Artificial Intelligence (AI) for Development Series

Module on

AI, Ethics and Society

July 2018

Work in progress, for discussion purposes
Comments are welcome!
Please send your comments on this paper at: gsr@itu.int by 30 July 2018

This module was prepared by Michael Best, Director, United Nations University Institute on Computing and Society (UNU-CS), Professor, Sam Nunn School of International Affairs and the School of Interactive Computing, Georgia Institute of Technology, under the direction of the ITU/BDT Regulatory and Market Environment Division, in collaboration with the ITU/BDT Telecommunication Division and under close coordination with the Chief of the ITU/BDT Infrastructure, Enabling Environment, and E-Applications Department. We would like to thank the ITU General Secretariat and the ITU Standardization Bureau for their contributions.

Contents

1. Introduction

Artificial Intelligence is growing exponentially in its impact on human society. While the field of scientific inquiry and technical progress is roughly seventy-years-old (if we pin its origin to the 1950 work of Turing and the 1956 Dartmouth workshop), it is only now that we see AI impacting many of our lives on a daily basis. AI appears in the foreground as we interact with some fluidity, through voice recognition and natural language processing, with digital assistants like Siri and Alexa. And AI is present for many of us in the background, for instance as we use a credit card and our bank applies an AI based fraud detection algorithm while approving payment. It is not just the frequency with which we might interact with an AI today that makes it ubiquitous, it is also its broad range of applicability: from healthcare to education to agriculture to manufacturing to transportation to leisure and more.

Thus, for many of us, AI is newly ubiquitous. For all of us, however, AI has multiple valences; it can be an instrument for social and individual advancement and pleasure, or it can be an instrument for social and individual ills and discontent. Put simply, AI is replete with vast rewards and manifest risks. For example consider this utopian/dystopian dialectic: AI will either be the source of a broadly enjoyed ethical leisure society or the cause of massive unemployment. As Yudkowsky (2008) trenchantly put it, "after the invention of AI, humans will have nothing to do, and we'll starve or watch television."

These two factors – a growing global ubiquity and an emerging set of risks and rewards – is why AI presents such a wide array of increasingly sticky ethical and societal concerns. It is why, in particular, policymakers and political institutions must vigorously join the public debate over AI systems. Ultimately, policymakers need to be willing to speak, learn and act to enhance the rewards and mitigate the risks of increasingly ever-present artificial intelligences. These two factors are what motivated more than 8000 AI researchers and developers, including the likes of Elon Musk, Stephen Hawking, and Margaret Boden (https://futureoflife.org/ai-open-letter/) to argue that "[t]here is now a broad consensus that AI research is progressing steadily, and that its impact on society is likely to increase…. Because of the great potential of AI, it is important to research how to reap its benefits while avoiding potential pitfalls."

A sign of this development of AI's impact, and its clear benefits and potential pitfalls, is the growth of the ITU's AI for Good Global Summit (https://www.itu.int/en/ITU-T/AI/Pages/201706-default.aspx), where industry, academia and government have been exploring the privacy, security and ethical questions that arise from AI.

These three factors – the growing scale and scope of AI, its bivalenced risks and rewards, and the central role for policymakers including ICT regulators – are the core foundations of the ethical and social issues overviewed in this module. *The goal of this module is to help ICT regulators and policymakers consider a few of the many core ethical and social issues that are emerging due to AI systems; these issues are developed here as a series of values and the ways that AI can positively or negatively impact these values. This module is not designed to offer policy prescriptions but instead will try to surface relevant values-based ethical and social issues. In doing so it raises some of the core ethical and social questions that policymakers and regulators*

must understand, track, and at times influence as AI continues its remarkable growth and development.

2. A Brief History of AI ethics and society

As mentioned above, the birth of modern AI is rooted in work from the 1950s, primarily undertaken in the USA and UK (see Box 2.1 for some emerging AI activities from other regions). The paper that perhaps best marks the creation of AI is Allan Turing's landmark *Computing Machinery and Intelligence* (Turing, 1950). In it he introduced what has come to be known as the Turing Test, an imitation game designed to demonstrate artificial intelligent behavior by a machine. Embedded within this parturition of AI, Turing already contemplates its ethical and social implications, arguing that an AI in principle *could* "[b]e kind, resourceful, beautiful, friendly, have initiative, have a sense of humor, tell right from wrong…" (and presumably could also be the opposite).

Attempts to encode ethical guidelines for AI even predates Turing's seminal work. Most famously we have Isaac Asimov's (1950) Three Laws, which seek to circumscribe the actions of a robot (and by extension any artificial intelligence) to ensure its social benefit and self-preservation. The Three Laws read:

1. "A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws."

In a sense, Asimov gives us the first policy designed to ensure modern artificial intelligences behave in a safe and ethical way. Today it is hard not to look back at this early AI policy with some envy: if only current realities would us to rely on such a simple, short and elegant policy.

While AI technologies have matured and developed at a remarkably rapid pace, our consideration of the ethical and social implications of AI systems have grown slowly from these storied beginnings. For instance, a decade ago Yudkowsky (2008) asked why there "aren't more AI researchers talking about safety?". In his influential essay, he argues for a "Friendly AI," an ethical and constrained artificial intelligence which benefits humanity and society. In the decade since the publication of his call for an ethical AI, many more researchers are indeed talking about the social and ethical implications of AI. Nevertheless, ethical and social thinking about AI has *not* kept pace with the rapid technological developments at hand. Moreover, the policy and regulatory community has often remained at a distance to this small but growing community of AI ethicists. We hope that this will change and that AI ethicists will be informed by and in turn will help to inform ICT policymakers. Indeed, as we will often repeat in this module: AI is moving at such a pace that it is critical for ICT policymakers and regulators to stay abreast of its growth along with any concomitant ethical and social dimensions to AI. To do otherwise could put our ICT systems (and more) at risk. Consider this module just one invitation to this conversation between ICT policymakers and regulators and AI ethicists.

BOX 2 .1: ARTIFICIAL INTELLIGENCE STARTUPS IN AFRICA AND LATIN AMERICA

Historically, artificial intelligence has been a project primarily within the USA and UK. Recently, China has made considerable strides in developing its AI capabilities (see Box 5.b). As a response to these realities, twenty-four EU ministers have signed a pledge to develop a "European approach" to artificial intelligence research, development, investment, and legal and ethical policies (Stupp, 2018). Stakeholders in Asia, Europe and North America are competing for AI dominance, but what about Latin America and Africa?

Countries in Africa are looking at artificial intelligence as a means to create new jobs and develop skills for the workforce (Novitske, 2018). Some argue that emerging states can leapfrog into the AI revolution (Samans & Zahidi, 2017). But challenges have been identified in the development of AI on the continent including: 1) weak communications and complimentary infrastructures, 2) limited government engagement, 3) a need for AI training and capacity building, 4) persistent "brain drain" among AI experts especially in Africa, and 5) limited gender diversity among the AI workforce (Brandusescu, Freuler, & Thakur, 2017).

Today, a combination of universities, domestic businesses and organizations are supporting AI in a number of African nations. Many USA headquartered tech companies are also investing and mentoring entrepreneurs through incubators, hubs and competitions like the Google Launchpad Accelerator, IBM AI Xprize and Microsoft's 4Afrika initiative. The extensive mobile infrastructure and embracing of mobile money creates ripe conditions for AI research and use. Companies in Africa are experimenting with AI to solve a variety of problems across sectors from finance, healthcare, manufacturing, agriculture and others. Here are some examples:

- In Nigeria, an AI start-up, Ubenwa, has developed a mobile app which uses a newborn's cry to detect childbirth asphyxia (http://ubenwa.com/). According to their website, Ubenwa's machine learning system takes an infant's cry as input, analyses the amplitude and frequency patterns of the cry and provides an instant assessment of whether the newborn has birth asphyxia or is at risk of it. Ubenwa has recently started conducting clinical validation exercises at hospitals in Nigeria and Canada.

- The Kenyan company, Farmdrive, is addressing the problem of financial exclusion faced by millions of family farmers across rural Africa (https://farmdrive.co.ke/). These smallholders are often excluded from the traditional banking system and face barriers to access capital. In 2014, Farmdrive was founded by two Kenyan women to leverage the power of disruptive tech to bridge the gap between small-scale farmers and financial institutions. According to their website, Farmdrive uses machine learning to construct alternative credit profiles for farmers, and decision-making tools for financial institutions that combine environmental data (weather and climate patterns, soil data), economic data (annual income, market data), agronomic and satellite data. By verifying and augmenting the self-reported financial history of the farmers with exogenous data, Farmdrive reduces risk for the banks. The company connects with its end user through a farmer-facing app that runs over SMS, allowing farmers access even should they lack data connectivity or data-enabled feature phones (LHoFT, 2017).
-

- According to Accenture, 78% of South African executives state that they need to boost their organization's competitiveness by investing in AI (Schoeman, Moore, Seedat, & Chen, 2017). One South African company, Data Prophet, strives to address this market need by providing artificial intelligence consulting services to South African, as well as international businesses in manufacturing, retail and financial sectors (https://dataprophet.com/). It's product suite, Omni, uses machine learning models to predict defects, faults or other quality control criteria in manufacturing. Convolutional neural networks and image recognition are also used to automate visual quality control inspection. According to their website, Omni further identifies and prescribes ideal process variables to optimize operating yield in manufacturing, thus improving the quality and efficiency of the business. Apart from Omni, Data Prophet also offers an AI chatbot called Mentat. The chatbot learns responses to customer service queries, escalating only unsolved queries to a human representative. The company claims to improve their clients call volume and costs by more than a fifth.

Similarly, across Latin America, many enterprises are using AI in novel ways to solve some of the toughest local and global challenges:

- *Operação Serenata De Amor* is an open-source and crowd-funded artificial intelligence project, pioneered by Brazilian data scientist Irio Musskopf (https://serenata.ai/). The project uses a combination of public data procured from government websites and agencies, and information from Google, Foursquare, Yelp and other websites, to "audit public accounts" and support citizen control of public finances and engagement with public officials. Brazilian government agency *Quota for Exercise of Parliamentary Activity* (CEAP) receives over 20,000 reimbursement claims every month from Brazilian Congress members (https://jarbas.serenata.ai/dashboard/chamber_of_deputies/reimbursement/). An AI system, named Rosie is used to analyze this enormous quantity of data and flag any irregularities or suspicious activities. The system then reports its findings to the lower house of Brazilian parliament and also flags it publicly on twitter, holding elected legislators accountable to Brazilian citizens (Matsuura, 2017). The *Serenata de Amor* team plan to gather and make public information such as the wealth of politicians, the donations received by campaigns, bills already proposed, and expenses for work and district projects, ahead of Brazil's national elections in 2018 (Monnerat, 2018). The organization further plans to develop a robot-journalist that will be able to write short articles about the bills that were flagged by Rosie.

- Based in Chile, Not Company is a food tech company that aspires to resolve world hunger, climate change and unsustainable food production through transforming the food we eat (http://www.thenotcompany.com/). The company has developed *Giuseppe*, an AI system that analyzes the molecular structure and properties of popular food ingredients and develops sustainable plant based substitute recipes for popular animal products like meat and milk (Al Jazeera, 2017). The company uses a machine learning algorithm to produce products that mimic taste, texture, nutrition and appearance of animal products, but require only a fraction of resources to produce and is more environmentally friendly (Baer, 2016).

Table 2.1

Country	Sector	Company Name	Project Description
Nigeria	Health	Ubenwa	Mobile app to detect childbirth asphyxia from a newborn's cry.
Kenya	Agriculture	Farmdrive	Profiling and decision support for credit to small-scale farmers.
South Africa	Consulting	Data Prophet	Manufacturing quality control and customer service chatbot.
Brazil	e-Government	*Operação Serenata De Amor*	Tools to enhance government accountability and transparency in their public expenditures and financial activities.
Chile	Food	Not Company	Application to develop plan based vegetarian substitutes for popular animal based recipes.

3. Is AI Different?

Above, we argue that AI is bivalenced – the technology presents both risks and rewards. But this surely is not unique to AI as, indeed, all technologies have both potential positive and negative impacts. Most insidiously, technologies also generally have unintended consequences. In this way, while designed for positive social impact they may instead unexpectedly result in negative outcomes. The ICT industry is replete with examples of this bivalenced tension. For example, we all know of the amazing social and economic benefits that have arisen from mobile telephony. But we also should reflect on the safety issues of texting or talking while driving; the criminal applications of anonymous "burner" phones, and so forth.

