HUMAN RIGHTS, ETHICS & ARTIFICIAL INTELLIGENCE

Challenges for the Next 70 Years of the Universal Declaration of Human Rights

CONFERENCE REPORT
January 2018

HARVARD Kennedy School
CARR CENTER
for Human Rights Policy

INTRODUCTION

In early December 2018, the Carr Center for Human Rights Policy, the Edmond J. Safra Center for Ethics, and the Berkman Klein Center for Internet and Society hosted an inaugural conference that aimed to respond to the Universal Declaration of Human Rights' 70th Anniversary by reflecting on the past, present and future of human rights. The conference was organized by Carr Center Faculty Director Mathias Risse.

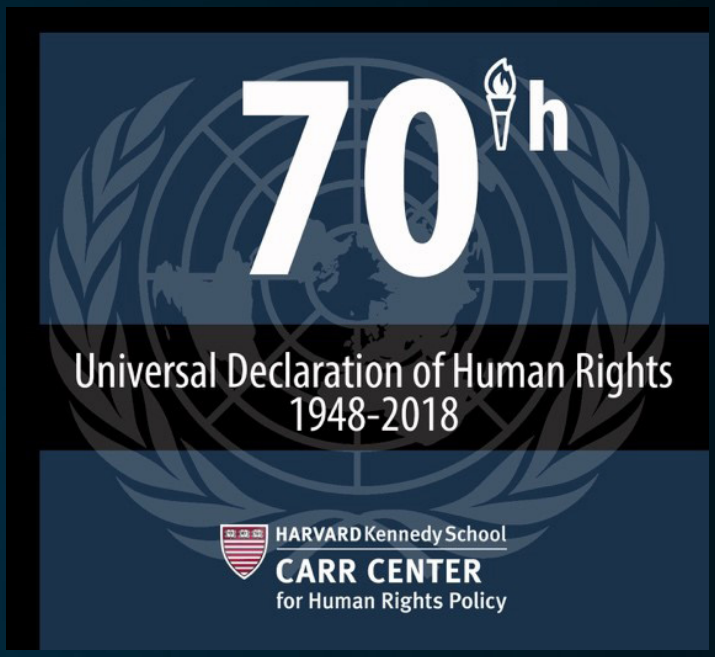
The Conference sought to:

1. Discover how the institutional protection of distinctly human life can be assured in an era of exploding technological innovation.

2. Explore how innovation engages human values and a broad range of ethical issues that arise from our ways of using technology and perhaps eventually from sharing our lives with forms of technology that for now are merely imaginable.
3. Bring together many emerging initiatives in this domain at Harvard to make sure the ethical dimensions of these changes take center stage in Harvard's agenda.

This inaugural conference brought together practitioners and academics from different industries, disciplines, and professional practices. This discussion on Human Rights, Ethics, and AI explored the ways modern and emerging technology could redefine what human rights will look like in the next 70 years. Speakers discussed the latest advances in AI technologies, the need for ethics, and what human rights could look like in the advent of such technologies.

CONFERENCE BACKGROUND



"Now is also the time to look ahead to the next 70 years, an era likely marked by technological breakthroughs at an unprecedented scale."

December 10, 2018, marked the 70th Anniversary of the Universal Declaration of Human Rights. The human rights movement with its moral, legal and organizational ramifications has been one of the great innovations of the 20th century. Much has been achieved, but the world continues to fall short of human rights ideals in numerous ways.

Now is also the time to look ahead to the next 70 years, an era likely marked by technological breakthroughs at an unprecedented scale. Artificial Intelligence is increasingly present in our lives, reflecting a growing tendency to turn for advice or turn over decisions altogether, to algorithms. The effectiveness of algorithms is increasingly enhanced through Big Data: the availability of an enormous amount of data on all human activity.

The key idea behind human rights is that power can be used in ways that protect and advance basic human interests, especially those of the most vulnerable and that there is a global responsibility to make sure of this. New technology greatly helps with these goals, but there are also dangers. In the short term, these dangers range from the perpetuation of bias in algorithmic reasoning to use of technology to create "alternative facts." In the longer term, we must worry about ownership of data, increasing power of tech companies, changes in the nature of work and perhaps eventually existential risk through the advent of entities that are not alive in familiar ways but might be sentient and intellectually and even morally superior to humans.

This conference explored ways in which technological change will affect human life, especially the rights designed to protect that life, in the next 70 years and beyond. At stake is nothing less than the future of human life and its organizational possibilities, what that life will look like, and whether it might eventually be threatened at an existential level by its own innovations.

Harvard Kennedy School
79 JFK Street, Cambridge, MA 02138
Wexner Building, 4th Floor, Room 434AB
Rubenstein Building, 4th Floor, Room 414AB

Day 1, Nov. 29

5:00 PM - 6:30 PM
Opening Keynote

6:30 PM - 9:00 PM
Reception & Dinner

Day 2, Nov. 30

8:15 AM - 9:00 AM
Registration & Light Breakfast

9:00 AM - 10:15 AM
From Ethical Challenges of Intelligent Systems to Embedding Ethics in Computer Science Education

10:15 AM - 10:30 AM
Break

10:30 AM - 11:45 AM
Accounting for Artificial Intelligence: Rules, Reasons, Rationales

11:45 AM - 12:00 PM
Break

12:00 PM - 1:15 PM
The End of Human Exceptionalism?

1:15 PM - 2:15 PM
Lunch

2:15 PM - 3:15 PM
Flash Presentations by Members of the Inaugural Techtopia Group

3:15 PM - 3:30 PM
Break

3:30 PM - 4:45 PM
How to Get Empowered, Not Overpowered, by Artificial Intelligence

4:45 PM - 5:00 PM
Break
5:00 PM - 6:15 PM

What's Worse: Machine Learning That Doesn't Work, Or Machine Learning That Does?

