explainable machine learning entails, we recommend Doran et al's article "[What Does Explainable AI Really Mean? A New Conceptualization of Perspectives](#)". Doran et al 2017, What Does Explainable AI Really Mean? A New Conceptualization of Perspectives, <https://arxiv.org/abs/1710.00794>

What can companies do to avoid intentional discrimination using ML?

Action	Impact
Ensure diversity in ML development teams	Bring different perspectives together; afford insights into which model types and data types may need to be considered in order to design an ML system that is both accurate and non-discriminating
Create an internal code of conduct based on human rights framework	Guide those involved in designing and interpreting ML-generated decisions to understand when there is a violation of the code of conduct that constitutes discrimination
Create an incentive model for adherence to human rights guidelines	Encourage people to avoid discrimination to create a company culture that promotes human rights and seeks to eliminate both intentional and inadvertent discrimination
Promote transparency and understandability in ML	Allow users or those monitoring ML systems to understand when a model is built with discrimination as a desired outcome and hold the relevant parties accountable. Provide a mechanism for a safe feedback from the audience to which AI is delivered.
Map out risks	Have a sense of what could go wrong in order to identify what stages will require human-in-the-loop checks and how to leverage dynamic testing; determine what set of indicators could be used to detect discrimination and might be helpful for dynamic testing
Include dynamic testing	Determine how algorithms are performing according to a chosen set of indicators that reflect non-discrimination in order to course correct (either by changing the training data, the target variables, parameters, cost functions, or other elements of the ML application) if necessary
Calibrate models to include fairness criteria where appropriate	Create automatic checks and balances in the ML system that might be able to prevent discrimination even when it is intended; balance a model's success according not only to accuracy but also to fairness and non-discrimination
Train ML designers/developers and AI leaders on human rights responsibilities	Have leaders and data scientists who are able to translate ethics into code that runs the ML systems; minimize risk of inadvertent or blatant discrimination
Restrict ML deployment in cases where it is judged incongruous with human rights	Protect people from discriminatory outcomes in the most sensitive application contexts; limit human rights abuses

Appendix 3: Principles on the Ethical Design and Use of AI and Autonomous Systems

	Asilomar Principles (Ethics and Values) on Safe, Ethical, and Beneficial use of AI*	FATML Principles for Accountable Algorithms	IEEE Principles on Ethically Aligned Design*
Safety/Security/ Accuracy (Verifiability)	Safety – AI systems should be safe and secure throughout their operational lifetime, and verifiably so where applicable and feasible	Accuracy – Identify, log, and articulate sources of AI error and uncertainty throughout the algorithm and its data sources so that expected and worst-case implications can be understood and inform mitigation procedures	Human Benefit (Safety) – AI must be verifiably safe and secure throughout its operational lifetime
Transparency/ Explainability/ Auditability	Failure Transparency – If systems cause harm, it should be possible to ascertain why Judicial Transparency – If systems are involved in key judicial decision-making, an explanation that is auditable by a competent human authority should be made available	Explainability – Ensure that algorithmic decisions, as well as any data driving those decisions, can be explained to end users and other stakeholders in nontechnical terms Auditability – Enable interested third parties to probe, understand, and review the behavior of the algorithm through disclosure of information that enables monitoring, checking, or criticism, including through the provision of detailed documentation, technically suitable APIs, and permissive use of terms	Transparency/Traceability – It must be possible to discover how and why a system made a particular decision or acted in a certain way, and, if a system causes harm, to discover the root cause
Responsibility	Responsibility – Designers and builders of AI systems are stakeholders in the moral implications of their use, misuse, and actions	Responsibility – Make available externally visible avenues of redress for adverse individual or societal effects, and designate an internal role for the person who is responsible for the timely remedy of such issues	Responsibility – Designers and developers of systems should remain aware of and take into account the diversity of existing relevant cultural norms; manufacturers must be able to provide programmatic-level accountability proving why a system operates in certain ways
Fairness and Values Alignment	Shared Benefit – AI technologies should benefit and empower as many people as possible Shared Prosperity – The economic prosperity created by AI should be shared broadly, to the benefit all of humanity Non-Subversion – The power conferred by control of highly advanced AI systems should respect and improve, rather than subvert, social and civic processes	Fairness – Ensure that algorithmic decisions do not create discriminatory or unjust impacts when comparing across different demographics	Embedding Values into AI – Identify the norms and elicit the values of a specific community affected by a particular AI, and ensure the norms and values included in AI are compatible with the relevant community Human Benefit (Human Rights) – Design and operate AI in a way that respects human rights, freedoms, human dignity, and cultural diversity
Privacy	Personal Privacy – People should have the right to access, manage, and control the data they generate, given AI systems' power to analyze and utilize that data Liberty and Privacy – The use of personal data by AI must not unreasonably curtail people's real or perceived liberty		Personal Data and Individual Access Control – People must be able to define, access, and manage their personal data as curators of their unique identity