We have also argued that AI is expansive in its scale and scope. It can reach across most sectors and elements of society and appear in all number of possible systems. But is this unique to AI? Surely a similar argument could be made about, for instance, the Internet or mobile telephony. These technologies also have a broad range of applicability from healthcare to education to agriculture to manufacturing to transportation to leisure and more.

The mere fact that AI is bivalenced and broad is not, in and of itself, an argument that AI is socially or ethically a markedly different technology from those that have preceded it. The internet and mobile phones are bivalenced and broad as well. Nevertheless, there are some ways that AI *might,* in fact, be unique and different with potentially profound ethical and societal import.

First, we cannot lose sight that, definitionally, the goal of AI is the creation of intelligent machines. And intelligence qua intelligence is, in many ways, a different form of "technology." A set of AI thinkers have noted that the path towards intelligent machines may lead us to greater-than-human intelligence. The emergence of this sort of super artificial general intelligence could pose ethical and social challenges hitherto unknown. Indeed, some argue that the rise of a superintelligence (Bostrom, 2014) will present an explosive "existential AI threat" (Good, 1966),

or “singularity” (Vinge, 1993), beyond which it is impossible to predict the outcomes or even the continued existence of humanity. Undeniably, it is plausible reasoning that underpins this perceived threat. An AI might increase in its intelligence with vast speed due to, for instance, recursive self-improvement (Good, 1966): an AI is designed to learn efficiently, it applies this efficient learning to its own capacity to learn, which when enhanced becomes even better at learning efficiently, and so forth.

An AI explosion due to recursive self-improvement is a cautionary tale of technologies out of control. And some researchers have responded to this possible threat with a call for a moratorium on any relevant research pathways (Metzinger, Bentley, Häggström, & Brundage, 2018). Indeed, such a moratorium, and related approaches, are likely to be sensible policy responses.

Even if this feedback loop does not result in an AI explosion, the potential speed of AI change that it drives might be faster than we experience with other ICT areas. How can we be sure that ethical and social policies keep up with this pace of technological change? How can ICT policymakers position themselves to respond when and as required to AI development? A first step is for all ICT policymakers to commit to remaining knowledgeable and up-to-date with the cutting-edge state of AI ethics and society.

Beyond learning feedback loops, and the potentially unusual pace of change that results from this feedback, there are further ways that AI might be different from other current or emerging ICTs. For instance, AIs capacity to enhance or replace human agency with machine agency (and in doing so subsume moral agency directly), might be different. And any ability to outsource ethics itself to AI, with machine based ethical decision making and behavior, would be different. The ability for an AI to develop true autonomy could also be different. And so forth.

A superintelligent singularity or a moral or ethical AI are, indeed, ways that artificial intelligence might be different from other technologies today. But these are also directions that have more than a hint of a dystopian science fiction Hollywood film; they are not a likely reality. Therefore, beyond these few paragraphs, this module will not further consider these existential ethical and social implications for AI, instead focusing on more likely threats and promises. In doing so, we will be treating AI as more similar than dissimilar to other major recent technological innovations (e.g., the Internet and mobile phones).

What this overview should make clear is that we all are struggling to predict the future of AI and our gaze onto its potential impacts on humanity is hazy at best. The fact that AI is complex and its future is not entirely clear underpin the need for ICT regulators and policymakers to stay abreast of its development and conversant with any emerging social and ethical concerns. ICT policymakers will need to develop systems for multi-stakeholder consultation from which emerge dynamic, responsive, and as required, predictive and admonitory AI policies and public processes. It is only through an ongoing ethically-informed policy engagement that the best forms of AI are likely to flourish, and the most negative potential impacts of AI can be attenuated.

BOX 3.1 : THE AI & ETHICS ADMONITION:
AI is changing with enormous speed and is affecting many different elements of human society all at once. It is impossible to perfectly predict all of the many ways AI will impact the systems, infrastructures, and ethical and social areas of concern to the world's ICT policymakers and regulators. In order to respond quickly and effectively to ethical and social issues that arise out of new AIs, and in order to be proactive and prudent, ICT policymakers must remain up-to-date on AI social and ethical issues, engage in real-time and continuous multi-stakeholder and cross-institutional consultations and engagements on these issues, and maintain nimble and responsive policy mechanisms and procedures.

4. Framing AI, Ethics and Society

There are many potential ways to organize, or frame, the reciprocal role of AI on ethics and society. Different organizational structures would best reveal certain societal and ethical properties of AI and would most usefully structure a conversation on this topic.
For instance, one might taxonimize the various extant and nascent AI technologies, list for each technology the various societal and ethical considerations and contemplate the various policy positions relevant to this technology. For example, autonomous vehicles and drones could be considered as a stand-alone technology, as could natural language processing systems. While there is a natural ease to this particular framing device, there is a risk that it would become techno-centric allowing the systems to carry the conversation when it is the ethical and social issues that are most salient to policy decisions. Further, a technology focused approach might limit the policy responses, for instance requiring a specific regulation for each new technical innovation.

Another option is to organize the discussion around various economic sectors, for instance health, education, military, transportation, etc. But many social and ethical issues are cross-cutting; they impact multiple sectors at once. And, indeed, ICT policymakers are often themselves cross-cutting given the all-encompassing nature of communication and information infrastructures. Thus, sectors would seem to be an unparsimonious organizing principle.

Instead, in this module, we constitute a values-based organizing framework. One advantage of leading with values is that it naturally privileges what is most ethically and socially salient to a values-driven policymaker in order to arrive at a values-driven society. For instance, privacy is a widely regarded value, and a values framework would allow us to place privacy front-and-center in our considerations. Additionally, many values are rooted in certain universal principles including the Universal Declaration of Human Rights, adopted by the General Assembly of the United Nations and adhered to by member states. Admittedly, though, a values-based approach is not without its challenges as many values are not entirely universal in how they are defined or prioritized. For example, anti-discrimination of protected groups is a widely regarded value, though different cultures might differ in which groups they most assiduously strive to protect.

Nevertheless, we believe that a values framework offers the best chance at surfacing key ethical and social issues and so in this module we develop a bivalenced values framework for AI, ethics and society. Each *value* is presented and overviewed as it relates to artificial intelligence. For each value, we then offer examples of *rewards* – ways that AI might positively engage this value – along with *risks* – ways in which AI presents a challenge or threat to the value. We will also

explore *connections* – ways in which ICT policymakers and regulators have already considered how this value interfaces with types of information and communication systems. We also include some of the salient or most representative *questions* that each value exposes. As we repeatedly mention, AI is moving quickly and is affecting many parts of our lives. Some aspects to AI are new, and all aspects are dynamic. We are just now getting a handle on some of the most important ethical and social questions posited by these emerging AI technologies; it may be too early to have answers to many of these questions.

For this module we will examine just a few of the many critical values pertaining to artificial intelligence: 1) livelihood and work; 2) diversity, non-discrimination and freedoms from bias; 3) data privacy and minimization; 4) peace and security. We note that this list of values is quite similar to values that the IDRC have identified as potentially at risk, particularly in the Global South, due to the rise of artificial intelligence (M. L. Smith & Neupane, 2018).

4.1. Livelihood and Work

Value	Rewards	Risks	Connections
Livelihood and Work	Economic growth; new forms of work; expanded leisure	Enormous global labor disruptions; expanding unemployment and job competition	Similar claims were made about e-commerce, the internet generally, etc.

For most, work holds an intrinsic value. For all, a secure livelihood is paramount. Truly, livelihood and work encompass a set of human values which AI can and should positively support. But AI is also seen as a potential source of work and livelihood dislocations. This section will explore the positive and negative valences of AI as it relates to human livelihood and work.

Significant technological disruptions are commonplace in market economies going back to the Industrial Revolution and even before. And debates regarding the negative versus positive influence of technological change on work go back as far. The negative impact of technology on work, the displacement effect, occurs when human labor is replaced or undermined by technological systems. The positive effect of technology on work, the productivity effect, happens when demand for labor increases due to innovation and technological automation.

In general economists argue that while technological innovation often have a short-term displacement effect, in the long run, new technologies create an overall positive productivity effect on labor (Petropoulos, 2017). But, could AI be different compared to our experience with other technologies? Are we entering a new period of Artificial Unemployment (Boddington, 2017) different from what we have experienced from other technological changes? Or is AI creating new opportunities? Since the speed and scale of technological change and the scope of areas affected by AI is so vast, it may be harder to predict just how significant AI will be on labor, livelihood and work; AI may indeed be different than other technologies in this regard (Calo, 2017). As an earlier GSR discussion paper argued, “these applications are still moving from the lab to adoption, it is not feasible yet to quantify their impact at a macro-economic level,” (Katz, 2017). In this way, our admonition for ICT policymaker to be prepared, our call that they remain

up-to-date on the social and ethical issues associated with artificial intelligence, is particularly manifest when it comes to livelihood and labor issues.

4.1.1. Risks

Labor economists and related researchers have noted potentially massive workforce disruptions due to emerging AI technologies, including downward pressures on both the availability of jobs as well as on wages. According to one study, one-ninth of jobs in the USA could be affected due to self-driving transportation alone (Beede, Powers, & Ingram, 2017) and overall 47% of US workers have jobs at “high risk of potential automation” (Frey & Osborne, 2017). Are potentially half of all workers at risk of dislocation due to AI enabled automation technologies? Compared to earlier technology-driven disruptions, this would be an entirely new level of workplace dislocations.

Even if a substantial share of the US economy is impacted by AI, it is not clear if a similar scale of change will affect workplaces globally. Worldwide, McKinsey argues that “[w]hile about half of all work activities globally have the technical potential to be automated by adapting currently demonstrated technologies, the proportion of work actually displaced by 2030 will likely be lower, because of technical, economic, and social factors that affect adoption,” (Manyika et al., 2017). A recent World Bank World Development Report also warns of potentially enormous labor disruptions especially in low- and middle-income countries, but sites mitigating factors that should attenuate the impact (World Bank, 2016). The report states that “[t]wo-thirds of all jobs could be susceptible to automation in developing countries in coming decades, from a pure technological standpoint.” But it goes on to argue that “large-scale net job destruction due to automation should not be a concern for most developing countries in the short term,” due to the lower cost of labor (which creates less market pressure for labor substitution) and the slower pace of technological adoption in these economies. Nevertheless, the percentage of the economy susceptible to AI enabled automation in the Global South, even when taking wages and technology diffusion delays into account, is generally above 40% of the employment market (World Bank, 2016).

One sector of importance to many economies of the Global South, and at particular risk from automation, is the range of offshoring activities such as call centers and back-office business processing facilities (Katz, 2017). For example, as speech recognition and natural language processing systems continue to mature, their capacities may replace many corporate offshored call centers.

Studies have noted not just geographic variation on the potential impact of AI enabled automation on labor, but variation among the genders. The World Wide Web Foundation’s report on AI in low- and middle-income countries notes that automation based reduction in job opportunities will create even more pressures on women for employment as “men compete with women for fewer jobs,” (World Wide Web Foundation, 2017).

While AI may be a net benefit to elements of the economy, its negative impacts may be more widely felt. "[J]ob loss is more salient to people—especially those directly affected—than diffuse economic gains, and AI unfortunately is often framed as a threat to jobs rather than a boon to living standards," (Stone et al., 2016).

4.1.2. Rewards

It is fair to say that AI technologies will impact a large percentage of the global labor market; just when, how much, and for whom is a matter of some debate. But impacts will be felt by a lot of the economy and not in the far distant future. This impact on labor, however, need not be entirely negative; history tells us that new technologies serve to displace labor but usually (with time) in a way that is a net gain to overall employment. So short-term disruptions historically lead to productivity enhancing longer-term growth.

A common example, sited in many of the reports referenced above, is that of the bank ATM. When bank ATMs first came onto the scene, there was considerable concern as to the impact they would have on teller employment. The worry was that ATMs would automate away the bank teller job, displacing a non-trivial number of workers in many economies globally. Instead, while ATMs did disrupt the banking industry, overall they have had more positive than negative impact on employment. As an article in The Economist put it, "Replacing some bank tellers with ATMs, for example, made it cheaper to open new branches, creating many more new jobs in sales and customer service," (The Economist, 2016b).