Day 3, Dec. 1

9:00 AM - 10:00 AM
Light Breakfast

10:00 AM - 11:15 AM
Transcending the Brain? AI, Radical Brain Enhancement, and the Nature of Consciousness

11:15 AM - 11:30 AM
Break

11:30 AM - 12:45 PM
Human Rights and Artificial Intelligence: The Long (Worrisome) View?

12:45 PM - 2:00 PM
Lunch

2:00 PM - 3:15 PM
Panel Discussion: Artificial Intelligence and Human Rights Law in the Here and Now

3:15 PM - 3:30 PM
Break

3:30 PM - 4:45 PM
Contextualizing Calls for AI Transparency and Fairness

4:45 PM - 5:00 PM
Break

5:00 PM - 6:15 PM
Responsible Business in the Digital Economy: What's New

6:15 PM - 6:30 PM
Closing Remarks

SKEWED OR RESCUED?: THE EMERGING THEORY OF ALGORITHMIC FAIRNESS

CYNTHIA DWORK

Gordon McKay Professor of Computer Science at the John A. Paulson School of Engineering and Applied Sciences at Harvard

In this talk, Professor Dwork introduced the emerging theory of algorithmic fairness and the challenges experts face in ensuring that machines make objective decisions. Using data in new ways to make more accurate predictions or enabling new services, intelligent systems offer the hope of overcoming the limitations of our own decision-making. However, with this they bring questions about key questions, such as if fairness mechanisms are implicit.

Professor Dwork has been working on the problem of algorithmic fairness since 2010. She set out to discover fairness in algorithms, such as in the decision making of advertising platforms in light of differential unfairness. She discussed how algorithms, though right most of the time, are inherently unfair when they are wrong. Algorithms must be more aware of the constituent populations it makes decisions about.

Machine learning algorithms try to label things, such as cats and dogs. They could also be trained to learn about large populations and provide a likelihood of something. Dwork argued that an algorithm is no smarter than the data that is fed to it—if there is a bias in the data, then the algorithm will adopt that bias. Furthermore, humans must help the algorithm

along the way in order to be accurate and fair—historical data is not enough.

So how can we construct fair algorithms? We must first define what unfairness is and know how to prevent it—by doing so we can discover what is fair. We must then mathematically construct algorithms that are fair according to the definition. Lastly, we must prove that systems built from fair pieces are fair in toto. Composition matters.

In her lecture, Professor Dwork focused on two types of algorithms: binary classification algorithms, which classify each person as a positive (high risk of disease) or negative (low risk), and probability estimation algorithms, in which each output has a different probability of an outcome ($f(x) = p$, where each individual might have a different value for p). We should also think about the choices of features in algorithms. We should ask, is there something not included that should be included?

"How can we construct fair algorithms? We must first define what unfairness is and know how to prevent it."

Professor Dwork also discussed the need for group fairness in algorithms and its weaknesses. She mentioned that group fairness properties are statistical requirements. Statistical parity is achieved when demographics of people assigned positive (negative) classification are the same as the demographics of the general population. Individual fairness overcomes some of the weaknesses of group fairness, but it is difficult to get metrics to achieve individual fairness.

Before concluding, Professor Dwork also outlined a few additional approaches to achieving fairness. Algorithms could learn a "fair representation" which stamps out sensitive information but retains sufficient information to permit standard training. There could be a fairness calibration for very large numbers of large, overlapping groups. Lastly, they could be causality based; for example, it could rule out effects being "caused by" being in "S."

In her final remarks, she considered how computer science can help in algorithmic fairness through multi-calibration. Through multi-calibration, an algorithm can learn biases in multiple, possibly intersecting groups, and can cover traditionally underrepresented groups. This will tell you if there has been a group in a larger pool that has been mistreated. But computer science cannot manage without wise inputs. Defining fairness is a political process. Different definitions are appropriate to different contexts, and fairness will be radically different to different people, which makes it very difficult to design fair algorithms.

OPENING REMARKS

Mathias Risse

Lucius N. Littauer Professor of Philosophy and Public Administration, HKS
Faculty Director, Carr Center for Human Rights Policy

"Throughout this conference, we will explore what human rights challenges lie ahead in light of technological innovation."

Faculty Director of the Carr Center for Human Rights Policy, Mathias Risse, opened the conference by looking forward at the future of human rights for the next 70 years, asking questions such as, 'how are we falling short?' and 'what lies ahead?'

Technological innovation is only going to increase over the next 70 years. "Throughout this conference, we will explore what human rights challenges lie ahead in light of technological innovation," he said. Risse continued by elaborating on the content of upcoming discussions, and outlining the program for the next two days.

FROM ETHICAL CHALLENGES OF INTELLIGENT SYSTEMS TO EMBEDDING ETHICS IN COMPUTER SCIENCE

BARBARA GROSZ

Higgins Professor of Natural Science, John A. Paulson School of Engineering and Applied Sciences, Harvard University

In light of the increasing success of AI, Grosz believes that the ethical path forward is through combined human and AI systems. As a result, Grosz is interested in building ethics principles into all of computer science.

Grosz defined AI in two ways: AI Science and AI Systems. AI Science is the development of a computationally based understanding of intelligent behavior that has two prongs: cognitive science (understanding intelligence) and the development of theories, models and algorithms that embody that intelligence. AI Systems, on the other hand, incorporate AI methods to enable systems to reason intelligently and act autonomously through interpreting language, learning, drawing inferences and making decisions.

