ARTIFICIAL INTELLIGENCE, THE GREAT DISRUPTOR

Coming to Terms with AI-Driven Markets, Governance and Life

David Bollier

A Report on the Second Annual Aspen Institute
Roundtable on Artificial Intelligence

THE ASPEN INSTITUTE

Communications and Society Program
Charles M. Firestone
Executive Director
Washington, D.C.
2018

The Aspen Institute
2300 N Street, NW
Suite 700
Washington, DC 20037

Published in the United States of America in 2018
by The Aspen Institute

Printed in the United States of America

18/003

Contents

Foreword

In 2017, artificially intelligent (AI) technologies surged into the popular discourse for its advancements — such as autonomous vehicles and predictive analytics — to critiques of potential biases, inequity and need for transparency. Growth in dataset sizes, increased computing efficiency and enhanced techniques in neural networks and machine learning are all factors in the success and pervasiveness of AI systems. These very same factors, however, also contribute to its limitations and biggest challenges. The prevailing unknown remains: to what extent do AI technologies affect our society's basic institutions, daily practices and cultural norms?

For the second year, the Aspen Institute Communications and Society Program convened thirty thought leaders from across disciplines and various fields to question the many ways in which AI may or may not impact the world. Participants of the Roundtable on Artificial Intelligence, held in August 2017, challenged the powerful narrative of AI's growing dominance and inevitable influence on today's society. To facilitate this thinking, the Roundtable encouraged participants to re-situate the human as the focal point, asking the question: *In what ways do AI innovations enhance and or limit personal human autonomy?* Discussions traversed the philosophical to the applied, and raised deep, fundamental questions on how to guide the trajectory of AI "with meaning and dignity for humans."

The following report, "Artificial Intelligence, The Great Disruptor: Coming to Terms with AI-Driven Markets, Governance and Life," authored by David Bollier, brings to light various nuanced issues in assessing tradeoffs between the benefits of AI and the potential negative social consequences.

The report is divided into four sections, which reflect the key themes roundtable participants addressed. First, "The Far-Reaching Disruptions of AI," highlights key technological trends across nine industry verticals and calls into question the techno-deterministic narrative so widely ascribed to AI systems. Second, "The Co-Evolution of AI and Humanity," reflects on four historical critiques of technology and begins to unpack the human and machine relationship.

In the third section, "The Perils of Predictive Analytics in Criminal Justice," Bollier provides a concrete example of the complications

embedded in AI systems, such as automatic sentencing programs and predictive policing. As several participants voiced, a move towards eliminating human discretion to an automated justice fails to account for structural inequalities and obfuscates accountability. The section also cites several citizen-data projects that aim to neutralize biases in data and algorithms.

Finally, in the fourth section, "AI Governance and Its Future," the report outlines the shortcomings of AI governance models in both the U.S. and around the world, leaving open questions on how best to evaluate AI innovations according to pre-existing standards, and under whose authority. Recommendations on developing a "soft law" versus strict regulation may help keep pace with the speed of change in AI. Most notably, the discussion on AI governance by design offers a starting framework to assure accountability for AI that maps onto the very basic questions of *who, what* and *when.*

At the end, the conference confirmed that the impact of AI technologies will be measured not only by units of its adoption but by the breadth of human success and failure. We hope the report provides the humanistic lens through which industry, government, academia, civil society and citizens can better address the future of the interaction between human and machine.

Acknowledgments

ARTIFICIAL INTELLIGENCE, THE GREAT DISRUPTOR

David Bollier

ARTIFICIAL INTELLIGENCE, THE GREAT DISRUPTOR

Coming to Terms with AI-Driven Markets, Governance and Life

David Bollier

Over the course of seventy years, efforts to develop artificial intelligence (AI) have encountered no fewer than four "AI winters" in which confidence in the technologies virtually collapsed. Over the past five years, however, various forms of AI have surged ahead with astonishing speed — multiple initiatives to build autonomous cars, factory automation, predictive analytics for human behavior, personalized marketing, algorithmic trading, civil infrastructure for "smart cities," medical diagnostic techniques, supply chain logistics and scores of other applications.

The power, scope and resilience of the current AI boom has persuaded many observers that society may be approaching an inflection point in history. As machines learn to learn — supported by copious, affordable computer memory, storage, connectivity, datasets and related technologies — AI is experiencing a powerful renaissance. The economics are so compelling, the technology so powerful, and the applications so diverse, that many observers now regard artificial intelligence as an inevitable, profoundly transformational force.