Some authors have focused on the productivity enhancing possibilities that come when AI is partnered directly with humans around some particular work task. According to their report, "Gartner is confident about the positive effect of AI on jobs. The main contributor to the net job growth is AI augmentation — a combination of human and artificial intelligence, where both complement each other," (Gartner, Inc., 2017). Gartner is predicting that ultimately the job-creating benefits of AI will overwhelm any labor disruptions. They expect that "[i]n 2020, AI becomes a positive net job motivator, creating 2.3 million jobs while only eliminating 1.8 million jobs. In 2021, AI augmentation will generate $2.9 trillion in business value and recover 6.2 billion hours of worker productivity." Similarly, according to the panel report of the One Hundred Year Study on Artificial Intelligence, "AI will likely replace tasks rather than jobs in the near term and will also create new kinds of jobs. But the new jobs that will emerge are harder to imagine in advance than the existing jobs that will likely be lost…. It is not too soon for social debate on how the economic fruits of AI technologies should be shared," (Stone et al., 2016).

Another entirely radical viewpoint, articulated by some (perhaps on the fringe), is that AI's work supplanting capacities will be so unlimited as to render human labor itself superfluous. In this scenario, human activity will be taken over by leisure, art, interpersonal interactions, and rest.

Returning to the sober analysis of The Economist, they conclude by asking, "who is right: the pessimists (many of them techie types), who say this time is different and machines really will take all the jobs, or the optimists (mostly economists and historians), who insist that in the end technology always creates more jobs than it destroys? The truth probably lies somewhere in between" (The Economist, 2016a).

4.1.3. Connections

ICT policymakers and regulators will find some of these concerns (and promises) reminiscent of similar arguments lodged against e-commerce, the internet generally, and indeed ICTs more

generally. Market displacements, dislocations, and indeed growth are not new phenomena to ICT stakeholders. In order to manage the oncoming labor changes driven by AI, Manyika and Spence (2018) call us to attend to three priority areas:

1) skills and training,
2) labor market fluidity and dynamism (for instance ensuring that workers can easily move between jobs),
3) and income and transition support for those displaced.

Even though this list has been newly created to account for AI's emerging impacts, it should not be entirely unfamiliar to ICT policymakers who have looked to make similar accommodations for the labor impacts of previous information and communication technologies.

4.1.4. Key Questions to Consider for This Value

Are AI systems different enough from other technological change to upset the usual patterns where labor productivity effects ultimately outweigh labor displacement effects?
Will AI develop so quickly across so many economic sectors all at once that our economies will struggle to adapt and develop in time?
Can ICT infrastructure help to enhance specific labor productivity effects while mitigating displacement effects?
How will the ICT sector itself (e.g., customer support operations or network design) be assisted by AI engines?

4.2. Diversity, non-discrimination and freedoms from bias

Value	Rewards	Risks	Connections
Diversity, non-discrimination and freedoms from bias	Systems to support linguistic diversity, pre-literacy, physical disabilities, etc.	Learned bias (racial, gender, etc.); systems that unduly privilege majority populations	Universal service and common carriage; TTY and emergency response

In her seminal book, *Machines Who Think*, Pamela McCorduck (2004) maintains how, "I'd rather take my chances with an impartial computer," than cast her lot with a potentially biased human. We as a global community value diversity and eschew bias and discrimination against protected categories. Can we, however, claim that computers are indeed impartial? Are algorithms always free from discrimination and respectful of diversity? According to Kate Crawford, "[s]exism, racism and other forms of discrimination are being built into the machine-learning algorithms that underlie the technology behind many 'intelligent' systems that shape how we are categorized and advertised to," (Crawford, 2016).

One step in ensuring an AI is free of bias is for its designers and developers to be diverse and bias-free. Box 4.2 overviews the diversity challenges within the AI design and research communities. In addition, for an AI to be impartial also requires that the underlying data that

informs and trains the AI be non-discriminatory and inclusive. For example, machine learning error rates are usually inversely proportional to training data size. Therefore, a minority subpopulation within a group is at significant risk for increased error rates in an AI's decision making if its representation within the training data is small. If data is sampled evenly based on population sizes than small populations (minorities) will be weakly represented in the data and therefore subject to heightened error rates.

Put simply: the risks for bias in AI is probably greater due to the qualities of its datasets than for any "hand coded" biases of its algorithms. As the 100 year Study Panel put it, "though AI algorithms may be capable of making less biased decisions than a typical person, it remains a deep technical challenge to ensure that the data that inform AI-based decisions can be kept free from biases that could lead to discrimination based on race, sexual orientation, or other factors," (Stone et al., 2016).

BOX 4.2: WHERE ARE THE WOMEN? GENDER DISPARITIES IN AI RESEARCH AND DEVELOPMENT

The artificial intelligence community has a diversity problem. Microsoft researcher Margaret Mitchell has called AI a "sea of dudes" (Boddington, 2017). Kate Crawford, also a Microsoft researcher and NYU professor, asserts that AI has a "white guy problem" (Crawford, 2016). Crawford goes on to articulate why this matters: "Like all technologies before it, artificial intelligence will reflect the values of its creators. So inclusivity matters — from who designs it to who sits on the company boards and which ethical perspectives are included.
Otherwise, we risk constructing machine intelligence that mirrors a narrow and privileged vision of society, with its old, familiar biases and stereotypes."

The low level of female presence among AI researchers, developers and thought leaders might best epitomize this diversity challenge. Hannah Wallach, yet another Microsoft based AI researcher, has guessed that the entire field of machine learning is only 13.5% female (Weissman, 2016). To support those women who are already in the field, and increase the number of women who enter it, she co-founded the Women in Machine Learning (WiML) initiative (http://wimlworkshop.org). Since 2006, WiML has held regular events and now puts on an annual flagship workshop co-located with the NIPS conference.

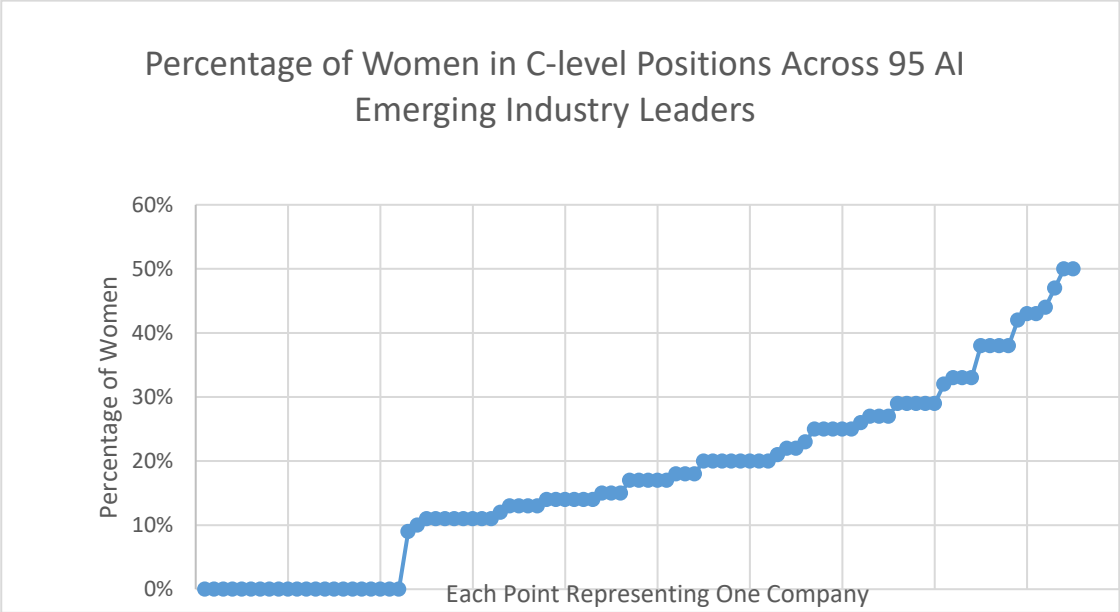
Wallach's estimate is depressingly low and underlines an enormous diversity challenge across the field. To better amass evidence as to this gender disparity, we have accumulated data on women participation in leadership among top AI companies, as well as scholarly presence among the top USA based university computer science faculty. Our new study finds that women represent a paltry 18% of C-level leaders among top AI startups across much of the globe and just 22% of faculty in top USA based university AI programs. While these percentages are slightly better than Wallach's overall industry estimate, we take no solace in them; clearly, females are overwhelmingly underrepresented among AI scholars and corporate leaders.

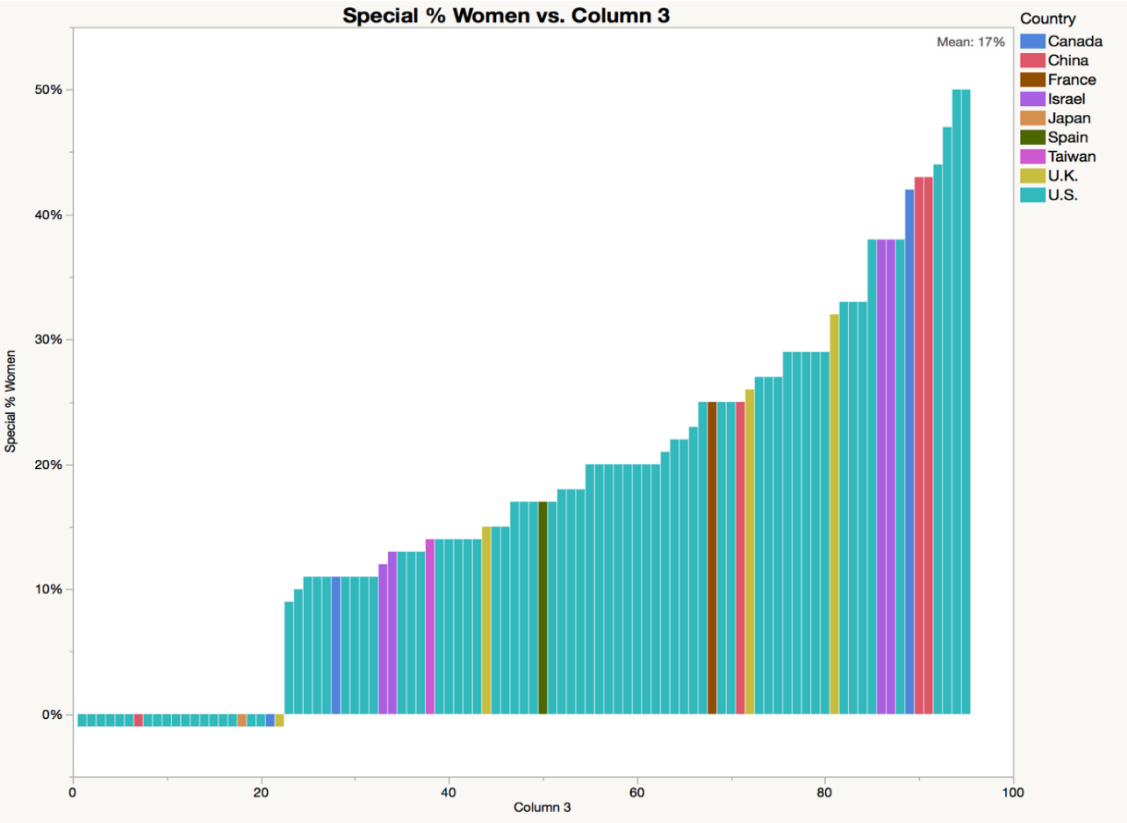
To calculate the percentage of women in executive management at leading AI startups we began with CB Insights' 2018 "AI 100", their ranking of the top 100 promising start-ups in Artificial Intelligence (https://www.cbinsights.com/research/artificial-intelligence-top-startups/). This ranking includes seed stage startups along with more advanced stage companies, and inclusion

criteria comprised factors such as investor profiles, technical innovation, business model, funding history and valuation. The executive board for each company, as listed on their website, was used to calculate the number of women in leadership positions. In case the website did not mention the executive management then searches across LinkedIn was used to establish those in leadership positions. Our calculations do not consider the business's Board of Directors, investors or advisors when gaging women in leadership positions.