She clarified that AI is not just deep learning and that there are two types of AI methods: Model-Based and Data Dependent. Model-based AI methods are rooted in symbolic AI (logical reasoning, decision-theoretic reasoning, game theoretic) and in probability (Bayesian reasoning, sequential decision-making models and algorithmic game theory). Data Dependent models, based on statistical reasoning and data, focus on neural nets and deep learning. She argues that while model-based methods can explain reasoning with general theories and model semantics, it will still require handcrafted methods with expert input. As for the Data Dependent methods, Grosz also argues that we will need lots of data and some experts at some point, as the results are descriptive, not causal.

Grosz then provided the good news of AI, namely in the field of fraud detection, conversational agents, translation, provision of recommendations, self-driving vehicles and surgical robots. She states in many ways, AI exceeds human accuracy and the progress is boundless (AI can better distinguish rare breeds than humans can) yet limitations do exist. For example, AI often has dialogue gaps because humans rarely clarify pronouns for extended periods of time. Furthermore, there are ethical challenges from learning from bad sources.

> "Trying to get rid of people is a mistake. For a system to be smart, it has to work well with people. Every citizen should demand that. A crucial change of approach is needed: focus on AI complementing people rather than replacing people."

Grosz argues that with the benefit of increased accuracy through AI systems, it is critical that AI works with people and does not replace them, as the sum of the two is greater than its parts alone.

She argues that AI + human intelligence provides better machine learning results and that they should work together during the training, evaluation, and execution of AI systems.

In light of the Universal Declaration of Human Rights, Grosz observes that the declaration provides all positive commands—what one should do. However, it is important to discuss the ways in which AI can help us. Ethics requires more than just harm avoidance. For example, health care coordination can be boosted with efficient information sharing powered by AI. Grosz says that we should use AI to do good, not just avoid harm as ethical principles concern doing good and not just avoiding doing evil. She believes that values should be decided by people and focusing on "robot takeover" scenarios could distract from issues that are important now.

Grosz highlights that people are not talking about ethics in computer science classes. In her own class, CS108, she found that those who were charged with making revenue from ads while building a social media company did not even think to design a system that was both efficient but also ethical. Ethics must be considered from the start. Grosz said the goal is to integrate ethics into CS curriculum, providing ethics expertise deep and as broad as computer science. Students must learn to identify and reason clearly about the ethical implications of technology while they are learning ways to develop and implement algorithms, design interaction systems, and code, and not as a separate endeavor.

Grosz believes we should "embed ethics" into the teaching of CS courses. She imagines a curriculum that has Ethics Ph.D.'s and post-docs identifying ethics problems in classes and addressing them in class sessions with learning activities. Grosz also believes that ethics is everyone's responsibility. If we are not in CS, it is important to learn enough to ask the right questions. If one is a technologist—learn enough about ethics to incoporate that knowledge into your work.

ACCOUNTING FOR ARTIFICIAL INTELLIGENCE: RULES, REASONS, RATIONALES

SOLON BAROCAS

Assistant Professor, Department of Information Science, Cornell University

Professor Solon Barocas believes that algorithms implicate basic human rights because they are often secret and opaque black boxes that are rarely made transparent, operate on the basis of correlations rather than causality and produce predictions rather than explanations, have behavior that could lack intelligibility and foreseeability, and challenge established ways of being informed.

Effectively, algorithmic decisions are arbitrary. As a result, Barocas calls for machines that explain themselves. In this session, Barocas covered machine learning problems in explanation, the legal requirements for explanation, the techniques for furnishing explanations, the value of explanations, the limits of intuition as a basis for normative evaluation, and documentation as an alternative to intuition.

Barocas reminded us of *The Code of Fair Information Practices* from 1973, where it was stated, "There must be no personal data record-keeping systems whose very existence is secret." From a human rights perspective, we care about why decisions are made the way they are. It is very different to answer questions on how decisions are made and why the

decision is made that way. Essentially, it's rules and reasons versus rationales.

There are also legal requirements for explanations, initially not about how decisions were made by companies but for how information was collected. Companies were required to give us adverse action notice but were not required to reveal the full logic of decision making or the justification for making the decision that way. It is critically important to provide meaningful information about the logic involved.

There are several techniques one could use to describe the operation of a model. We could purposefully orchestrate the learning process such that the resulting model would be interpretable. We could also apply special techniques after the model creation to either approximate the model in a more readily intelligible form or identify features that are most salient for specific decisions. Lastly, we could also provide tools that allow people to interact with the model and get a sense for its operation.

It is important to build interpretable models to understand the logic of decision making. When rules are given, we need to be able to understand why or at least a set of reasons for the particular outcome. Neither the law nor the techniques listed above, however, go beyond describing the operation of the model.

There are some core values to the explanation of AI systems. There is an inherent value of explanation because the fears about a system that lacks explanation are visceral. Giving people explanations is also pragmatic because explanations are actionable. There is an ongoing debate about the best way to explain decisions in an actionable way, which seems to be converging on the belief that explaining specific outcomes is the right approach. The focus has therefore shifted to the different methods by which specific decisions can be explained: the main factors in a decision, the minimum changes required to switch the outcome of a decision, and similar cases with divergent outcomes or divergent cases with similar outcomes. Lastly, there is a justificatory value in explanation in that if we give reasons that are good enough, we can make better decisions.

Machine Learning Explanatory Problems

There are different forms of machine learning explanatory problems in AI systems:

- When the existence of some decision making process is simply unknown, the system is secret.

- When the existence of some decision making process is known but the actual operation is not, the system is opaque.

- When a decision making process is disclosed but defies practical inspection, even by its developers and experts, the system is inscrutable.

