Input to United Nations Special Rapporteur on Extreme Poverty and Human Rights
Regarding Visit to the United States
October 2017

This input to the Special Rapporteur is provided respectfully by the following parties:

- Edward W. Felten, Robert E. Kahn Professor of Computer Science and Public Affairs, Princeton University, and former Deputy United States Chief Technology Officer;
- Bendert Zevenbergen, Visiting Professional Specialist, Center for Information Technology Policy.

We respond specifically to the following item in the Special Rapporteur's request for input:

> *"There is an increasing debate worldwide on the impact of new technologies on societies, including in the area of Artificial Intelligence, robotics, Big Data and algorithmic decision-making. How do these developments affect the human rights of those living in poverty in the United States? The Special Rapporteur is interested in learning how these technologies may affect civil and political rights as well as economic and social rights."*

Definition of artificial intelligence ("AI")

The concept of AI has been proven to be notoriously difficult to define. A basic though popular definition of AI refers to "intelligence exhibited by machines" or "the science and engineering of making intelligent machines." These definitions assume that 'intelligence' is clearly defined itself, though it, too, is ambiguous. No commonly agreed upon definition of artificial intelligence currently exists. Therefore, the scope of what is and isn't considered to be AI is flexible.

A distinction must be made between *narrow* AI and *general* AI (or 'artificial general intelligence'). Narrow AI is created to address specific application areas, where machines typically outperform humans in terms of speed, accuracy, and efficiency. Successful applications of narrow AI can be found in many parts of society. General AI requires systems to exhibit intelligent behavior that is (at least) as broad, adaptive, and advanced as a person across the full range of cognitive tasks. While it is unlikely that general AI will be achieved within the next few decades, it is expected that specific tasks performed by humans can and will be replaced by narrow AI applications on an ongoing basis.

Why discuss AI now?

Algorithms are increasingly replacing humans in decision making or calculations. This opens up new opportunities for efficiency and progress in areas such as health, education, commerce, communication, energy, and the environment. Citizens and social groups are also increasingly scrutinized by computers and algorithms in areas such as personal finance, tax, insurance, immigration, and law enforcement.

AI has been in development since the 1930s and became a separate field of research in the 1950s. While the AI research community has experienced wide swings between optimism and pessimism over the years, there has been steady progress on underlying technical challenges. There is now a sufficient the sociotechnical constituency to bring AI into broad use: computing power and capacity has increased significantly, vast datasets exist to train AI systems, and funding appears to be readily available, primarily from industry. Innovations and deployment of AI systems to mediate an increasing amount of activity warrants a cautious regulatory and informed approach.

AI risks and concerns

Many fears and doom scenarios about the deployment of AI in society have been expressed in the media, academic literature, and government communications. The most extreme fears have received disproportionate attention. While it would be unwise and unjustified to dismiss these fears as mere science fiction, one should recognize that these fears are speculative and receive considerable skepticism within the technical community. It is important to address expressed fears, but also to state the relevant assumptions and technological forecasts of these scenarios.

Some of the shorter term and less extreme concerns are well justified, however, given current technological progress. First, the outputs of individual systems have been shown capable of exhibiting unfair, irresponsible, and even racist tendencies. While these properties are not necessarily programmed into the systems wittingly, they may result from various causes including disproportionate representation in datasets, replication of past biased practices of human decision makers, and emergent properties of facially neutral statistical methods. More broadly, there is concern about the difficulty of ensuring that a complex AI system will behave consistently with the values of the organization deploying it or of society generally.

AI systems may also surface issues that have been latent in the past, such as how to trade off goals of efficiency or accuracy against the goal of fairness and nondiscrimination. To the extent that an AI system requires putting numerical weights on the importance of these factors, or requires defining precise numerical metrics for bias, it may force a more precise discussion of issues that had been treated in generalities before.

However, the proprietors of an AI system may not want to reveal their source code or training data, thereby making the decision processes of AI systems opaque. For this reason or other technical reasons, an AI system may appear to be a “black box” which may be difficult to understand, monitor, or govern. The increasing reliance of sectors or society as a whole on AI systems that were inscrutable would pose risks to public values such as social mobility, equality, inclusion, and autonomy. When the systems that govern society are developed by a few, without the possibility of public scrutiny, then the goals specified by those few may come to dominate public and private life. Systematic governance of society by AI systems may lead to a society where the values those AI systems are trained to optimize are automatically given preference over other values.