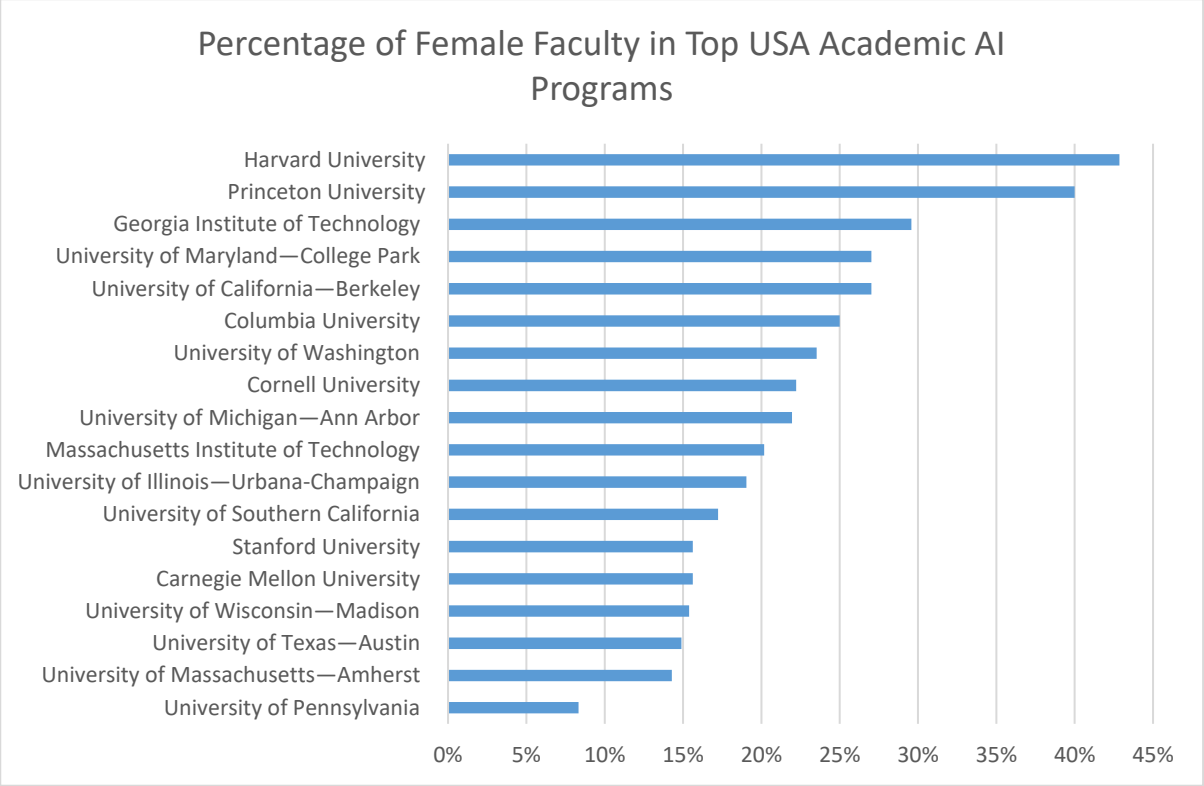
The AI 100 list includes companies from the USA, Canada, the UK, France, Spain, Japan, China, Taiwan, and Israel. We were able to establish the gender balance among executive management for all but five of these 100 companies. In only one instance was a C-level manager identified as non-binary and, for this calculation, they were not categorized.

Of the 95 companies we were able to establish data on, only two have an equal number of women to men in their C-level positions (e.g., gender parity) and none are majority female. Three in five have less than 20% women in their leadership team and one in five have no females at all. As stated above, females overall made up 18% of these AI leaders.





To compute the percentage of female professors at top US-based university AI programs we started with the US News & World Report 2018 ranking of best artificial intelligence graduate programs (https://www.usnews.com/best-graduate-schools/top-science-schools/artificial-intelligence-rankings). Using their list of the top 20, we calculated the number of faculty members (including adjuncts) from each university's website. For universities with AI Labs, we determined the faculty gender makeup directly from the lab's staff listing (e.g., Stanford University). In cases where the university did not have a separate AI Lab, the faculty's research interest, as stated on their website, was used as the inclusion criteria (e.g., Carnegie Mellon University). Some universities (e.g., Columbia University) subdivided research interest into AI, machine learning, robotics, etc. In these cases, the faculty for each related area was aggregated. We were able to obtain faculty gender information for all but two (UCLA and Cal Tech) of the top 20 programs. While the average, as noted above, was 22%, the percentage of female AI faculty ranged from a low of 8% (University of Pennsylvania) to a high of 43% (Harvard). No university had achieved gender parity among its AI faculty.



4.2.1. Rewards

First, let's consider a couple of the many ways that robust AI systems can support diversity and increase the potential for minority and exploited groups to thrive. For example, AI-based natural language translation and voice recognition systems can have a significant impact in countries with multiple languages, especially for those who communicate in a minority language which may reduce their political or economic engagement. This is the case in countries, such as Brazil or Mali, where the Portuguese or French language holds significant power. It is also true for countries with substantial linguistic diversity, such as Indonesia and Myanmar, where the majority language of Basa or Burmese holds significant power. Natural language translation systems, especially those from minority to majority languages or from local to European languages, has the promise to enhance diversity and support minority rights and inclusivity.

Of course for language translation to support linguistic diversity, natural language AI systems must become available for more languages. According to Google, their translate feature is currently available for over 100 languages (https://translate.google.com/intl/en/about/languages/index.html). If we consider Myanmar, a country of enormous linguistic diversity, their majority language of Burmese is available (though media reports have given the service mixed reviews (Weston, 2015)). However, beyond Burmese, none of the country's other major language groups are available. To belabor the obvious, natural language translation technologies can support linguistic inclusivity, but only if the language translation systems for a globally diverse set of languages are made available.

A related language technology, speech recognition, can also support diversity by providing text-to-speech readers for pre-literate populations and people with visual impairments. As the Web Foundation has put it, "These systems could also have an impact in places with high levels of illiteracy, allowing people to engage with the government or public service provision interfaces by spoken rather than by written means," (World Wide Web Foundation, 2017).

Beyond linguistic diversity, AI systems can help support other communities that are part of our diverse populations. For example, autonomous vehicles can help to enhance freedoms among people who have reduced mobility due to age or disability (Stone et al., 2016).

Policymakers and regulators have an interest in supporting diversity and ensuring inclusive access and benefit from ICTs across their populations. In this way, they should support ICT infrastructure and services that are inclusive and diversity-enhancing.

4.2.2. Risks

Bias and discrimination of AI systems against protected groups, including racial and gender groups, has received considerable recent attention. In an important piece of investigative journalism in the USA, ProPublica analyzed an AI risk assessment technology, called COMPAS, developed by the Northpointe software company (Angwin, Larson, Mattu, & Kirchner, 2016).

BOX 4.2.2: BIAS AND AI, THE CASE OF COMPAS

COMPAS is a system to assess the recidivism risk among criminal inmates in the USA being considered for parole. The COMPAS risk assessment system receives on input a set of features pertaining to an inmate housed within the US criminal justice system. This feature set includes personal details such as the inmate's education level, drug problems, age at first adjudication, number of prior arrests, etc. These inputs are then used to calculate a recidivism risk score, an attempt to predict the likelihood that the individual will reoffend in the future. An analysis from the independent news organization, ProPublica, showed that the likelihood of *false positives* – instances when the COMPAS system predicted future crime when in point of fact no such criminal act went on to occur – was higher for black individuals than for white individuals. As the ProPublica journalists described it, "The formula was particularly likely to falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants," (Angwin et al., 2016).

Note that this type of bias against black inmates was present even though race was *not* one of the features input into the system. In the USA, race is strongly correlated with other features that were input into the COMPAS technology; these surrogate features could be said to have statistically "stood in" for race in the system's analysis. Due to this surrogacy, simply excluding a protected feature from the AI system's inputs is not enough to prevent the statistical inference of this very feature. According to a Rand study, "The standard safeguard against algorithmic disparate impact effects is to hide sensitive data fields (such as gender and race) from learning algorithms. But the literature on modern reidentification techniques recognizes that learning algorithms can implicitly reconstruct sensitive fields and use these probabilistically inferred proxy variables for discriminatory classification," (Osoba, 2017).

The Northpointe corporation responded vigorously to ProPublica's reporting, highlighting independent studies arguing that COMPAS was fair in this particular way: the proportion of people classified as high-risk for recidivism who do reoffend is identical across racial groups (Corbett-Davies, Pierson, Feller, & Goel, 2016). In other words, Northpointe says their algorithm is not racially biased because their *true-positive* rates are identical across racial groups. ProPublica, in contrast, says that the COMPAS algorithm *is* racially biased because their false-positive rates are not identical across racial groups (and, specifically, blacks are more likely to be falsely classified as a future offender than whites).

According to Barocas and boyd (2017), "computer scientists and statisticians have debated the different qualities that an intuitive sense of fairness might imply: that a risk score is equally accurate in predicting the likelihood of recidivism for members of different racial groups; that members of different groups have the same chance of being wrongly predicted to recidivate; or that failure to predict recidivism happens at the same rate across groups." Scholars have noted that mathematically it is impossible to maintain all of these forms of fairness simultaneously; being fair in one way means necessarily abandoning another fairness criteria (Kleinberg, Mullainathan, & Raghavan, 2016).

The COMPAS example brings up many issues related to AI bias and fairness. How do we measure fairness and what are the trade-offs between various fairness values? What about when an AI is involved in life impacting critical-decision making areas, such as parole decisions within the criminal justice system? Does it need to be held to a higher standard? Who is accountable if the AI algorithm is free from bias but the system "learns" to be discriminatory due to its training data?

Many other examples of learned bias and discrimination in AI systems have appeared in the literature. As legal scholar Ryan Calo (2017) summarizes, "[t]he examples here include everything from a camera that cautions against taking a Taiwanese-American blogger's picture because the software believes she is blinking, to an image recognition system that characterizes an African American couple as gorillas, to a translation engine that associates the role of engineer with being male and the role of nurse with being female." These examples demonstrate the difficulties associated with bias and discrimination beyond critical-decision making machine learning systems.

4.2.3. Connections

Bias and non-discrimination are not entirely new to ICT policymakers and regulators. For instance, universal service provisions (USP) and public carrier principles are inclusive and non-discriminatory policies at their foundations. Some universal service obligations stipulate that all licensed operators must provide service of some minimal quality to all comers who meet some minimal requirements. Put simply, operators under such provisions are not allowed to discriminate against anyone making a legitimate request for their service.

Similarly, TTY and telecommunication relay services, obligated telecommunication facilities in many localities, support diverse and inclusive access to these benefits of ICT systems. These policies require that communication operators provide text-based services for users with hearing or voice disabilities.

In these ways, regulators have experience with ways to ensure that ICT providers are non-discriminatory in their practices and that they are inclusive in their services. The fact that demands on inclusivity and freedom from bias become heightened when dealing with life critical and high-risk systems also has some familiarity as it touches on some provisions of emergency response regulations (e.g. 999 or 911 services). Responding to the heightened need for care in critical decision making, Recital 71 (http://www.privacy-regulation.eu/en/r71.htm) of the European Union's General Data Protection Regulation (GDPR) states that people have the right to not be subject to decisions made by information processing systems, including AIs, that are critical in nature or have legal affect.

4.2.4. Key Questions to Consider for This Value

How can our policies best support AI systems that promote diversity and inclusivity?
How do we measure fairness and detect discriminatory action within an AI?
When an AI is involved in a life impacting critical-decision can we impose heightened requirements that it perform without discriminatory bias?
How can ICT operators use AI engines to help ensure non-discriminatory behavior and support diversity?

BOX 4.2.4 : CAN WE TRUST AI? - THE NEED FOR AI EXPLAINABILITY

"We are increasingly relying on machines that derive conclusions from models that they themselves have created, models that are often beyond human comprehension, models that 'think' about the world differently than we do," (Weinberger, 2017). With these words, technology author David Weinberger opens his essay on Alien Knowledge and the rise of AI systems whose decisions are beyond human explainability and understanding.

Consider a deep learning image analysis system designed to label pictures with terms that describe the predominant object on display (e.g., a "banana" or "car" or "tiger"). The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) is an annual competition around such a task, where competing systems machine label a publicly available image dataset (Russakovsky et al., 2015). In 2015, using a deep neural network learning algorithm, machine image labeling technology exceeded human performance in the ILSVRC competition. This technology, which performs beyond human capacities at this particular task, is an example of black-box AI; it is a form of Weinberger's Alien Knowledge. Its ability to classify images is based upon the computer system learning thousands of weights across a multi-layered neural network. The image classification that results from the application of these weights, upon the input of an image's many pixel values, admits to no human-understandable explanation. The system can correctly classify a tiger image as such, but it does so for no reason we can articulate; it is not because the object on display is "furry" and "with whiskers," for example. We cannot say exactly why the neural network weights correctly classify the image of the tiger. We can merely state that as an

outcome of processing a set of initial pre-labeled training images its many weights were set and hat it now, empirically, succeeds in labeling these cat images correctly.