- Lastly, Barocas says that when a decision making process is disclosed but it defies intuition, the system is non-intuitive.

THE END OF HUMAN EXCEPTIONALISM?

STEPHEN WOLFRAM

Founder and CEO, Wolfram Research

Wolfram believes that technology is the harnessing of what exists in the world to achieve human goals. Before, humans harnessed the physical world, mining for minerals and liquid crystals. Now, humans can also mine the computational universe. Modern machine learning mines algorithms from the computational universe, not looking at what is out there incrementally but at all of it. How do we tame the ocean of computational possibilities and map it into human goals? How do we tell AI what to do? There is a computational ocean out there, and machine learning can tap into it. But we need to know how to get relevant information to help solve human problems. These are the questions we should be asking for useful purposes.

Computational language is made to take what humans want and compute it. It can map the world into something we can actually compute with. For every human endeavor, x, there is a computational version of x. At high levels of sophistication, when you are out of the domain of rules, you very quickly end up where computational sophistication is the same at a universal level. For example, computational sophistication at a sufficiently high enough level is the same in weather applications, simple applications, and even human brains, a concept known as universal computation. This comes to show that it is not true that only humans can attain sophisticated computations—anyone or anything with enough computational sophistication

"If we are making AI's that run the world, what is the contract that humans need to make with AI's?"

can. This form of intelligence that we have is the first example of alien intelligence, and we must realize that humans are not special in an intellectual way. We think we have a purpose but is there human specialness? The answer is no. We think our intellect is the pinnacle of achievement, but it is not true. We are important because we have a special history throughout evolution and cultural progression. But there is an ocean of possible computation out there, and at certain levels of computational sophistication, there is compatibility and universality.

So, what are the computational implications for humans and democracy? We do not yet have widespread computational literacy—perhaps at some point, but we will eventually have general computational literacy. People could theoretically express their emotions not by voting but by writing a computational essay. An AI could then use everyone's computational essays to figure out the best way to run the government. But there are important questions we must answer first. If we are making AI's that run the world, what is the contract that humans need to make with AI's? How do we think about doing something like that? How do we discuss computational representativeness? How can we know what humanity wants?

These kinds of questions Wolfram claims must be answered before a completely computational future for humanity arrives.

INAUGRUAL TECHTOPIA GROUP – FLASH PRESENTATIONS

The following four flash presentations were given by undergraduate and graduate students as part of Techtopia @ Harvard University. Techtopia's mission is to illuminate and expand Harvard's teaching and research efforts on the ethics and governance of emerging digital technologies and to equip students with the interdisciplinary problem-solving skills required to tackle the pressing social, political, ethical, and legal issues at this intersection. It is a new program led by Faculty Director of the Berkman Klein Center for Internet and Society, Jonathan Zittrain, in an effort to bring together faculty and students.

JENNY FAN

Harvard Graduate School of Design

Jenny Fan is designing collaborative adjudication systems for moderating dangerous speech online. She argues that information disorder on social media platforms is an embodiment of the tragedy of the digital commons typically solved by collective action or regulation. Information disorder can be tackled by regulation. She considers several existing types of regulation. Regulation by markets in the digital space is based on attention driven business models reliant on ad revenue. Content moderation is curated to drive users into niche platforms. Regulation by law provides platform community standards and an appeals process through the *Santa Clara Principles of Transparency and Accountability in Content Moderations*. There is also regulation by code, such as rule-based moderation systems, algorithmic/AI efficacy, dog-whistles, coded language, and shadowbans. Social norms are also regulatory although not all social norms transfer digitally.

In the digital sphere, scale prevents consistent moderation standards across communities. There are also meta modernization problems—who checks the fact-checkers? What should the ethics of rating users on trustworthiness look like? What are the rights and responsibilities of the citizens who inhabit cyberspace?

LUIS VALLES
Harvard Business School

Luis Valles asks if recommendations on digital platforms erode the freedom of choice. If they do, he wonders how digital businesses can align profits with ethics. Businesses use data to generate user recommendations on their platforms. Business models introduce recommendation bias by defining what data is and is not gathered. Algorithmic structures introduce recommendation bias by how they analyze data. Algorithm type combined with data provides a prediction with statistical significance but can miss key relationships and key variables.

Depending on the method you use, you introduce new biases. How can biased recommendations erode our freedom of choice? It is our choices that drive behavior. Recommendation algorithms do not capture changing priorities over time. Under this system, it becomes very hard to change your own behavior. The solution will require new business action and different user engagement with platforms. Companies need to design their business structure and algorithmic structure to limit ethical risk and users need to evaluate AI output. Valles' next steps are to partner with digital platforms, define a model for users to assess algorithm output, meet users face to face for feedback, and define a preliminary point of view of ethical risks driven by a firm's specific business structure and algorithmic structure.

JOSHUA SIMONS
Harvard Graduate School of Arts and Sciences

Joshua Simons is studying the politics and political implications of machine learning in democratic life. He says that the choices you make in the process of designing a machine learning system are political. For example, in the Facebook Trustworthiness Score, choices here are target variables—it is how Facebook as a company defines trust. Explaining a model is not sufficient for accountability. It may even distract and detract from it. Is it enough for Facebook that it shows you how it displays its newsfeed? They must do more and showing you the mechanisms simply distracts consumers from demanding more.

For accountability, Simons argues that institutions must justify their choices in the design and integration of machine learning models. They must justify how they make these choices. Since machine learning design is political, political tools require political oversight in a world of algorithmic bureaucratic systems.

IRENE SOLAIMAN

Harvard Kennedy School

Irene Solaiman is studying cybersecurity in elections, security, and privacy in intelligent systems. She believes the balance of data protection and the promotion of research is the key to privacy in the age of AI.