And yet this future is arriving with no small measure of uncertainty, trepidation and resistance. Even in their nascent stages of development, AI technologies are likely to radically reshape most sectors of the economy, including transport, energy, healthcare, retail and beyond. AI also raises new questions for government in rethinking economic policy, trade, national security and the future of work. AI technologies are likely to affect the basic institutions of democracy and government, everyday social practices and culture, and our deepest sense of what humanity is and should be.

To take stock of the many ways in which AI may change the world or not, the Aspen Institute Communications and Society Program convened thirty leading technologists, industry executives, social scientists, policymakers, public-interest advocates, and others, at the second annual Roundtable on Artificial Intelligence. The event, from August 6-8, 2017, in Aspen, Colorado, was moderated by Charles M. Firestone, Executive Director of the Communications and Society Program. This report, by rapporteur David Bollier, is an interpretive summary of the most significant themes discussed at the gathering.

The Far-Reaching Disruptions of AI

In an opening presentation, Wendell Wallach, an author and scholar at Yale University's Interdisciplinary Center for Bioethics and senior advisor to The Hastings Center, spoke about emerging AI technologies as "a fourth industrial revolution." Because of the broad scope of AI applications and their capacity to "amplify everything else," Wallach believes that we may be "at a major inflection point in history." He noted that AI has enormous capacities to shape and mold human behavior, and perhaps every segment of life.

> **. . . the self-driving car [is] "an apt metaphor for what we're dealing with — technology is moving into the driver's seat as a primary determinant of humanity's destiny."** – *Wendall Wallach*

This potential is exciting but also fraught with great risks, he said, because "the rapid pace of technological innovation and scientific discovery associated with AI is increasing the pressure on us to respond, often with little or no capacity for reflection." Wallach regards the self-driving car as "an apt metaphor for what we're dealing with — technology is moving into the driver's seat as a primary determinant of humanity's destiny. We are being challenged as to whether we can shape the trajectory of that future to some degree, with relatively weak tools."

Wallach said that the rise of AI in its current forms raises serious questions about some fundamental principles of the Enlightenment, such as the sovereignty of human rationality and the role of individualism as the foundation of governance. He noted, for example, that the fields of behavioral economics and evolutionary psychology "are revealing that humans are not rational agents, and that we are prone to systematic errors and biases. Furthermore, our behavior can be highly determined and easily manipulated, which suggests that we as individuals have very weak will." As AI facilitates new forms of "weaponized narratives" and propaganda via social media, said Wallach, "we are seeing major assaults on Enlightenment traditions."

These general developments pose three major challenges, said Wallach: to evaluate AI innovations in terms of existing ethical criteria; to determine whether those criteria still apply; and to "nudge the trajectory of AI and indeed all emerging technologies toward a future with meaning and dignity for more humans." Society will soon be asked to consider what tradeoffs it is willing to make for the benefits of AI, and whether and how to mitigate the risks and negative social consequences.

What is Driving AI Innovation Today?

To learn more about what forces are propelling AI forward, Naveen Rao, Corporate Vice President and General Manager of the Artificial Intelligence Products Group at Intel Corporation, cites three primary, interrelated drivers of AI today: dataset sizes, Moore's Law, and demand. Datasets have vastly grown in size over the past twenty years, said Rao, as Moore's Law has enabled computers to process and store data more efficiently. He noted that computer hard drives in the late 1990s may have had 80 megabytes; now an inexpensive flash drive contains 32 or 64 gigabytes.

As for Moore's Law, Rao shared a chart showing the relentless climb in computing efficiency as "computational substrates" have shifted from mechanical systems and relay switches to the vacuum tube, transistor and integrated circuit. The result has been dramatic improvements in the number of computations per second as measured in constant dollars.

Thanks to these trends, AI technologies can increasingly outperform human beings in tasks that were previously thought to be beyond the

reach of machines. For example, in 2010, computers attempting to identify one image from among five drawn from ImageNet, a database of some 1.2 million images, failed 30% of the time — as opposed to a 5% failure rate for humans. But by 2012, computer error rates were down to 16%, a phenomenal improvement, thanks to "deep learning" techniques that enable a machine to learn from its errors. By 2015, computers were exceeding humans in tests to correctly identify images. "The time it takes for a neutral network machine-learning algorithm to train on a dataset has fallen precipitously," said Rao. This in turn is reducing the amount of time and expertise needed to use such systems. Twenty years ago, it took a major company like Yahoo! to serve 100 million users; today a startup with only a handful of people, such as Instagram, can do that.