Black-box (or Alien Knowledge) artificial intelligences have been the source of stunning recent successes and sit at the core of many of the most powerful AI systems in image analysis, speech recognition, natural language processing, game playing (e.g., Go, chess, or arcade games), etc. But they also are increasingly the source of consternation among some technology ethicists; in some instances, an impenetrable black-box might not be good enough and instead we may demand explainability of a system's decision. As Barocas and boyd (2017) ask us: "When is the ability to meaningfully interrogate a model sufficiently important to justify some cost in performance? What kinds of decisions—and real-world effects — drive data scientists to develop a model that they can explain, even if its decisions might be less accurate as a result?" In other writings, they answer their own questions by insisting that all critical life affecting applications demand decisions from algorithms that are fully explainable. In essence, they advocate a moratorium on black box decisions applied within these critical domain areas.

In a subsequent Wired article, Weinberger (2018) seems to consider Barocas and boyd's question on when an AI's outputs can and should be explainable, even at a cost in performance. He argues that, given the potential positive social benefits of powerful AI, explainability is an unnecessarily excessive goal if it comes at the cost of efficiency or effectiveness. "Demanding explicability sounds fine, but achieving it may require making artificial intelligence artificially stupid. And given the promise of the type of AI called machine learning, a dumbing-down of this technology could mean failing to diagnose diseases, overlooking significant causes of climate change, or making our educational system excessively one-size-fits-all. Fully tapping the power of machine learning may well mean relying on results that are impossible to explain to the human mind." Instead, for Weinberger it is enough for an AI to "meet its marks"; in other words, to empirically perform up to requirements and expectations (for accuracy, safety, etc.) independent of its knowability.

Other technologists have sought a middle ground: systems that offer the benefits of explainability while avoiding any algorithmic "dumbing-down" to achieve it. For instance, our colleagues at Harvard's Berkman Klein Center for Internet and Society (Doshi-Velez et al., 2017) call for what some others have referred to as algorithmic auditing. An audit is a method to systematically probe a black-box system with inputs designed to, in essence, reverse engineer an algorithmic explanation for its output choices. For instance, consider this example: "[S]uppose that the legal question is whether race played an inappropriate role in a loan decision. One might then probe the AI system with variations of the original inputs changing only the race. If the outcomes were different, then one might reasonably argue that race played a role in the decision," (Doshi-Velez et al., 2017).

TABLE 4.2.4: How to Trust AI: A Taxonomy of Know ability

Solution Meet its mark	Example Proponent	Reward	Risk
Ensure that AIs meet expectations on performance, safety, etc.	Weinberger, 2018	Does not risk pessimizing efficiency or effectiveness; embraces unknowable "Alien knowledge"	May mean some critical-decisions are made which cannot be explained; cannot ensure that decisions did not turn upon inappropriate bias
Auditability. Audit algorithms through varying input features	Doshi-Velez et al., 2017	Could get the best of both worlds: explainability with the power of deep learning black-boxes	Probably will not work in many cases; may be better at detecting specific cases of bias versus ruling out all potential bias
Explainability. Create new algorithms that are optimal and explainable	Angelino et al., 2017	Explainability without any of the performance detriments	Probably will not work in many cases; can cost in terms of efficiency and effectiveness
Red-line high-risk critical domain areas and demand explainability	Barocas & boyd, 2017	Ensures all critical-decisions arise from explainable systems	Excludes the potential benefits of AI from red-lined critical-decision domains

One concern with this input-varying approach to explainability is that it relies on an ability to vary eatures which may not be directly available given the very high-dimensioned inputs applied in deep learning applications. Indeed, the raw data input into these deep learning systems may be completely free of human-discernable features (such as race) and instead composed of, for instance, just a long series of pixel values. In these cases, the deep learning architecture relies on representation-learning methods that compose multiple levels of increasingly abstract data representations none of which, from input to output, are human discernable (LeCun, Bengio, & Hinton, 2015). Furthermore, consider the cases when the potential features of concern are not pre-known to those who wish to audit the algorithm, or when multiple features influence the

output through unexpected combinations such that single feature audits will never quiet reveal the full decision-making explanation.

Alternatively, some technologists have tried to find AI solutions that perform as well as deep learning neural networks while also producing outputs that are transparent and easily understood by humans. Angelino and co-authors (2017) have developed a decision list predictive model and applied it to the same recidivism datasets used by the COMPAS tool introduced in the ProPublica analysis described in Box 4.2.2. They propose an algorithmic approach that can offer a rule-based human explanation of its decision processes along with a "certificate of optimality," proving that the algorithm is as accurate and efficient as alternative black-box solutions. While this type of approach may not succeed in all or even most AI application areas, it could be explicitly demanded within domains of high-risk and critical decision making; domains where, perhaps, no Aliens need apply.

4.3. Data Privacy and Minimization

Value	Rewards	Risks	Connections
Data privacy, protection and minimization	Privacy preserving decentralized systems; privacy expert systems; new privacy protecting policies	User profile data breaches; de-anonymization and privacy concerns that arise from basic digital records.	Data privacy concerns with mobile operator user data.

Today's newspapers are replete with stories of personal data loss at the hands of large online data brokers and social media platforms. These stories surface many issues of data privacy, protection and minimization (the principle that data acquired should be limited to just what is necessary for the particular purpose at hand and shall not grow beyond what is necessary). As discussed above and in previous modules, many of today's most significant AI systems are made possible through the acquisition and analysis of large corpora of (potentially personal) data. And the range and depth of personal data acquired by AI systems are on the rise. For example, voice-driven conversational assistants, such as Alexa (Echo), Siri, and Cortana, may be more likely to know detailed private information such as what you are eating. What is clear is that users, and policymakers, are increasingly sensitive to privacy issues that arise from artificial intelligences. Indeed in a recent survey, "[a]lmost three-quarters (71%) [of respondents] say they don't want companies to use AI that threatens to infringe on their privacy, even if it improves the customer experience," (genpact, 2017).

4.3.1. Risks

Many recent headline-grabbing incidents illustrate some of the challenges related to AI systems which acquire, store and analyze large amounts of personal user data. First are risks associated with the very business models underlying some online platforms, and how these business

practices incentivize the acquisition of highly personal user data and make possible user data releases, both accidental and purposeful. A second goes to the kind of personal information that can be inferred indirectly from data released by users.

The release of millions of social media users' profile information to a data science and machine learning corporation, Cambridge Analytica, has been widely reported across major news media. (The Guardian has been at the forefront of some of this reporting, see https://www.theguardian.com/news/series/cambridge-analytica-files.) This user profile data was used by clients of Cambridge Analytica, including high-profile political campaigns within the USA, to micro-target persuasive advertising to specific and precisely modeled user populations. In this specific incident, a reported breach of user data along with sophisticated machine learning approaches came together to produce powerful targeted communications that may have had considerable political impact. This raises ethical and societal concerns for AI. First are privacy concerns that arise from the user profile data release, generally without the knowledge of the users involved. Second is the role of machine learning in influencing a nation's political processes (a topic we will not further address in this module).

Jose Marichal (2012) argues how the acquisition and disclosure of private user data, made famous in the Cambridge Analytica story, is not a misfeature of social media platforms but instead is central to their business model. He simplifies these platforms to a system for connections and disclosures. *Connections* are mostly realized through social media's facility to network friends. These friends are self-selected, and research has shown mainly consist of intimate, strong-tie relations existing offline as well as online. Communication exists within this network of close connections, and it is these user communications that Marichal refers to as *disclosures*. The architecture of disclosure is the platform's purpose-built environment to systematically and, in some ways, insidiously encourage its users to not simply disclose but increasingly to disclose personal revelatory data. To Marichal, social media has become the "perfect machine to get you to reveal intimate (if sometimes banal) details about yourself to others," and in so doing, reveal these very details to the platform and ultimately their clients and advertisers.

Social media systems have perfected this architecture not with degraded voyeuristic interest; it is simply their business model. They capture and commodify a portfolio of these disclosures, often through profile modeling, and sell this on to their advertisers. They have no prurient interest in your personal data, but it is the acquisition and analysis of this unique and highly personal dataset that has allowed social media systems to become the world's largest micro-targeted advertising platforms.

The Cambridge Analytica incident perfect illustrates the clash between this business model, predicated on acquisition of increasingly personal user data, and its associated privacy risks. Increasingly, policymakers and technology leaders are arguing that this incident underlines the need for ethical AI privacy standards and regulations (Hern, 2018). We are now constantly reminded that we have "entrusted the most intimate parts of [our] digital life to a profit-maximizing surveillance machine," (Roose, 2018).

While the Cambridge Analytica story is predicated on the release of intimate *private information*, up to and including even private instant messages, it has also been shown that private information can be discerned even from basic *public information* such as "Likes" (Kosinski,

Stillwell, & Graepel, 2013). In the Cambridge Analytica case, intimate private user details were released to a third party (for the purpose of political persuasion). But researchers have now concluded that even basic seemingly non-private information, some of which may be by default on offer to the public at large, can be used to infer our most intimate of private details through machine learning techniques. Therefore, a data breach of private profile data may not even be required for an entity to obtain personally revelatory information about social media users.

In their study, Kosinski and co-authors (2013) determined that "Facebook Likes, can be used to automatically and accurately predict a range of highly sensitive personal attributes including: sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender…." This information can be gleaned through machine learning techniques even though users never imagined they were releasing such information and certainly never consented to such a release. In the hands of an advertiser, one can imagine this information being used to target promotions that may be unwanted or even harmful when received by the platform user. And if used by other actors, information about sexual orientation or political views might pose an even more dire and direct threat to an individual.

4.3.2. Rewards

Sophisticated machine learning techniques along with social media platforms' architectures of disclosure working in concert have created a substantial challenge to our individual privacy. Happily, AI technologies are also attempting to enhance user privacy that has increasingly been put at risk. For instance, Thomson Reuters has teamed up with IBM's Watson division to develop the Data Privacy Advisor, an expert system able to provide specialized advice to privacy professionals on their obligations across multiple jurisdictions (B. Smith & Al-Kohafi, 2018). The system works through an IBM Watson conversational interface allowing users to pose questions through natural language queries.

A few additional privacy-reclaiming approaches have been suggested (and to some extent are underway) by policy and technology leaders. This includes a set of responses around the putative monopoly control that some of the largest social media platforms may enjoy. Breaking up these large companies, and the enhanced competition that this might introduce, may have privacy enhancing effect as customers select for platforms which meet their privacy needs. Second, technologists have increasingly been developing and piloting decentralized social media platforms. These initiatives include platforms such as Mastodon (https://joinmastodon.org/) which bills itself as "the world's largest free, open-source, decentralized microblogging network." While there are some doubts as to whether these new systems can take hold, grow their user base, avoid being gamed by bad actors, and generally succeed (Barabas, Narula, & Zuckerman, 2017), nonetheless, among some there remains hope in platform decentralization.

Finally, there has increasingly been a call for increased data privacy regulations, and the European GDPR offers a glimpse at this regulatory approach. The GDPR directs that when companies collect user data, they must:

- tell them what they are using it for,

- minimize the amount of data they collect and keep to that needed just for their expressly articulated purpose
- tell people just what data they have on them,
- allows users to correct or have removed any of their data held by the company,
- and explain the logic they have used for any decision made based on the user data (Meyer, 2018).

These provisions have privacy preservation and data minimization (and explainability) at their core. As the GDPR roles out it is likely to drive changes to machine learning systems and their use of large user profile datasets. It is worth considering if the GDPR feature set, for instance, could limit future Cambridge Analytica type events.

4.3.3. Connections

Privacy concerns, made manifest when AI is applied to social media profile data, has a lot in common with the privacy issues that impact telecommunications operators. While at first blush it may seem that mobile operators hold relatively basic digital records of their users, such as cell tower derived user position data, a relatively small amount of mobile location data can be used to uniquely identify individuals. In this way, even anonymized data can be relatively easily de-anonymized. Researchers have shown that "the uniqueness of human mobility traces is high, thereby emphasizing the importance of the idiosyncrasy of human movements for individual privacy. Indeed, this uniqueness means that little outside information is needed to re-identify the trace of a targeted individual even in a sparse, large-scale, and coarse mobility dataset," (de Montjoye, Hidalgo, Verleysen, & Blondel, 2013). In turn, this locational data can be used to infer private details of the individual (Blumberg & Eckersley, 2009).

Thus the data privacy concerns that arise out of AI systems (including those based on social media profile data analysis) has significant similarities to the data privacy concerns already present with telecommunication user data (including mobile location data). Both data sets when subject to powerful analysis can turn even the public and seemingly most benign information into deeply personal details. This challenge will grow even for operator held user data sets as the capabilities of AI-driven analytics expands. AI engines, applied to an operator's user data, may result in intimate private user information almost at the scale held by social media platforms.