Solaiman argues for public-private collaboration and regulation to protect personal data and privacy. Information is less secure than it was five years ago. Individual cyber hygiene is compromised. The US does not have a national, cohesive policy on privacy. HIPAA is a law that can interfere with FERPA which can conflict with COPA. These laws move in the right direction, but they are outdated and there is little enforcement. In the European Union on the GDPR, data belongs to the user, who has the right to be informed, of access, to rectification, to erasure, to restrict processing, to data portability,

"How can biased recommendations erode our freedom of choice?"

and to object in relation to automated decision making and profiling. Should we be blindly creating new technologies without considering ethical implications? Can regulation spur innovation as well?

Solaiman says for average citizens, the most important thing to do is to clean up their cyber hygiene. For the United States, we need a cohesive and coherent regulatory framework like how we regulate food. We need to incorporate labels, rules, and quality control. This framework also needs to have teeth. The US government has an obligation to protect their citizens.

HOW TO GET EMPOWERED, NOT OVERPOWERED, BY ARTIFICIAL INTELLIGENCE

MAX TEGMARK

Professor of Physics Massachusetts Institute of Technology

The cosmic perspective on life is both humbling and inspiring. Life's beginning was simply biological. Then, it became cultural. Now, life is technological. The parameters for life now ask the questions: Can it survive and replicate? Can it design its own software? Can it design its own hardware?

Humanity is now journeying forward into an AI future as a community. In rocketry, you not only need power, but you also need to be able to steer. The same is true for AI—the steering is as important as the destination. Intelligence is the ability to accomplish complex goals. The key insight behind AI is that intelligence is all about information processing—and it can be biological (human) or artificial.

AI has the power to save lives on the road and through improving hospital care. There is better science with AI, with more accurate diagnoses. AI

"Humanity is now journeying forward into an AI future as a community. In rocketry, you not only need power, but you also need to be able to steer. The same is true for AI—the steering is as important as the destination."

defeated humans at the game Go, which defied millennia of human intuition. So how far will this progress go? Is Artificial General Intelligence (a superintelligence in all areas of life) a crazy scientific idea? A recursively self-improving AI is certainly possible—chief technologists and philosophers both believe that AGI will be achieved in decades. This may or may not occur—but if this happens, then what? Humanity could just build machines and not worry about complications and consequences. We face two options: we can be complacent or we can be ambitious. Tegmark opts for the latter, envisioning an inspiring, high-tech future, where we can steer AI to help humanity flourish.

As we look forward to the next 70 years of the Universal Declaration of Human Rights, it would be wise to consider a Universal Declaration of Rights—without "Human." We must protect not only animal rights but also the rights of AI. Conscious entities and societies should have the freedom to pursue their own goals unless this conflicts with an overriding principle. As long as no one else suffers from it, an entity (not specifically human) should have the freedom to think, learn, communicate, own property and to not be harmed, a right to do whatever doesn't infringe on the freedoms of others.

Tegmark believes that humanity must be proactive instead of reactive in terms of the advent of new technologies. When humanity invented fire, we also invented the fire extinguisher. When we invented the car, man was able to invent seatbelts and traffic lights. When mankind invented nuclear power and synthetic biology, we were going beyond the threshold of safety. When AGI is invented, we will be well beyond the threshold of safety. And we cannot be behind the threshold, so we need a new strategy. AI turning evil and turning conscious are mythical worries. The true worry is AI turning competent with goals misaligned with the goals we have for humans.

Where do we want to go? Few discuss this because we are so focused on the future of work and killer robots. AGI is coming, and we need to decide what we want this future to be like. If we are complacent, this will lead to a world of massive misery. But if we are ambitious, everyone is better off. By building AI that empowers us rather than overpowers us, we can have a better future.

WHAT'S WORSE: MACHINE LEARNING THAT DOESN'T WORK OR MACHINE LEARNING THAT DOES?

JONATHAN ZITTRAIN

George Bemis Professor of International Law, Harvard Law School, HKS Faculty Director, Berkman Klein Center for Internet & Society

Professor Jonathan Zittrain explores a legal perspective on artificial intelligence to understand the fairness of machine learning. In law, the legitimacy of the system depends on us not knowing how it works, which is a feature, not a bug. It is also part of the social acceptability of the role of the jury not to explain itself. But how machines make decisions and how human experts make decisions can be vastly different. Which way is fairer?

How machine learning works introduces new problems in fairness. We must consider that something can be fair from the machine learning system's point of view but unfair from the individual's point of view. For example, if someone has the right to an individualized sentence, what would that mean for the machine learning system? Will it rely on generalizations of people who are similar to sentence the individual? Are people being treated fairly if a machine learning algorithm assigns a higher-risk rate to people with red shirts, based on the fact that 9 out of 10 people with red shirts tend to be violent? Is it fair to set a high bail for that individual even though the individual has a 1 in 10 chance of being safe?

People's freedoms are at stake. People have a right to an individualized judgment. But there are many ways to achieve 'fair.' In light of this, we must seek proscriptive rather than prescriptive interventions in machine learning.

TRANSCENDING THE BRAIN? AI, RADICAL BRAIN ENHANCEMENT, AND THE NATURE OF CONSCIOUSNESS

DR. SUSAN SCHNEIDER

Director, AI, Mind and Society Group, The University of Connecticut, Distinguished Scholar, US Library of Congress, Institute for Advanced Study, Princeton, New Jersey

Dr. Susan Scheider opened her presentation remarking that AI will be the product of design but that humans will be the designers.