Notes: (*) Both the Asilomar and IEEE guidelines include additional principles that relate to human control, the avoidance of lethal autonomous weapons arms race, and long-term capability questions regarding the beneficence of artificial general intelligence and superintelligence.







Appendix 4: Areas of Action Matrix for Human Rights in Machine Learning

● Fairness
 ● Active inclusion
 ● Right to Understanding
 ● Access to Redress

Identifying human rights risks linked to business operations

Need		Action	Tasks	Actors
Assess Wider Impacts	●	Organize research	<p>Hire and designate people to oversee research on ways to maximize the benefits of ML while preventing human rights violations.</p> <p>Assign individuals/ taskforce to build out a strategic approach to preventing negative outcomes in ML</p>	<p>Businesses involved in developing or deploying machine learning systems, starting with company leadership that makes the strategic decisions for how AI should be developed/ deployed. Companies like Google, Microsoft, Amazon, and Facebook are already involved in dedicated research efforts in this space.</p> <p>Public-sector entities involved in deploying machine learning systems.</p>
	●		<p>“Before releasing an AI system, companies should run rigorous pre-release trials to ensure that they will not amplify biases and errors due to any issues with the training data, algorithms, or other elements of system design. As this is a rapidly changing field, the methods and assumptions by which such testing is conducted, along with the results, should be openly documented and publicly available, with clear versioning to accommodate updates and new findings.”⁵⁴</p> <p>Map human rights risks throughout the life cycle of machine learning products, from development to deployment and use; this mapping should take into account risks inherent in machine learning, including data bias and inadequate data, and must include the intended uses and the potential for human rights abuses in each case</p> <p>Update human rights risks for each new use case of a ML application</p>	<p>Businesses involved in developing or deploying machine learning systems</p> <p>Public-sector entities involved in deploying machine learning systems, such as has been done by New York City where an initial mapping of risks led the City Council to consider a bill to ensure transparency and testing of algorithmic decision-making systems</p>





⁵⁴ “The 10 Top Recommendations for the AI Field in 2017” AiNow Institute <https://medium.com/@AINowInstitute/the-10-top-recommendations-for-the-ai-field-in-2017-b3253624a7>

Develop and/or enhance Industry Standards			Develop or augment standards to evaluate fairness, inclusion, and accountability in machine learning	Companies with the capacity to do so should partake in industry-wide efforts to arrive at a common understanding and set of standards for fairness and non-discrimination and dignity assurance in machine learning.	Businesses involved in developing or deploying machine learning systems. This might look similar to approaches taken to develop standards for fairness in trade in the Fair Trade movement. Public-sector entities involved in deploying machine learning systems
			Build out harmonized standards for data labelling	Companies should work together to develop and use a widely understood transparency taxonomy to label data accurately that will allow companies licensing in data to know what they are getting.	Businesses whose commercial model involves licensing datasets
			Develop standards to track the provenance, development, and use of training data sets throughout their life cycle ⁵⁵	Designate appropriate employees (data scientists, data anthropologists, etc.) to develop better records for how a training data set was created and maintained Continue to examine existing training data sets and work to understand potential blind spots and biases that may already be at work ⁵⁶ This should also include evaluation of ethical outcomes delivery. Gartner research predicted that by 2019, more than 10% of IT hires in customer service will mostly write scripts for bot interactions. As such, it is not just input data but also output has a potential for discrimination.	Social scientists and measurement researchers within the AI bias research field (like AI Now, Partnership on AI, IEEE, FATML)