4.3.4. Key Questions to Consider for This Value

Can platform decentralization and private sector competition solve many of our data privacy woes?
How will policy responses, including the GDPR, support user privacy requirements?
Can privacy processes already in place for operator data assist us as AI grows in its analytic capabilities and data volume, sources, and services expand?
What are the new privacy risks and data protection imperatives for ICT operators when applying AI analytics to their large user datasets?

4.4. Peace and Physical Security

Value	Rewards	Risks	Connections
Peace and physical security	Systems for inclusivity and trust-building; peacekeeping situational awareness	Lethal Autonomous Weapons Systems	Special policy needs and realities in conflict stressed environments

Silicon Valley companies are increasingly experiencing internal debates as companies develop programs relevant to peace and warfare sectors. For instance at Google thousands of employees wrote a letter to their CEO calling for a company moratorium on "warfare technology" ," (Blumberg & Eckersley, 2009; Shane & Wakabayashi, 2018). These employees letter is indicative of a growing community of AI stakeholders calling on a moratorium on AI enabled warfare.

Alternatively, proponents have noted the promise AI holds for peace-preserving influences (for instance in fair resource distribution and climate mitigation) and conflict and crises response (Best, 2013). As one writer has put it, "AI holds much promise to enable the international community, governments and civil society to predict and prevent human insecurity. With increased connectivity, more sophisticated sensor data and better algorithms, AI applications may prove beneficial in securing basic needs and alleviating or stopping violent action," (Roff, 2017).

4.4.1. Risks

Even before Google employees sent a letter to their CEO, a large number of AI researchers (along with many thousand other endorsers) signed a letter calling for a moratorium on the development of AI-driven autonomous weaponry (https://futureoflife.org/open-letter-autonomous-weapons/). This letter concludes with the statement that, "[s]tarting a military AI arms race is a bad idea, and should be prevented by a ban on offensive autonomous weapons beyond meaningful human control." As AI's relevance to warfare continues to grow, technology stakeholders are increasingly expressing concern especially around the development of Lethal Autonomous Weapons Systems (LAWS). LAWS are systems which, according to Regina Surber (2018), "once activated, would, with the help of sensors and computationally intense algorithms, identify, search, select, and attack targets without further human intervention."

Apprehension arises around the reduction of human control, which may make warfare seem lower risk or "easier," position AIs in situations that rely on human morality and judgment, and confuse issues of accountability. Some have called to moderate the autonomy of these AI systems, requiring instead "meaningful human control" or MHC. "At its most basic level, the requirement for MHC develops from two premises: 1. That a machine applying force and operating without any human control whatsoever is broadly considered unacceptable. 2. That a human simply pressing a 'fire' button in response to indications from a computer, without cognitive clarity or awareness, is not sufficient to be considered 'human control' in a substantive sense," (Roff & Moyes, 2016). Among many scholars, a consensus may hold that people must always be meaningfully in the loop over kill decisions (Calo 2017).

The UN Convention on Certain Conventional Weapons (UN CCW), also known as the Inhumane Weapons Convention, has taken up the issue of AI-enabled autonomous weaponry. In 2014 UN

CCW convened its first informal Meeting of Experts; more recently they have stood up a Group of Governmental Experts (GGE) on LAWS. The GGE has continued to meet and explore the legality of LAWS under international law, methods for assigning responsibility and deciding questions of accountability with these autonomous systems, and consideration of the various international normative principles challenged by these technologies (Surber, 2017). This UN work is in its early stages, and much is needed to move the group towards an agreed to political declaration or international treaty.

However, others have argued that there is no practical method to restrict the development of autonomous AI enabled weaponry. Cummings (2017) argues that "[b]anning an autonomous technology for military use may not be practical given that derivative or superior technologies could well be available in the commercial sector." If the commercial sector is already undertaking an "arms race" to be the first to develop robust autonomous systems (e.g., driverless cars) then it might be practically impossible to restrict these systems from being applied in warfare settings.

4.4.2. Rewards

Can AI serve as a tool to reduce conflict and wage peace? Or will emerging AI's have unintended consequences that exacerbate war or, when placed in the wrong hands, actively erode peace?

The AI and Peace Consortium, currently incubated between Georgia Tech and Harvard's Berkman Klein Center, aims to explore the relationship of AI to peacemaking and peacekeeping through policy, social scientific and computational means. Collaborators will pursue novel research studies and interventions and convene stakeholders, scholars and decision-makers at workshops. Indeed the AI systems for inclusivity mentioned above are examples of the potential for AI to help with conflict mitigation, trust building, and post-conflict reconciliation (Best, Long, Etherton, & Smyth, 2011).

The UN Peacekeeping communities have also looked to AI to assist them in their mission to promote and establish peace in conflict stressed areas. Many have noted how peacekeeping suffers from insufficient access to and ineffective use of digital technologies (e.g., see Stauffacher, Weekes, Gasser, Maclay, & Best, 2011). Since the Brahimi Report (2000), which argues peacekeeping must be brought into the information age, operations have used ICTs but struggled to capture their full capabilities or keep pace with their rapid change. An ongoing program of relevant UN departments is exploring a broad and far-reaching platform for peacekeeping situational awareness. This platform relies on AI capabilities in data capture and validation; tracking, sensors and data integration; analysis; and visualization. Indeed, AIs could be responsive to three traditional security and peacekeeping challenges: "the inability to know about threats in advance; the inability to plan appropriate courses of action to meet these threats; and, the lack of capacity to empower stakeholders to effectively respond," (Roff, 2017). Indeed, Roff argues that artificial intelligences could automate most of the tasks associated with peacekeeping logistical support, supply chain management, forecasting and planning, and so forth.

4.4.3. Connections

In other venues we have examined the telecommunications policy process in conflict stressed environments (Best, 2011; Best & Thakur, 2009). We find that while there are many similarities in ICT policymaking between conflict stressed environments as compared with other locations, there are also differences. In particular, in conflict and immediate post-conflict states, policymaking has to contend with a weak and nascent institutional environment, intra-governmental competition, limited human and technical resources, the contested role of international actors such as the World Bank, and the dominance of elite groups in decision-making. While some of these factors are not unknown to many countries, especially in the Global South, they can be particularly germane in conflict-stressed environments.

The social and ethical concerns that arise out of AI systems applied in peace and security areas are likely to test many parts of a policymaking process. The distinctive risks associated, for instance, with LAWS means that policy and regulatory responses probably cannot just wait to "see what happens" but instead will need to be proactive and respond early. If a policy response is to be successful, traditional ICT infrastructures will almost certainly have to play a direct role.

4.4.4. Key Questions to Consider for This Value

Should AI, at least in the form of lethal autonomous weapons, be banned or tightly regulated?
How can we encourage the development of peace-enhancing AI systems?
Can traditional ICT infrastructures be brought to bear towards both of the above questions – somehow encouraging the best forms of peaceable AI while excluding the worse forms of AI supported warfare?
What are the special concerns for ICT policymakers operating within conflict stressed regions?

5. Conclusions

In this module we have proposed a bivalenced values framework for AI and have explored the risks and rewards associated with AI systems for four core example values: livelihood and work; diversity and non-discrimination; privacy and data minimization; and peace and security. Of course, this is just a starting set of values of importance to and impacted by AI. There are many other values salient to AI which should be considered when formulating policy and regulatory responses to emerging systems and that regulators and policymakers should keep in mind when examining the AI sector. The sections above offer just a handful of the many salient values that we hold, and which are impacted by and should drive ethical decisions around artificial intelligence technologies. Future work in AI, ethics and society might undertake similar analysis but for other values such as: 1) Economic freedoms and wealth; 2) democratic rights and civic engagement; 3) food security and healthy living; 4) leisure and entertainment 5) climate resilience; and 6) literacy and education.

In this module, we have also reviewed ways in which AI systems connect to systems and infrastructures well known to ICT policymakers and regulators. While there are ways that AI feels new – and without question it is particularly broad in scale and scope and is advancing with

unusual (nearly unprecedented) speed – it nonetheless shares plenty of features we encountered with the growth of mobile telephony or the internet. As we reflect on these connections between AI's social and ethical import and related values we encountered with other ICT infrastructures, we find a plethora of ways that AI impacts on areas already mandated to ICT policymakers:

- The ICT sector as a target or beneficiary of AI. For example, customer data retained by mobile and internet service providers can be subject to powerful de-anonymizing AI analysis increasing the import of data security and privacy among operators.
- The ICT sector as a tool for supporting the best forms of AI and responding to the worst. For example, operators may be best able to assist other stakeholders in identifying and responding to potentially harmful AIs released onto their networks.
- The ICT sector as a set of businesses directly employing AI, potentially in ways that have policy and regulatory relevance. For example, consider how much of operator customer support may move away from human agents (including offshored call offices) to AI chatbots.

While these are examples of ways AI is related to existing core ICT regulatory and policy areas, it is likely that in many locations ICT policy stakeholders will be asked to take on even more direct consideration of emerging AI issues. In order to be respond to existing mandated areas, and be ready for increasing and new considerations, ICT policymakers must remain informed, nimble, and conversant around the various social and ethical aspects of artificial intelligence. To do so, they must engage in real-time learning and consultation among multi-stakeholder cross-institutional coalitions.

This is already happening across a number of jurisdictions, among multi-lateral and professional societies, and within various companies. Box 5 overviews just some of the emerging policy reports and acts that are emerging.

With these connections in mind, and putting it simply, there are many many ways that artificial intelligence is already touching on areas of concern to ICT policymakers and regulators, and these associations are likely to grow, not diminish, with time. If useful, apply this module's bivalenced values framework to interrogate some of the many ways that AI is or may soon impact human ethics and society in areas relevant to ICT policymakers and regulators. It is these realities that drives us to our AI and Ethics Admonition. If they are not already, it is critically important that ICT policymakers and regulators engage with the many areas of ethical and social concern that AI touches.

BOX 5: A SAMPLE OF AI ETHICAL AND SOCIAL POLICY INITIATIVES

In the last few years a variety of initiatives have been launched to explore the ethical and social implications of AI, and to formulate policy responses to them. These initiatives can be loosely classified into three categories: governmental initiatives; initiatives started by tech companies; and those started by non-governmental organizations, academia and professional associations.

Example Governmental Initiatives

The US Government's *Preparing for the Future of AI Report* recommends re-evaluation of existing regulation with an eye towards adopting it to AI, as well as striking a balance between boosting

innovation, the costs and benefits to regulation and compliance, and the needs of public safety and fairness (National Science and Technology Council, 2016). The report also acknowledges that AI systems will need to transition cautiously from laboratories to real-life human environments, in order to avoid unsafe and unforeseen situations. Recently a FUTURE of Artificial Intelligence Bill was also introduced in US Congress that recommended forming a *Federal Advisory Committee On Development And Implementation Of Artificial Intelligence* (FUTURE of Artificial Intelligence Act, 2017). The bill sought to identify and eliminate possible bias in selection and processing of data used by AI algorithms, enhance the diversity in algorithm development, and identify applications of technology that could possibly have adverse consequences. It also explores how AI innovation will affect the privacy of individuals, create socio-economic changes, and how the government can best adopt AI to improve its own efficiency.

The European Union recently published its *Statement on AI, Robotics and 'Autonomous' Systems* that proposes nine fundamental principles for governing AI and further calls for creating an internationally recognized, common, ethical legal framework (European Group on Ethics in Science and New Technologies, 2018). It warns that the absence of a common AI regulation framework can result in "ethics shopping", and the relocation of AI development to regions with lower ethical standards. In April 2018 EU member states signed the *Declaration of Cooperation on Artificial Intelligence* and agreed on creating an ethical and legal framework based on fundamental rights and values enshrined in the EU charter including "privacy and protection of personal data" (European Council, 2018).

United Kingdom's House of Lords published a comprehensive report *AI in UK: Ready, Willing and Able?* (House of Lords' Select Committee on Artificial Intelligence, 2018), that discussed supporting and strengthening the AI industry through policy and education, as well as managing the loss of employment due to automation. It highlights the need for better legal and technical mechanisms allowing users to tailor control of their personal data while protecting privacy, instead of total data openness or total data privacy. The report endorses establishing data-trusts for the ethical sharing of data between organizations, and to counter data monopolization by a few technology companies globally. It also recomends developing a cross-sector ethical code of conduct, or "AI code" across private and public sectors, that includes creating an ethical board across companies pursuing AI development or use. The report argues that a blanket AI regulation will not be appropriate, and instead advocates for a more sector-wise regulation approach. Finally, the report also calls on research councils to request from university applicants demonstrations of implications of their AI research, and its potential misuse, along with measures taken to prevent misuse.