When it comes to changing the human mind, we must be cautious as there are stakeholders who must all be involved. She recalls the Jetsons, a cartoon where a human family lives surrounded by AI and technology. In this example, the family is unaltered by the surrounding AI. In reality, AI will not just change the world, it will change us. And as technology moves to the domain of enhancement, such as neural prosthetics and creating artificial hippocampi, we have to think about the future of the mind.

Evolution did the first wave of mind creation—now Schneider posits that businesses are leading the second wave. Will Facebook decide how that works? Again, all stakeholders must be involved.

In considering how the mind can be enhanced, one must consider consciousness.

Consciousness is essential to mindedness. And we must consider not only if AI can have consciousness but if it can have the felt quality, inner experience that humans have. Which intelligent systems are selves? Which are just conscious minds?

We must learn to identify conscious beings—human, non-human, as well as the combination of both. If non-conscious machines supplant biological intelligences, the singularity would be a nightmare. It would be the end of consciousness on Earth. It is not enough to only consider the rights of

machines—there are future of humanity issues at stake.

Dr. Schneider says that this issue reaches a cosmic level as there is a good chance that other species on other planets enhance their brains or have already been supplanted them with AI. More and more civilizations become more and more intelligent. We cannot control this flowering of consciousness on other planets—but we can perhaps do so on ours, with public dialogue.

There are two approaches we can take to achieve this public dialogue. The "wait-and-see" approach takes the middle of the road position. It concerns prospects for conscious machines but stops short of techno-optimism. The other approach is to adopt "techno-optimism," an idea that believes that AIs can be conscious based on the idea that brains are simply information processing systems and that all mental functions are computations.

Another worry is that consciousness will eventually be outmoded. It may be more efficient for superintelligent AI to eliminate consciousness. For all we know, the smartest AI systems may not be conscious. AI companies could also cheap out on consciousness. The properties that give rise to sophisticated systems information processing may not be the same processing that yields consciousness. For example, the AI that defeated humans at the game Go won using non-brain like functions. We should not assume that AI will need to have conscious processing.

We will need some provisional tests to know whether AI is conscious or not. Dr. Schneider outlined several examples of such tests, such as the ACT Test and the HAL 9000. However, these tests in and of themselves are not always appropriate and so a new test will need to be devised or multiple cross-checking tests used.

In light of mind enhancement and the nature of consciousness, Dr. Schneider says that soon, the unenhanced may no longer be the most intelligent beings on Earth.

HUMAN RIGHTS AND ARTIFICAL INTELLIGENCE: THE LONG (WORRISOME) VIEW?

Mathias Risse

Lucius N. Littauer Professor of Philosophy and Public Administration, HKS
Faculty Director, Carr Center for Human Rights Policy

Professor Mathias Risse opened his segment with a reading from German philosopher, Martin Heidegger, to prepare humanity for the human rights implications of major technological breakthroughs. He outlines several medium-term worries as well as longer-term worries.

In the medium-term, the worry is this: today, philosophers and activists treat rights as separate from distributive justice. Such an approach fails to make human rights a stable idea worth pursuing and as such, global distributive justice and human rights might lose out. It is critical to help strengthen global and distributive justice now so that we do not end up in a society that has a massive gap between the poor and the wealthy, as seen in the film *Elysium*.

We must seek to protect human rights in the medium term internationally because as technology improves, there will either be more inequality or not. For example, as technology grows increasingly capable, people will be replaced. The underclass might no longer be needed to support upperclasses and so the wealthy would transcend and the underprivileged would be hidden from view.

Focusing on human rights exclusively is neither philosophically plausible nor strategically advisable.

We need to get more serious about social and global distributive justice now or we will lose any serious prospects of having them realized in the long run and we also lose a plausible possibility to have human rights realized.

Risse warns that we are woefully unprepared philosophically to deal with situations such as this. How should we think about morality? How can we predict what a superintelligence would do? There would need to be different kinds of "moral statuses," running the gamut from simple ethical impact agents to full ethical agents, who have free will. What rights should be given to each class of ethical agents? Androids could also have adjustable independences and morality. Humans have a mind that machines would not have—does that give humans rights that machines would not have? Humans also have souls. Not many philosophers address this issue because it introduces so many more problems. Or is duality more physical, such as in emergentism where when matter is put together correctly, a mental body naturally occurs.

In conclusion, Risse states that there is a lack of agreement on moral foundations. We do not yet fully understand what moral comprehension is. In the medium term, technology will transform society. In the longer-term, machines must be given moral considerations.

Technology is not merely a set of tools but shapes the world we live in. Human rights will be under siege and be majorly shaped by global and societal distributive justice.

Lastly, technology must be used to advance human life rather than impoverishing it—and we must be prepared to deal with more, and multiple types of moral status for humans and machines.

ARTIFICIAL INTELLIGENCE AND HUMAN RIGHTS LAW IN THE HERE AND NOW

CAN YEGINSU

Barrister, 4 New Square Chambers, Columbia Law School

Yeginsu points out that the Universal Declaration of Human Rights is simply the starting point for international human rights law—it is not a legally binding instrument. It is the template and source for all human rights laws that are binding regionally, domestic, and internationally. However, unlike other areas of law, the human rights law themselves are very short. Tax codes are often 500 pages or more and the Universal Declaration was simply one page. This approach was intentional so as to make the Declaration more general and give judges interpreting rights flexibility so that human rights could remain alive in the advent of technology. Very few of these rights are absolute—most are qualified. Individual rights need to be balanced with other rights, such as public health and national security.

Human rights law is really a talk on balancing. We must ask ourselves how does one achieve a fair balance? This used to be the preserve of justices, but it is now a much wider question. How does the language of the Universal Declaration of Human Rights vary in regional and international communities? There is not just one body of law interpreting the document, says Yeginsu.