⁵⁵ Language and concept from the [AiNow Institute 2017 Report](#)

⁵⁶ [Ai Now 2017 Report](#)

Taking effective action to prevent and mitigate risks

Need		Action	Tasks	Actors
Enhance Company Governance		Create an internal code of conduct based on human rights framework	<p>Businesses should work internally to “define and promote the use of a code of conduct for responsible use of data and algorithms” with an explicit priority around preventing human rights abuses.⁵⁷ There may be opportunities to integrate such a code of conduct into existing workplace conduct guidelines, such as sexual harassment protocol. In this model, new principles would be drafted to outline what qualifies as discrimination for the relevant products being developed within the company, and how each person involved in a discriminatory model would be considered and held responsible.⁵⁸</p> <p>When possible, consortia of businesses should work together to verify whether these internal codes of conducts align with a wider vision for non-discriminatory applications of machine learning. We hope that the work of many organizations seeking to establish standards for non-discrimination will soon yield a shared understanding of what these internal codes of conduct should look like.</p>	<p>Businesses involved in developing or deploying machine learning systems (starting with company leadership) working together with governments and, where viable, multi-sector oversight bodies, such as the consortia already at work in the inclusive/fair AI space (AI Now, the Partnership on AI, etc.) and think tanks/NGOs focused on “AI for Good” and ethical applications for AI</p> <p>Public-sector entities involved in deploying machine learning systems</p>
		Create an incentive model for adherence to human rights guidelines	<p>Work with management teams across the business verticals to integrate the internal code of conduct into employee incentive models from training programs through to C-level management; models of this include employing causal reasoning⁵⁹</p> <p>Exert influence on the relevant industry to embrace openness, accountability, and human rights in their machine learning applications</p>	<p>Businesses involved in developing or deploying machine learning systems, starting with company leadership that makes the strategic decisions for how AI should be developed/ deployed.</p> <p>Public-sector entities involved in deploying machine learning systems</p>
		Develop or augment standards to evaluate fairness, inclusion, and accountability in machine learning	Companies with the capacity to do so should partake in industry-wide efforts to arrive at a common understanding and set of standards for fairness and non-discrimination in machine learning.	<p>Businesses involved in developing or deploying machine learning systems, starting with company leadership that makes the strategic decisions for how AI should be developed/ deployed. This might look similar to approaches taken to develop standards for fairness in trade in the Fair Trade movement.</p> <p>Public-sector entities involved in deploying machine learning systems</p>
		Develop standards to track the provenance, development, and use of training data sets throughout their life cycle ⁶⁰	<p>Designate appropriate employees (data scientists, data anthropologists, etc.) to develop better records for how a training data set was created and maintained</p> <p>Continue to examine existing training data sets and work to understand potential blind spots and biases that may already be at work⁶¹</p>	<p>Company leadership that makes the strategic decisions for how AI should be developed/ deployed</p> <p>Social scientists and measurement researchers within the AI bias research field (like AI Now, Partnership on AI, IEEE, FATML)</p>

⁵⁷ Citing “Algorithmic Accountability: Applying the concept to different country concepts” from the Web Foundation

⁵⁸ Idea put forward by Richard Socher in an Interview on 9/13/2017.