In 2017, China published its *New Generation AI Development Plan* (State Council, 2017), calling on financial and state resources to develop AI ensuring a "first mover advantage". The Plan also discusses challenges that AI will bring to Chinese society in employment, social stability, security risks and changing norms in international relations. It discusses developing a multi-level ethical framework for governing human-machine collaboration and deepening international cooperation to create artificial intelligence laws and regulations. The Development Plan recommends strengthening AI risk assessments with a long-term focus and establishing security monitoring & early warning mechanisms for AI.

Example Corporate Initiatives

Over the last years many Tech majors have developed principles and policies for AI research and development. For example, Microsoft states its four principles as: Fairness, Accountability, Transparency and Ethics (https://www.microsoft.com/en-us/ai/our-approach-to-ai). It speaks of honoring societal values and diversity of experience, though does elaborate on how it would implement these principles.

IBM has also articulated and AI position, declaring, for instance, that it will only develop artificial intelligence systems that augment human ability and not have any independent agency (https://www.ibm.com/blogs/think/2017/01/ibm-cognitive-principles/). It has also announced that IBM will be transparent about when and for what purposes it uses AI. IBM has proposed algorithmic responsibility and explanation based systems as a trust-building means to address bias in computer decision making (Banavar, 2016). They have also established an internal IBM Cognitive Ethics Board to advise and guide AI development and deployment.

In 2017 Google established the *DeepMind Ethics & Society* program, which has identified six key ethical challenges: Privacy, Transparency and Fairness; Economic Impact; Governance and Accountability; Managing AI Risk; AI Morality and Values; and Global Challenges (https://deepmind.com/applied/deepmind-ethics-society/research/). Their Responsible Development of AI document (https://www.blog.google/topics/ai/ai-principles/) recommends development of AI systems whose benefits will outweigh their risks. It states that Google will not pursue AI technology that supports surveillance, contravenes international law and human rights, or that can be used as a weapon. It also states that it will incorporate privacy safeguards and offer control over use of data and that its AI systems will affirm accountability by providing "appropriate opportunities for feedback, relevant explanations and appeal". Google has further stated that it will develop AI in accordance to the prevalent best practices and monitor all AI technologies post their deployment.

Technology companies have not only articulated individual AI principles, they have also been collaborating towards shared AI social and ethical policies. For example, the *Partnership on AI* is an initiative founded by Facebook, Amazon, Apple, Google, DeepMind, IBM and Microsoft (https://www.partnershiponai.org). The organization aims to advance public understanding and provide an inclusive platform for discussion and engagement with key stakeholders. It opposes "development and use of AI technology that would violate international conventions or human rights".

Other Example Initiatives

Among professional associations, the IEEE has convened a diverse set of experts to publish *Ethically Aligned Design* (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2017). The report explores the application of ethical principles to AI, as well and includes legal frameworks for accountability. It also explores "embedding values into autonomous systems"; and concludes that such a system will need to be trained in values specific to the community it is deployed in, as well as in norms relevant to the operation it is designed for. The ACM US Public Policy Council (2017), meanwhile, has released a *Statement on Algorithmic Transparency and Accountability* that advises transparency of data used to train AI systems, as well as explainability of their decisions. The statement also suggests auditing systems

in case of harm, redressing groups adversely affected by algorithms, and holding the algorithm producing company accountable.

The *Future of Life Institute* has been at the forefront of exploring ethical challenges posed by AI. Asilomar Principles, published by the Institute, have been endorsed by thousands of AI researchers and reject the creation of an AI with undirected intelligence (https://futureoflife.org/ai-principles/). The principles advocates for people's right to access, manage and control their data and against any use of private data that would curtail real or perceived liberties. The principles also insist that if am AI system could cause harm then the cause must be identifiable. They address society at large calling for sharing of economic prosperity created through AI innovation, and strengthening social and civic processes using AI systems, rather than subverting them.

TABLE 5

			Type				Values Considered				
Sector	Country	Title	Policy	Act	Principles	Livelihood Jobs	Diversity Anti-Bias	Accountability	Privacy	Peace	Security
Government Initiatives	USA	Preparing for the Future of AI	X			X	X	X	X	X	X
	USA	FUTURE of AI Act		X		X	X	X	X	X	
	EU	Statement on AI, Robotics and Autonomous Systems	X			X	X	X	X	X	X
	EU	Declaration of Cooperation on Artificial Intelligence			X	X		X	X	X	
	UK	AI in UK	X			X	X	X	X	X	X

	China	New Generation AI Development Plan	X			X		X		X	X
	Microsoft	Our Approach to AI			X		X	X		X	
	IBM	Transparency and Trust in the Cognitive Era	X			X		X	X		
Corporate Initiatives	Google	Responsible AI Practices			X	X	X	X	X	X	X
	Facebook, Amazon, Apple, Google IBM, Microsoft	Partnership on AI			X	X	X	X	X	X	X
Other Examples	IEEE	Ethically Aligned Design Version 2	X			X	X	X	X	X	X
	ACM	Statement on Algorithmic Transparency and Accountability	X				X	X	X		
	Future of Life Institute	Asilomar AI Principles			X	X	X	X	X	X	X
	AINow	AINow Report 2017	X			X	X	X	X	X	

BOX 5. B : CHINA'S GREAT LEAP INTO AI

Realizing the potential of AI technology, the Chinese government has placed AI-related technologies as one of the strategic priorities for the next decade. President Xi Jinping, in his Report at the 19th Chinese Communist Party National Congress, declared that China is to become a "science and technology superpower"; four months before in July 2017, the Chinese State Council published the Next Generation Strategic Plan for AI technologies, in which it

specifically vocalized China's goal "to achieve global leadership in AI theories, technologies, and general application, as well as becoming a major AI innovation center worldwide" by 2030.[1] [2]

China is looking at AI as an enabler of the "Chinese Dream of the Great Rejuvenation of the Chinese People" and a crucial part to building "an innovative country.[3] Specifically, the State Council recognizes the profound impact AI technology can make in international geopolitics, economic prosperity, and societal development. For the Communist Party leadership, China has to undertake a national strategic initiative if it is to compete among the top international AI actors.[4]

Politically, the Chinese government considers leadership in AI technology a tool to "improve national competitiveness," especially as China "currently faces a complicated scene of national security and national competition."[5] In his 2017 and 2018 Annual Government Work Reports, Premier Li Keqiang enunciated China's plan to "secure core technology, develop top talent, and enforce high standards" in the near future.[6] Specific to the promotion of AI-related policymaking, the State Council plans to carry out research on the legal issues of AI, including its impacts on civil and criminal liability, privacy and intellectual property protection, safe use of information, and system accountability and transparency.[7]

At the societal level, the Chinese government believes the application of AI technology is a "new opportunity for societal construction."[8] The government considers AI technologies to be instrumental in the current "fully developing a moderately prosperous society."[9] The State Council plans to utilize AI technology to advance various societal issues ranging from education, medical care, environmental protection, urban management, legal counsel, and providing care

[1] 习近平在中国共产党第十 九次全国代表大会上的报告 (Xi Jinping's Report at the 19th Chinese Communist Party National Congress)
http://cpc.people.com.cn/n1/2017/1028/c64094-29613660-7.html

[2] 国务院印发《新一代人工智能发展规划》 (The State Council Issues "Next Generation Artificial Intelligence Development Strategic Plan")
http://www.gov.cn/xinwen/2017-07/20/content_5212064.htm

[3] 国务院关于印发新一代人工智能发展规划的通知 (Notice of The State Council's Issuance of Next Generation Artificial Intelligence Development Strategic Plan)
http://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm

[4] Ibid.

[5] Ibid.

[6] 2018 李克强总理政府工作报告 (Premier Li Keqiang's Annual Government Report in 2018)
http://www.xinhuanet.com/politics/2018lh/2018-03/22/c_1122575588.htm

[7] 国务院关于印发新一代人工智能发展规划的通知(Notice of The State Council's Issuance of Next Generation Artificial Intelligence Development Strategic Plan)
http://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm

[8] 国务院关于印发新一代人工智能发展规划的通知(Notice of The State Council's Issuance of Next Generation Artificial Intelligence Development Strategic Plan)
http://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm

[9] Ibid.

for the elderly.[10] The government hopes to leverage the perception, prediction, warning, and analyzing capabilities of AI systems based on big data to take an active role in policy-making and improve its ability of social management and stability maintenance.[11]

In the private sector, the three largest Chinese technology companies, Baidu, Alibaba, and Tencent are all making strong investments into AI research and development. Baidu, the Chinese search engine, is the best-equipped and most advanced in AI development. Like Google, Baidu has unique advantages in algorithm and data collection, and has a natural inclination to take a lead in AI technologies. Yanhong (Robin) Li, the Chairman and founder of Baidu, has declared AI to be its next "utmost priority" and expressed his optimism about AI as the "lever to new economic prospects" in a recent state media interview.[12] In February, the Chinese State Department requested Baidu to spearhead the National Engineering Lab on Deep Learning Technology and Application with collaborators including Tsinghua University, Beijing Aeronautics and Aerospace University, China Institute of Information, and others.[13] Baidu has currently developed two open-source platforms, DuerOS (Baidu's virtual assistant application) and Apollo (an open-source AI solutions platform) which is offering pilot tools for application development in finance, education, and medical services.[14]

Another major player in the private sector, Alibaba, is a relatively new actor in the development of AI. Unlike its American counterpart, Amazon, Alibaba—the biggest online retailer in China—established its own AI department just two years ago as an extension of its e-commerce platform. As of now, Alibaba's AI technologies are not yet at a level to compete with other global major players such as Microsoft or Google, and still primarily serve as a part of its powerful cloud computing and e-commerce network.[15]

Tencent, the developer of instant messaging services WeChat and QQ, with over 600 million daily active users, has also begun to take advantage of its data resources to become more vocal in the AI theater. Tencent's highest ranked leadership pays a significant amount of attention on its AI department. Tencent's CEO, Zhiping Liu, has repeatedly claimed that AI is a core technology in all of Tencent's products.[16] Tencent encourages every team on every project to expand its involvement in the AI sector, and to apply AI core technologies. At the same time, Tencent is building an experimental AI lab to research fundamental AI technologies. Currently, Tencent has

[10] Ibid.

[11] 国务院关于推进"互联网+"行动的指导意见(The Guiding Opinions Regarding the "Internet Plus" Initiative from the State Council)
http://www.gov.cn/zhengce/content/2015-07/04/content_10002.htm

[12] 李彦宏委员：用人工智能"撬开"关于未来的想象(Commissar Yanhong Li: Using AI to "Crack" Imaginations on the Future)
http://www.gov.cn/xinwen/2018-03/04/content_5270502.htm

[13] http://www.tsinghua.edu.cn/publish/thunews/9659/2017/20170303142537710950910/20170303142537710950910_.html

[14] 百度公开 AI 生态开放战略(Baidu Announces Its AI Ecosystem Strategies)
http://www.xinhuanet.com/tech/2017-07/06/c_1121271415.htm

[15] https://www.leiphone.com/news/201805/5sM1zwCCE1IBo5j7.html

[16] https://ai.tencent.com/ailab/腾讯总裁刘炽平：人工智能具有战略意义，加码投入不急于短期回报.html

invested a significant amount of capital in voice recognition, image recognition, computation visualization, voice processing, and deep learning.[17]

With national strategic leadership, a clear aim to become a global AI leader, and a number of highly invested major corporations, China is emerging as an AI behemoth. The many ways this will influence social and ethical issues of AI remain unclear.

[17] https://www.leiphone.com/news/201704/x1wlWNDfDZJqo3xz.html

References

ACM US Public Policy Council. (2017). *Statement on Algorithmic Transparency and Accountability*. Washington, DC: USACM. Retrieved from file:///Users/michaelbest/UNU/AI4Peace&Society/2017_usacm_statement_algorithms.pdf

Al Jazeera. (2017, June 3). Beyond meat: The end of food as we know it? Retrieved April 18, 2018, from https://www.aljazeera.com/programmes/talktojazeera/2016/02/meat-artificial-food-160205152233913.html

Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., & Rudin, C. (2017). Learning Certifiably Optimal Rule Lists. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 35–44). ACM.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*, *23*.