NANI JANSEN RAVENTFLOW

Director, Digital Freedom Fund

Being asked if the rights from the Universal Declaration have stood the test of time, Raventflow answers yes. Because it was drafted as a high level, technology independent document, she says that the Universal Declaration can still be applicable to a more technological society. So what will international human rights law look like in the

advent of technology? Raventflow says that it must be exercisable in the digital sphere and people must have the rights to understand tech, to opt out of profiling and request algorithm overwrite, and not to be judged by a machine. But are these new rights? Or specifications of rights in the framework already? Are the original rights flexible? These are all critical questions that we must ask today.

AMY LEHR

Director of Human Rights Initiative, Center for Strategic and International Studies (CSIS)

Amy Lehr says that the short-term human rights issues are really urgent. AI is used in surveillance and threatens our ability to even consider longer-term things. The UN Guiding Principles (UNGP) can be useful.

They were developed as a mapping exercise with Harvard professor John Ruggie. Human rights law applied to cities and states but because of globalization, companies needed to do more. Corporate Responsibility was created to respect human rights—companies should not infringe on human rights. This is a floor, not a ceiling. The UNGP called for companies to have some process to conduct human rights due diligence and impact assessments. This has been very impactful because companies are using the UNGP. It makes it very practical for businesses to know where to start and provides a good baseline for companies.

However, there are weaknesses large companies must think about. Large companies are developing processes to deal with human rights abuses, but they are often buying from companies who do not have those processes. How do we incentivize those companies to comply and behave ethically? Going further, when are companies responsible for remedying or walking away from a country that is not properly human rights compliant? When it comes to using AI, how can you know if you have a right to remedy if you don't even know if you were violated? The UNGP offers a good start but corporations still have a lot more to do.

MICHAEL KARAMIAN

Human Rights Manager, Microsoft

Microsoft is approaching their own way of protecting human rights, says Karamian. The day to day responsibilities in the human rights policy team at Microsoft include identifying who is most at risk in accessibility, freedom of expression, online safety, data privacy, and terrorist content online. Microsoft is also asking how it can counterbalance rights online. There is human rights related due diligence across the whole business, from managing risk to assessing risk, to tracking implementation and communicating what happens after. Most corporations only assess risk but not much more than that—a major weakness. Online tech users number in the billions. How do you engage with the users in human rights due diligence? The salient human rights issues are life and security and the future of work. Microsoft is not the only big company working hard on human rights due diligence. In the industry, there is great respect for companies doing so. Now we must learn how can companies collaborate and repeat the processes to coordinate better on human rights protection.

CONTEXTUALIZING CALLS FOR AI TRANSPARENCY AND FAIRNESS

JOHN BASL
Assistant Professor of Philosophy, Northeastern University

RONALD SANDLER
Chair and Professor of Philosophy, Director, Ethics Institute, Northeastern University

Basl and Sandler address the last mile problem in AI ethics, arguing for committee-based oversight to contextualize fairness and transparency. It will be a lot of work to translate what philosophers have done into something usable and practical. To do so, Basl and Sandler highlight several distinct kinds of fairness and transparency in AI systems, highlight the context/domain specificity of how those questions are answered, cover the difficulty in answering those questions within those contexts/domains, and why committee-based oversight is a promising tool to address these difficulties.

First, it is important to understand the various kinds of fairness in AI. Fairness in outcome depends on application or use, such as in using a predictive system (example: auto lending making existing inequalities worse). Operational fairness tackles how you put a certain chosen fairness into an algorithm. How fairness relates to and trades off with other values must also be considered. There is, for example, a low tolerance for bias in comparing recidivism predictions with medical health applications where predictive accuracy is more important. We must also ask if fairness can even be achieved by an algorithmic system. Not all problems can be solved by simple deliberation.

When it comes to transparency, there are multiple types of explanations that can be given by humans and AI systems. For example, there is a case base, where you could be matched to similar cases. In a sensitivity-based explanation, you would have cheaper insurance if you drove 10% less at night. It

is important to know whether explanations from AI systems match what humans give. If so, then this is a single standard view—a double standard is when an AI system has a different standard. Even if we adopt the single standard view, it would not be plausible because standards are still context dependent on which kind of explanation is most appropriate.

To contextualize fairness and transparency, committee-based oversight is a very promising tool. Strict compliance models are too restrictive or too lax. In situ ethics models where ethics personnel are embedded in technology companies is also promising. But oversight committees offer the most virtue. There are numerous extant problems to draw upon from the Institutional Review Board (IRB), the Institutional Animal Care and Use Committee (IACUC) and the Embryonic Stem Cell Research Committee (ESRO). Oversight committees are also well-suited to rapidly changing technological landscapes and are adaptable to a wide range of ethical challenges. They are also more democratic and tend to avoid conflicts of interests.

When it comes to AI, the ethical values espoused must be a respect for persons, justice, well-being, democracy, and moral status. To protect these, the oversight committee would be composed of technical, ethical, domain-specific, legal expertise as well as a public representative and a non-conflicted

"When it comes to AI, the ethical values espoused must be a respect for persons, justice, well-being, democracy, and moral status."

participant. Oversight committees would have an object of assessment and assessment tools such as protocols and proposals. They would have the power to prevent or halt the development of technologies, recall applications and revoke access to data, public disclosure, and to give ethics certifications. Institutionally, these committees can be positioned internally to a company or externally. External review boards would allow smaller companies to have resources like domain experts and ethicists.

Many additional challenges exist, particularly in funding and developing the capacity to implement these committee-based oversight teams. But as technology improves rapidly and algorithmic systems raise a host of domain/context specific ethical challenges, some tools must be adopted to protect people and maintain ethicality in the last mile.