⁵⁹ See Kilbertus et al., 2017

⁶⁰ Language and concept from the AiNow Institute 2017 Report

⁶¹ Ai Now 2017 Report

Take an inclusive approach to design	●	●	Engage stakeholders and domain experts in participatory manner	<p>Design engagement strategies to include users and stakeholders in defining the algorithmic features and parameters</p> <p>Detect and correct for data bias and ensure that data sets (including training data) are adequate by consulting with end users, their clients, and external domain experts</p> <p>“Intensify diversity, equity, and inclusivity efforts to go beyond human resources and allowed to effectively influence the approach towards product development and services provision by the company.”</p>	Corporations, designated task forces (where applicable), relevant stakeholders, and users of a service/product/tool/platform
	●	●	Ensure diversity in ML development teams	<p>Create an explicit internal commitment to inclusionary hiring, not just across the company but within the ML design/development teams</p> <p>Build diversity and inclusion principles into human resources practices and guidelines and set goals for each that are appropriate for the company's context and size; design and carry out sporadic check-ins or internal audits to evaluate how the company is doing in its diversity goals and outline steps that can be taken to promote diversity when needed</p> <p>Periodically add new team members and rotate in temporary team members from other areas to bring fresh perspectives to the AI team. (The problem of AI teams is not in grasping new ideas and technologies, but in established approaches that are difficult to change.)</p> <p>“Companies, universities, conferences and other stakeholders in the AI field should release data on the participation of women, minorities and other marginalized groups within AI research and development. Many now recognize that the current lack of diversity in AI is a serious issue, yet there is insufficiently granular data on the scope of the problem, which is needed to measure progress. Beyond this, we need a deeper assessment of workplace cultures in the technology industry, which requires going beyond simply hiring more women and minorities, toward building more genuinely inclusive workplaces.”⁶²</p>	Hiring managers/HR at companies developing and deploying ML systems
	●	●	Train ML designers/developers and AI leaders on human rights responsibilities	<p>Require training course/certification on human rights</p> <p>Expand curriculum for all ML architects/designers to include coursework on human rights and data science ethics</p> <p>Develop modular design programs for human rights in ML that can be adapted and integrated into existing ethics or non-discrimination curricula</p> <p>Invest in education for ML engineers in the global south to promote sustained participation from low and middle income countries. Examples of initiatives doing this include Data Science Africa, The CODATA-RDA School of Research Data Science, and Deep Learning Indaba.</p>	<p>Higher learning institutions</p> <p>Civil society organizations</p> <p>Corporations hiring ML architects/designers</p> <p>For example: the Blue Sky Agenda for AI 165 Education, a collection of ideas for ethics education in AI, seeks democratization of AI education and emphasizes inclusiveness in development toward the goal of respecting the values and rights of diverse populations¹</p>

Optimize ML models for fairness, accountability, transparency, and editability	●	●	Calibrate models to include fairness criteria	<p>In general, calibrating false positive and false negative rates in each group or population for which an algorithm is making decisions can help to equalize impacts. In other words, the individuals responsible for designing algorithms and weighting variables should ask the question, “When this system fails, who will it fail for, and how can we prevent that failure?”⁶⁴</p> <p>Johndrow and Lum’s article, “An algorithm for removing sensitive information: application to race-independent recidivism prediction,” provides a thorough analysis of how machine learning algorithms can be employed to augment fairness in AI-backed decision making.⁶⁵</p>	Businesses developing ML systems- starting with company leadership that makes the strategic decisions for how AI should be developed/ deploye
	●	●	Include dynamic testing	<p>Create and integrate quality assessment indicators that include fairness and accountability⁶⁶</p> <p>Where appropriate, integrate dynamic testing procedures to provide accountability either by⁶⁷:</p> <ul style="list-style-type: none"> – Employing cryptographic commitments (equivalents of sealed documents held by third party or in a safe place) – Fair random choices (a technique allowing software to make fully reproducible random choices) – Zero knowledge proofs (cryptographic tools that allow a decision-maker to prove that the decision policy that was actually used has a certain property without revealing either how the property is known or what the decision policy is) 	<p>ML development teams</p> <p>Teams tasked with monitoring and evaluating ML applications once they are implemented</p>

⁶⁴ Interviews with Cathy O’Neil and Joshua Cohen

⁶⁵ Johndrow and Lum 2017, An algorithm for removing sensitive information: application to race-independent recidivism prediction, <https://arxiv.org/abs/1703.04957>