AI is so explosive, said Rao, because "it really does apply everywhere. We're going to see it used across the board in the next five to ten years. To me, this moment actually feels very similar to the Internet twenty years ago." Based on Intel forecasts, Rao cited new examples of AI applications in nine industry verticals:

Consumer: Smart assistants, chatbots, search personalization, augmented reality, robots

Health: Enhanced diagnostics, drug discovery, patient care, research, sensory aids

Finance: Algorithmic trading, fraud detection, research, personal finance, risk mitigation

Retail: Support, experience, marketing, merchandising, loyalty, supply chain, security

Government: Defense, data insights, safety & security, resident engagement, smarter cities

Energy: Oil & gas exploration, smart grid, operational improvement, conservation

Transport: Autonomous cars, automated trucking, aerospace, shipping, search & rescue

Industrial: Factory automation, predictive maintenance, precision agriculture, field automation

Other: Advertising, education, gaming, professional & IT services, Telco/media, sports

As AI technologies improve, they will increase the efficiencies of scale for companies and reduce the costs of services. More people will be able to use a technology and fewer people will be needed for a given task, he said. For example, said Rao, AI will help automate healthcare processes that are currently costly, such as interpretations of an MRI scan. "That skill could be codified into an algorithm," he said. "Once that is done, the price will drop precipitously." Similarly, some of the "thought drudgery" associated with reviewing legal briefs and cases could be automated, said Rao, freeing up people from expensive routine tasks.

"AI is becoming the lens through which we view all data." – *Naveen Rao*

The exponential leaps in computing capacities are posing new challenges of their own, however, such as how to make sense of huge pools of data. "Our biggest computational problem today is actually data overload," said Rao. "We have too much data in the world that we actually don't know what to do with. If we froze the world today and gave 100 megabytes of data to every man, woman and child on the planet, it would take us thirty years to get through all that data. This problem is going to get 75 to 100 times worse in the next ten years as data-gathering capabilities get cheaper and better." This is why AI is so important today, said Rao: It addresses "the biggest computational problem that we face today, which is finding useful structure in data. AI is becoming the lens through which we view all data."

The most significant upshot of AI innovations is how they are changing interactions between people and data, and in turn, our larger society. There is little question that AI and humanity will need to co-evolve in the future, but how this should be negotiated and managed is an open question. There are also likely to be unintended consequences. We may thrill to the idea of AI systems helping us to filter information to suit personalized wants and needs, but belatedly discover that the same technologies can produce fake news, closed echo chambers of public opinion, and the erosion of a shared public reality. When human bodies are blended with biocompatible implants containing AI

capabilities — neuroprosthetics — and potentially even gene-modifications, difficult new complications arise.

Is an AI Juggernaut Inevitable?

In response to Rao's presentation, participants debated whether artificial intelligence would necessarily proceed in these general directions. For Reed Hundt, CEO of the Coalition for Green Capital, former Chairman of the Federal Communications Commission, and Intel board member, the computing trends outlined by Rao are "inevitable." "As computing architectures fundamentally change, the era of the general-purpose computer is over," said Hundt. "Specialized-purpose architectures that imitate the brain will start to populate the environment."

> **The most significant upshot of AI innovations is how they are changing interactions between people and data, and in turn, our larger society.**

Hundt noted that "while the aggregate volume of data right now is huge, the gathering of data in each of the industry verticals mentioned by Rao is only partially complete. I think it will be 100% complete in a really short period of time." The amassing of huge datasets subjected to AI analyses will be "fundamentally disruptive," he added. No matter what any individuals may want, all economies around the world are committed to improving productivity and creating wealth. AI will only intensify this trend, said Hundt. Rao, who also regards rapid AI growth as inevitable, sees it as "symbiotic" with humans, in the sense of "supporting positive, exponential human growth."

Several conference participants took issue with this vision of AI development, however. "The current lens on the technology and what it can do is pretty narrow," said Jean-François Gagné, Co-Founder and CEO of Element AI, the world's largest AI applied research lab. "It's fragile. It's limited. And there is a lot of danger that comes with that. We need to make sure that we have a smoother transition to more sophisticated systems. There are many gaps right now on all fronts. This is a moving horizon."

A confusing complication is that AI consists of many different technologies and functions, Gagné pointed out; it is not one, single phenomenon. When people talk about AI, are they referring to "augmented intelligence" to support humans in doing discrete tasks? Or automation that replaces human functions and jobs? Or an entirely new integrated layer of AI, an immersive reality for a workplace or social life?