RESPONSIBLE BUSINESS IN THE DIGITAL ECONOMY: WHAT'S NEW?

NIEN-HÊ HSIEH

Professor, Harvard Business School

Professor Nien-hê Hsieh asks critical ethical questions of business in light of the new digital economy and the advent of AI. First, he asks what changes by focusing on business. He argues that we are still motivated by greed—would bringing business into the problem make the ethical questions a lot worse? We must also consider what is technically possible (what are the harms and concerns that might arise?) and what is profitable. There is significant overlap but mutually exclusive areas exist.

So, what counts as responsible business? Can business and ethics be resolved or are they an oxymoronic null set? Milton Friedman says that there is one and only one social responsibility of business: to use its resources and engage in activities designed to increase its profits so long as it stays within the rules without fraud. Today, there are so many new views on what counts as responsible business: social enterprise conferences, creating shared value, conscious capitalism, social impact investing, stakeholder theory, and certified B corporations. Courts will uphold directors' decisions if those decisions are informed, untainted by conflicts, and taken in good faith that they are in the best interests of the corporation and its shareholders. This can be interpreted that increasing shareholder value is the priority but this is not required by law. Professor Hsieh suggests that the Milton Friedman view is not a bad place to start, it just needs a few additions. Former Dean of Harvard Business School (1908-1919), Edwin Francis Gay, says that "business is the activity of making things to sell at a profit—decently." Business ethics is how and how much we should profit. Is our profit reasonable? Are we earning them properly?

This can get presumptuous, so it is important to start with the basics. First, we emphasize the basic constraints that managers out to observe from the perspective of ordinary or everyday morality—'first, do no harm.' Then to undertake the basics of commercial activity in a way that is societally valuable, we must recognize that business is about doing well and doing good. Profit does not necessarily have to completely come from social value—such a view is problematic because

it doesn't recognize the societal value of profits. If social enterprises focus on mission statements, triple bottom lines, etc., then many other businesses can get off the hook. Reliability is not enough for trust--businesses need to have goodwill.

So, what changes in the digital economy? The question we must grapple with in business is if the risks taken in business are really unforeseen and unintended—or unforeseen but not part of the plan. Responsible business will have a real understanding of key technological developments, business opportunities and associate risks, will articulate the societal value proposition of the business opportunities, develop culture and systems in organizations to anticipate, identify, and avoid harms, and acknowledge the need to innovate the business model as much as the technology.

What is next for business ethics in light of technology? We don't have the 50 years that Nike had to improve business ethics given the scale, scope, and speed of these technological developments.

At business schools, we must develop cases that help understand technology opportunities and associated risks. There is also a lot of innovation that can be done on the business model level, but we are too focused on technological innovation.

CONCLUSION

Throughout this conference, the Carr Center for Human Rights Policy, in collaboration with the Edmond J. Safra Center for Ethics and the Berkman Klein Center for Internet and Society, together reflected on the past, present and future of human rights in light of new technologies and the development of artificial intelligence mechanisms. New technology such as AI will have deep lasting impacts on the future of human rights and ethical norms. In an effort to understand the future of rights, the Carr Center at the Harvard Kennedy School convened this conference to open a discussion between various experts and to facilitate this timely dialogue.

The discussion around artificial intelligence and technology affects and will affect all disciplines. By gathering experts from multidisciplinary fields such as business, policy, law, philosophy, computational mathematics, physics, and computer science, we learned that the assurance of institutional protection of distinctly human life is incredibly complicated in this era of innovation and brings forth many hard questions. We must also all collaborate together in an interdisciplinary and multi-disciplinary fashion to understand how to use or not use AI appropriately to protect human rights. New technology and AI are not theoretical - they are the future we must race to keep up with, should we want to harness their awesome power to ensure, not deny, human rights.

While some of the work posited in the conference was theoretical, i.e., 'Do we need to create an updated Universal Declaration of Human Rights that incorporates new types of sentient beings?' much of what was discussed has widespread implications on human rights today. Prison systems incorporate both human error and machine learning to determine outcomes for prisoners. Widespread use of data can affect everything ranging from one's credit, healthcare access, and voting ability. To protect humans, we must learn to develop and design ethical algorithms to avoid the perpetuation of bias and alternative facts. Regarding the future of work, we must also learn to harness intelligent systems to augment human work as we enter this shared reality with technology. Our learning institutions, particularly business schools and technology courses, must also embed ethical considerations into their curriculums.

Life ahead is a shared reality between humans and technology. AI has proven to be incredibly powerful and oftentimes more accurate than humans, but it is not without its limitations. Into the next 70 years, the world must discuss how we can define fairness, integrate ethics into our technology, understand explanatory problems of AI, seek transparency into artificial decision-making, and be empowered by AI. We must, as academics and practitioners but also as humans, grapple with these complex and vibrant questions now, as we seek to determine the norms for rights in the upcoming generations. We must move conversations between faculties forward to understand problems in the short-, medium-, and long-term, and engage with innovation while considering a broad range of ethics relating to how we as humans use and share our lives with such rapidly advancing technology.

As the Carr Center looks forward, our Artificial Intelligence and Technology Program will work to build bridges at both Harvard and within the AI & Tech community at large, serving as a convener and leader on these timely questions.

More information at:
www.carrcenter.hks.harvard.edu

HUMAN RIGHTS, ETHICS & ARTIFICIAL INTELLIGENCE

Challenges for the Next 70 Years of the Universal Declaration of Human Rights

CONFERENCE REPORT
January 2018

HARVARD Kennedy School
CARR CENTER
for Human Rights Policy