⁶⁶ Datta, Sen, and Zick “Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems” (<http://www.fatml.org/schedule/2016/presentation/algorithmic-transparency-quantitative-input>)

⁶⁷ All pulled from “Algorithmic Accountability” citing Kroll et al. (2017, February)

Being transparent about efforts to identify, prevent, and mitigate human rights risks

Need		Action	Tasks	Actors
Monitor and Refine Algorithms		Organize human oversight	Allocate increased resources to human monitoring and evaluation: “The methods and outcomes of monitoring should be defined through open, academically rigorous processes, and should be accountable to the public. Particularly in high stakes decision-making contexts, the views and experiences of traditionally marginalized communities should be prioritized.” ⁶⁸	Public-sector entities and businesses involved in deploying machine learning systems
			Where possible, designate an internal or partnered task force for oversight of human rights concerns in machine learning applications	Independent auditing bodies appointed by businesses developing or using ML
			Design and conduct audits, or hire an external firm to audit and to evaluate potential risks for discrimination related to: <ul style="list-style-type: none"> – Input data – Decision factors – Output decisions⁶⁹ Invest in quality controls to oversee data collection processes, including human-in-the-loop verification (e.g. involving human operators within AI-backed decision making systems) ⁷⁰ Undertake checks on an ongoing basis to ensure that the decisions/outcomes produced by AI systems are not biased, and to correct bias in the system	These might resemble similar auditing practices from other industries, such as the FLO certification model for fair trade standards, and the MSC chain of custody surveillance audits for sustainable products
		Monitor ML model's use across different contexts and communities ⁷¹	Assign technical experts, domain experts, and managers involved in the implementation of an AI model to ensure that the ML applications don't introduce errors or bias as cultural assumptions as domains shift ⁷² Create a process for monitoring systems throughout their life cycle	Technical experts and domain experts working with a given ML application
		Keep models up to date and contextually relevant	Depending on the context, models will need to be updated, whether with new training data, new parameters, new target variables, or other technical components; such updates should be prioritized and scheduled based on context	Data scientists, engineers, and algorithm designers that work within companies that are implementing machine learning systems

⁶⁸ “Top 10 Recommendations for the AI Field in 2017”

^{69, 70} Citing “Algorithmic Accountability: Applying the concept to different country concepts” from the Web Foundation

⁷¹ “Top 10 Recommendations for the AI Field in 2017” <https://medium.com/@AINowInstitute/the-10-top-recommendations-for-the-ai-field-in-2017-b3253624a7>

⁷² Campolo, Alex; Sanfilippo, Madelyn; Whittaker, Meredith; Crawford, Kate. Ai Now 2017 Report. Ai Now Institute. 2017. https://assets.contentful.com/8wprhvnpc0/1A9c3ZTCZa2KEYM64Wsc2a/8636557c5fb14f2b74b2be64c3ce0c78/_AI_Now_Institute_2017_Report_.pdf

<p>Establish channels to share ML impact transparently</p>	<p>● ●</p>	<p>Provide a mechanism for a safe feedback from the audience to which AI is delivered, and through it share publicly information on the process adopted to identify human rights risks, the risks identified, and the concrete steps taken to prevent and mitigate such risks; this could include publishing technical papers that explain the design of machine learning applications and how they work, and information in plain language aimed at communities impacted by the use of machine learning applications.</p> <p>Establish an open communication channel with a representative group of the people that this ML application could affect (could involve focus groups or other consultation processes)</p> <p>Identify ways to constrain the use of deep learning/neural networks that are inscrutable for decision-making functions that relate to people's rights</p>	<p>Businesses involved in developing or deploying machine learning systems</p> <p>Companies like Microsoft and Google are already deeply involved in partnerships to begin to understand how to best promote transparency</p> <p>Public sector entities that govern and regulate these businesses. For example: the European Union's new policies enshrining "the right to understand" in Big Data technologies</p>
<p>Measure, Evaluate, Report</p>	<p>● ●</p>	<p>Where machine learning has been used to make a decision that may directly impact the enjoyment of human rights, clearly disclose the use of AI to people impacted by such decisions and provide mechanisms for recourse</p> <p>When machine learning is used in circumstances where it interacts with the public and makes decisions that affect individuals legally or would have a significant impact on them, ensure that appropriate notices are provided (consent may be needed in some cases, in certain jurisdictions)</p> <p>As with the Responsible Research and Innovation (RRI) model, measure and report evidence of the positive social impacts that a ML system is having on society</p>	<p>Businesses involved in developing or deploying machine learning systems working together with governments and, where viable, multi-sector oversight bodies, such as the consortia already at work in the inclusive/fair AI space (AI Now, the Partnership on AI, etc.) and think tanks/NGOs focused on "AI for Good" and ethical applications for AI</p> <p>Public-sector entities involved in deploying machine learning systems</p>