Relevance of human rights framework

The human rights framework offers a substantive and internationally legitimate moral system that can be an important starting point for international debates about the design, engineering, and governance of AI technologies. Rather than merely debating which values to incorporate into systems and their governance, the human rights framework provides the base for questions about how to operationalize the variety of rights in AI systems. We agree that this viewpoint would benefit from more attention in discourse about AI. More scholarship exists on the intersection of information technologies and informational human rights, such as privacy and freedom of speech, than social and economic rights.

How human rights can be affected

As the availability of AI systems increases, there has been greater deployment of AI to replace decisions in public life, to shape markets, and to influence the distribution public resources. AI is also increasingly used in many private sectors, for example to make decisions about the employment or allocation of tasks and working hours. As the reliance on AI to make decisions about people and groups increases, the scope of affected rights also increases. Previously, when the Internet was mostly used as a communication and information retrieval system, the rights affected were mostly related to privacy and freedom of speech. The increasing use of AI to make consequential decisions about people and their legal status leads to deeper concerns about how to ensure justice and fairness on a broader scale.

When AI systems make decisions about the allocation of resources and opportunities (e.g. public health, law enforcement, tax collection, immigration), the increased technical mediation should be assessed from an administrative, legislative, and human rights point of view. The use of AI does not merely make decision-making more efficient, but it changes the operation of public institutions. AI systems can obfuscate the process of decision making. While AI systems can be designed merely as a decision support tool, it may prove to be difficult or risky to overrule the system's decision, thereby giving the system authority and power. Human rights will inevitably be affected.

On a positive note, AI systems can increase transparency, fairness, and governability, if used in the right way. For example, emerging research is showing that it is possible to get a sense of the level of bias and discrimination in AI systems, to design algorithms that resist bias, and to manipulate the inevitable trade-off between utility and fairness of a system to a desired or appropriate point. There is much promising research in the area of fairness in AI systems that will be particularly valuable for the design of systems going forward. There is a role for law and policy to set standards that guide this trade-off in engineering and governance of AI systems. An overview of these solutions would be beyond the scope of this response, but we are happy to discuss this emerging research field further.

The role of law, transparency, due process, and accountability

The tools currently at the disposal of policy makers and citizens to understand and challenge decision making and interferences with human rights may not be sufficient to appropriately address automated decision making by AI systems. To ensure the interests and human rights of citizens, the tools and interpretation of human rights must be assessed in the new socio-technological environment that is being created. However, due to the complexity and adaptability of AI systems, commonly suggested solutions such as simple transparency and auditability may not be sufficient by themselves to protect human rights.

AI is already subject to regulation, when it is used in existing products that are regulated, such as cars, aircraft, or procedures such as medical interventions. In these specific sectors, the impact of AI should be assessed based on the objectives of existing regulatory frameworks. If the use of AI increases or decreases the scope or magnitude of risks, the regulatory system may warrant review. It is important, however, that regulatory review does not retard the potential for socially beneficial innovation.

Systems that have an effect on the human rights of individuals should ideally be open and accountable. In some cases, however, it may be necessary to keep parts of systems secret to constrain any 'gaming' of a system. Designers of AI systems should be encouraged to build their systems in such a way that the possibility for scrutiny is considered from the start. Law and regulation have a role to pay to set the standards for the accountability and governability of AI systems, while discouraging manipulation of the auditing process.

Suggested contacts

- U.S. National Science and Technology Council Subcommittee on Machine Learning and Artificial Intelligence (This is an interagency committee of executive branch agencies within the U.S. government, which is charged with policy coordination relating to machine learning and AI: https://www.whitehouse.gov/sites/whitehouse.gov/files/ostp/MLAI_Charter.pdf
- Data Cabinet (the federal data science community of practice): https://ntis.gov/thedatacabinet/
- Partnership on AI to benefit people and society: https://www.partnershiponai.org/
- Fairness, Accountability, and Transparency in Machine Learning: http://www.fatml.org/
- Academic centers on AI in the U.S., including:
    - MIT: http://news.mit.edu/2017/mit-media-lab-to-participate-in-ai-ethics-and-governance-initiative-0110
    - Harvard: https://cyber.harvard.edu/research/ai
    - UC Berkeley: http://humancompatible.ai/