Asimov, I. (1950). *I, Robot*. Greenwich, CT: Fawcett Publications.

Baer, D. (2016, April 26). This startup is using machine learning to create animal product substitutes. *Business Insider*. Retrieved from http://www.businessinsider.com/chilean-startup-uses-machine-learning-for-meat-subsitutes-2016-4

Banavar, G. (2016). Learning to Trust Artificial Intelligence Systems - Accountability, Compliance and Ethics in the Age of Smart Machines. *IBM*.

Barabas, C., Narula, N., & Zuckerman, E. (2017, September 8). Decentralized Social Networks Sound Great. Too Bad They'll Never Work. *WIRED*. Retrieved from https://www.wired.com/story/decentralized-social-networks-sound-great-too-bad-theyll-never-work/

Barocas, S., & Boyd, D. (2017). Engaging the ethics of data science in practice. *Communications of the ACM*, *60*(11), 23–25. https://doi.org/10.1145/3144172

Beede, D., Powers, R., & Ingram, C. (2017). *The Employment Impact of Autonomous Vehicles*. US Department of Commerce.

Best, M. L. (2011). Mobile Phones in Conflict-Stressed Environments: Macro, Meso and Microanalysis. In M. Poblet (Ed.), *Mobile Technologies for Conflict Management: Online Dispute Resolution, Governance, Participation*. London: Springer.

Best, M. L. (2013). Peacebuilding in a networked world. *Commun. ACM*, *56*(4), 30–32. https://doi.org/10.1145/2436256.2436265

Best, M. L., Long, W. J., Etherton, J., & Smyth, T. (2011). Rich Digital Media as a Tool in Post-Conflict Truth and Reconciliation. *Media, War & Conflict*, *4*(3), 231–249.

Best, M. L., & Thakur, D. (2009). The Telecommunications Policy Process in Post-conflict Developing Countries: The Case of Liberia. *INFO Journal*, *11*(2), 42–57.

Blumberg, A. J., & Eckersley, P. (2009). On locational privacy, and how to avoid losing it forever. *Electronic Frontier Foundation*, *10*(11).

Boddington, P. (2017). *Towards a Code of Ethics for Artificial Intelligence*. Springer.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press Oxford.

Brahimi, L. (2000). *Report of the Panel on United Nations Peace Operations*. Retrieved from http://www.un.org/en/ga/search/view_doc.asp?symbol=A/55/305

Brandusescu, A., Freuler, J. O., & Thakur, D. (2017). *Artificial Intelligence: Starting the Policy Dialogue in Africa*. Washington, DC: World Wide Web Foundation.

Calo, R. (2017). Artificial Intelligence Policy: A Primer and Roadmap. *UCDL Rev.*, *51*, 399.

Clark, J. (2016). Artificial intelligence has a 'sea of dudes' problem. *Bloomberg Technology*, *23*.

Corbett-Davies, S., Pierson, E., Feller, A., & Goel, S. (2016, October 17). A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear. *Washington Post*. Retrieved from https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/

Crawford, K. (2016). Artificial intelligence's white guy problem. *The New York Times*.

Cummings, M. L. (2017). *Artificial Intelligence and the Future of Warfare*. London: Chatham House.

DataProphet. (2018, April). Retrieved April 13, 2018, from https://dataprophet.com/

de Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the Crowd: The privacy bounds of human mobility. *Scientific Reports*, *3*(1). Retrieved from http://www.nature.com/articles/srep01376

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., … Wood, A. (2017). Accountability of AI Under the Law: The Role of Explanation. *ArXiv Preprint ArXiv:1711.01134*.

European Council. (2018, April 10). Declaration of Cooperation on Artificial Intelligence.

European Group on Ethics in Science and New Technologies. (2018, March). Statement on Artificial Intelligence, Robotics and "Autonomous" Systems. European Group on Ethics in Science and New Technologies.

Frey, C. B., & Osborne, M. A. (2017). The future of employment: how susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, *114*, 254–280.

Fundamentally Understanding The Usability and Realistic Evolution of Artificial Intelligence Act, H.R. 4625 § (2017).

Gartner, Inc. (2017). *Predicts 2018: AI and the Future of Work*. Stamford, CT: Gartner, Inc. Retrieved from https://www.commerce-associe.fr/wp-content/uploads/predicts_2018_ai_and_the_fut_342326.pdf

genpact. (2017). *The consumer: Sees AI benefits but still prefers the human touch*. Retrieved from http://www.genpact.com/downloadable-content/the-consumer-sees-ai-benefits-but-still-prefers-the-human-touch.pdf

Good, I. J. (1966). Speculations concerning the first ultraintelligent machine. In *Advances in computers* (Vol. 6, pp. 31–88). Elsevier.

Hern, A. (2018, April 15). Cambridge Analytica scandal "highlights need for AI regulation." *The Guardian*. Retrieved from http://www.theguardian.com/technology/2018/apr/16/cambridge-analytica-scandal-highlights-need-for-ai-regulation

House of Lords' Select Committee on Artificial Intelligence. (2018). *AI in the UK: Ready, Willing, and Able?*

Katz, R. L. (2017). *Social and Economic Impact of Digital Transformation on the Economy* (GSR-17 Discussion Paper). Geneva: ITU.

Kleinberg, J. M., Mullainathan, S., & Raghavan, M. (2016). Inherent Trade-Offs in the Fair Determination of Risk Scores. *CoRR*, *abs/1609.05807*. Retrieved from http://arxiv.org/abs/1609.05807

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, *110*(15), 5802–5805.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

LHoFT, T. (2017, November 8). FinTech for All: Access to finance for Kenya's agriculture industry. Retrieved April 18, 2018, from https://medium.com/@The_LHoFT/fintech-for-all-access-to-finance-for-kenyas-agriculture-industry-f723ca420787

Manyika, J., Lund, S., Chui, M., Bughin, J., Woetzel, J., Batra, P., … Sanghvi, S. (2017). Jobs Lost, Jobs Gained: Workforce Transitions In A Time Of Automation. *McKinsey Global Institute*.

Manyika, J., & Spence, M. (2018, February 5). The False Choice Between Automation and Jobs. *Harvard Business Review*.

Marichal, J. (2012). *Facebook Democracy: The Architecture of Disclosure and the Threat to Public Life*. Farnham, UK: Ashgate Publishing Limited.

Matsuura, S. (2017, November 22). "Mesmo transparente, Brasil tem escândalos", diz criador de robô que analisa gastos públicos. Retrieved April 18, 2018, from https://oglobo.globo.com/sociedade/tecnologia/mesmo-transparente-brasil-tem-escandalos-diz-criador-de-robo-que-analisa-gastos-publicos-22097963

McCorduck, P. (2004). *Machines who think: A personal inquiry into the history and prospects of artificial intelligence*. AK Peters Natick, MA.

Metzinger, T., Bentley, P. J., Häggström, O., & Brundage, M. (2018). *Should we fear the future of artificial intelligence?* Brussels: European Union.

Meyer, D. (2018, May 25). AI Has a Big Privacy Problem And Europe's New Data Protection Law Is About to Expose It. *Fortune*. Retrieved from http://fortune.com/2018/05/25/ai-machine-learning-privacy-gdpr/

Monnerat, A. (2018, January 12). Data scientists in Brazil working on the country's first robot-journalist to report on congressional bills. Retrieved April 18, 2018, from https://knightcenter.utexas.edu/blog/00-19182-data-scientists-brazil-working-country%E2%80%99s-first-robot-journalist-report-congressional-b

National Science and Technology Council. (2016). *Preparing for the Future of Artificial Intelligence*. Committee on Technology. Retrieved from http://itlaw.wikia.com/wiki/Preparing_for_the_Future_of_Artificial_Intelligence

Novitske, L. (2018). The AI Invasion is coming to Africa (and It's a Good Thing). *Stanford Social Innovation Review*.

Osoba, O. (2017). *An intelligence in our image: the risks of bias and errors in artificial intelligence*. Santa Monica, Calif: RAND Corporation.

Petropoulos, G. (2017, April 27). Do we understand the impact of artificial intelligence on employment? Retrieved June 4, 2018, from http://bruegel.org/2017/04/do-we-understand-the-impact-of-artificial-intelligence-on-employment/

Roff, H. M. (2017). *Advancing Human Security through Artificial Intelligence*. London: Chatham House.

Roff, H. M., & Moyes, R. (2016). Meaningful human control, artificial intelligence and autonomous weapons. In *Informal Meeting of Experts on Lethal Autonomous Weapons Systems, UN Convention on Certain Conventional Weapons*.

Roose, K. (2018, March 28). Can Social Media Be Saved? *The New York Times*. Retrieved from https://www.nytimes.com/2018/03/28/technology/social-media-privacy.html

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., … Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Samans, R., & Zahidi, S. (2017). *The Future of Jobs and Skills in Africa - Preparing the Region for the Fourth Industrial Revolution*. World Economic Forum.

Schoeman, W., Moore, R., Seedat, & Chen, J. (2017). *Artificial Intelligence - Is South Africa Ready?* Accenture.

Shane, S., & Wakabayashi, D. (2018, April 4). 'The Business of War': Google Employees Protest Work for the Pentagon. *The New York Times*. Retrieved from https://www.nytimes.com/2018/04/04/technology/google-letter-ceo-pentagon-project.html

Smith, B., & Al-Kohafi, K. (2018, January 29). How Thomson Reuters and IBM are bringing AI to data privacy professionals. Retrieved May 2, 2018, from https://www.ibm.com/blogs/watson/2018/01/thomson-reuters-ibm-bringing-ai-legal-professionals/

Smith, M. L., & Neupane, S. (2018). *Artificial intelligence and human development: Towards a research agenda*. Ottawa, Canada: IDRC.

State Council. (2017, July). New Generation of Artificial Intelligence Development Plan. State Council.

Stauffacher, D., Weekes, B., Gasser, U., Maclay, C., & Best, M. (Eds.). (2011). *Peacebuilding in the Information Age: Sifting hype from reality*. Geneva, Switzerland: ICT4Peace Foundation.

Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., … Kraus, S. (2016). Artificial intelligence and life in 2030. *One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel*.

Stupp, C. (2018, April 10). Twenty-four EU countries sign artificial intelligence pact in bid to compete with US & China. Retrieved April 18, 2018, from https://www.euractiv.com/section/digital/news/twenty-four-eu-countries-sign-artificial-intelligence-pact-in-bid-to-compete-with-us-china/

Surber, R. (2017). *Artificial Intelligence: Lethal Autonomous Weapons Systems and Peace Time Threats*. Zurich, Switzerland: ICT4Peace Foundation.

Surber, R. (2018). *Artificial Intelligence: Autonomous Technology (AT), Lethal Autonomous Weapons Systems (LAWS) and Peace Time Threats*. Zurich, Switzerland: ICT4Peace Foundation.

The Economist. (2016a, June 25). Automation and anxiety; The impact on jobs. *The Economist*, *419*(8995).

The Economist. (2016b, June 25). March of the machines; Artificial intelligence. *The Economist*, *419*(8995).

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2017). *Ethically Aligned Design: A Vision for Prioritizing Humman Well-Being with Autonomous and Intelligent Systems* (No. Version 2). IEEE. Retrieved from http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433.

Vinge, V. (1993). The coming technological singularity: How to survive in the post-human era. Retrieved from https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19940022856.pdf

Weinberger, D. (2017, April 18). Our Machines Now Have Knowledge We'll Never Understand. *WIRED*.

Weinberger, D. (2018, January 28). Don't Make Artificial Intelligence Artificially Stupid in the Name of Transparency. *Wired*. Retrieved from https://www.wired.com/story/dont-make-ai-artificially-stupid-in-the-name-of-transparency/

Weissman, C. G. (2016, August 18). The Women Changing the Face Of AI. Retrieved April 10, 2018, from https://www.fastcompany.com/3062932/ai-is-a-male-dominated-field-but-an-important-group-of-women-is-changing-th

Weston, M. (2015, January 4). Digital Translation in an Analog Country. *Myanmar Business Today*, *12*(51). Retrieved from https://www.mmbiztoday.com/articles/digital-translation-analog-country

World Bank. (2016). *World Development Report 2016: Digital Dividends*. Washington, DC.

World Wide Web Foundation. (2017). *Artificial Intelligence: The Road Ahead in Low and Middle-Income Countries*.

Yudkowsky, E. (2008). Artificial intelligence as a positive and negative factor in global risk. *Global Catastrophic Risks*, *1*(303), 184.