Acknowledgements

Global Future Council on Human Rights 2016-2018

Dapo Akande	Professor of Public International Law	Faculty of Law, University of Oxford
Anne-Marie Allgrove	Head, Global Information Technology and Communications Industry	Baker McKenzie
Michelle Arevalo-Carpenter	Chief Executive Officer	IMPAQTO
Daniel Bross	Senior Adviser	Haas School of Business, University of California, Berkeley
Amal Clooney	Barrister	Doughty Street Chambers
Steven Crown	Vice-President and Deputy General Counsel	Microsoft Corp.
Eileen Donahoe	Executive Director	Global Digital Policy Incubator
Sherif Elsayed-Ali	Head, Technology and Human Rights	Amnesty International
Isabelle Falque-Pierrotin	President	Commission Nationale de l'information et des libertés (CNIL)
Damiano de Felice	Director, Strategy	Access to Medicine Foundation
Samuel Gregory	Programme Director	WITNESS
Erica Kochi	Co-Founder, UNICEF Innovation	United Nations Children's Fund (UNICEF)
May-Ann Lim	Executive Director	Asia Cloud Computing Association
Katherine Maher	Executive Director	Wikimedia Foundation Inc.
Marcela Manubens	Global Vice-President, Integrated Social Sustainability	Unilever
Andrew McLaughlin	Co-Founder and Partner	Higher Ground Labs
Mayur Patel	Executive Vice-President, Strategy and Corporate Development	Econet Media
Michael H. Posner	Jerome Kohlberg Professor of Ethics and Finance; Director, Center for Business and Human Rights	Stern School of Business, New York University
Esra'a Al Shafei	Founder and Executive Director	Mideast Youth
Hilary Sutcliffe	Director	SocietyInside
Manuela M. Veloso	Herbert A. Simon University Professor, School of Computer Science	Carnegie Mellon University

Global Future Council Fellow

Miles Jackson	Associate Professor of Law	University of Oxford
---------------	----------------------------	----------------------

World Economic Forum

Silvia Magnoni	Head of Civil Society Communities, Society and Innovation	World Economic Forum
Lisa Ventura	Project Specialist, Global Leadership Fellow, Society and Innovation	World Economic Forum

Key contributors

Jennie Bernstein	Urban Innovation Specialist	United Nations Children's Fund (UNICEF)
Sherif Elsayed-Ali	Head, Technology and Human Rights	Amnesty International
Erica Kochi	Co-Founder, UNICEF Innovation	United Nations Children's Fund (UNICEF)
Mayur Patel	Executive Vice-President, Strategy and Corporate Development	Econet Media



COMMITTED TO
IMPROVING THE STATE
OF THE WORLD

The World Economic Forum, committed to improving the state of the world, is the International Organization for Public-Private Cooperation.

The Forum engages the foremost political, business and other leaders of society to shape global, regional and industry agendas.

World Economic Forum
91–93 route de la Capite
CH-1223 Cologny/Geneva
Switzerland

Tel.: +41 (0) 22 869 1212
Fax: +41 (0) 22 786 2744

contact@weforum.org
www.weforum.org