Broad "inevitability narratives" about AI are not helpful in illuminating the challenges ahead, said Kate Crawford, Distinguished Research Professor at New York University and Principal Researcher at Microsoft Research. She warned that such perspectives ignore "the much richer, more complex history of AI" over the past two generations. "AI has gone through several distinct 'AI winters' where funding dried up. The field was disparaged for not producing the results that it had claimed it could produce. By telling these linear stories of inevitability rather than cyclical stories, we're losing a lot of important historical learning," Crawford noted. "We're in a big hype cycle, guys! It's lovely to talk about co-evolution of humans and machines, and brain implants and exponential human growth, but we're still a long way from this. Meanwhile, there are many things in the here and now that urgently need our attention." AI brain implants, to take one example, would be tremendously costly and end up creating a different type of class system in society, Crawford said. "What sort of work are we doing *in advance* to actually address these concerns?"

Other participants questioned the narratives of technological determinism, saying that they ignore any role for democratic or individual agency. "Where does consent come in?" asked Joy Buolamwini, Aspen Institute Guest Scholar and founder of the Algorithmic Justice League at MIT Media Lab. "Who is making the decisions? When does [AI] enhancement become entrapment?" In the same vein, John C. Havens, Executive Director at the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, questioned any talk about human/machine "symbiosis" "when people don't have access to their data." He explained, "It's not symbiotic co-evolution when, as a person, I can't go to a data owner and say, 'Please give me copies of my data so I can figure out what is there about me.'"

The Co-Evolution of AI and Humanity

Any consideration of AI technologies can benefit from reflecting on historical critiques of technology. Marc Rotenberg, President and Executive Director of the Electronic Privacy Information Center (EPIC), asked the group to consider the ideas of four big thinkers. He started with Gary Kasparov, the chess champion and activist who was a central figure in highly competitive human/machine chess competitions from 1996 to 2006. Kasparov had agreed to play the IBM Deep Blue computer program in a series of chess matches in 1996, and lost the first game, but won the match. The next year, he lost to Deep Blue. This "signaled a turning point" in human/machine chess, Rotenberg said, "because the world chess champion just stopped playing machines."

Kasparov was very upset about the triumph of Deep Blue, but twenty years later, he wrote *Deep Thinking*, a more measured, optimistic book about the experience. In that book, Kasparov wrote that while the disruptions caused by increasingly "intelligent" machines may be upsetting to humans (like himself) in the short term, "no matter how many people are worried about jobs, or the social structure, or killer machines, we can never go back. It's against human progress and against human nature. Once tasks can be done better (cheaper, faster, safer) by machines, humans will only ever do them again for recreation or during power outages."[1]

Rotenberg observed that Kasparov is basically optimistic about the future of AI, and personally supports the idea of augmented intelligence as the best path forward. Interestingly, in a later chess tournament on the website Playchess.com that mixed grandmasters, computer-assisted players known as "centaurs," and chess-playing computers, all four teams left in the quarter-finals were centaurs. The winner was a lower-ranked chess player who was a data scientist who understood how to work with computers.

Another landmark figure in the attempt to understand AI was the MIT computer scientist Joseph Weizenbaum, author of a 1976 book *Computer Power and Human Reason*. "Weizenbaum was trying to get people to think about what it is that separates man from machine," said Rotenberg. One key human attribute that computers simply don't have, said Weizenbaum, is *autonomy* — the capacity for passion, wisdom and independent desire, including the desire to preserve autonomy from intrusions by machines.

The French philosopher Jacques Ellul was less optimistic than Kasparov or Weizenbaum. Ellul published a hugely influential book in 1954 (published in English in 1964) called *The Technological Society,* which "described with tremendous insight what happens as we give over more human activities to *technique*, a term he used to refer to any complex of standardized means for attaining a predetermined result," said Rotenberg. For Ellul, technique is "the totality of methods, rationally arrived at and having absolute efficiency (at a given stage of development) in every field of human activity. Modern technology has become a total phenomenon for civilization, the defining force of a new social order in which efficiency is no longer an option but a necessity imposed on all human activity."[2]

Ellul foresaw the pervasive use of technique to optimize the outcomes of all social functions, including elections, and to start to control people's destiny. This is precisely what computers have increasingly sought to do — to gather and process vast amounts of information about individuals in order to "make decisions" that affect their lives. To counter this risk, Ellul, influenced by many European thinkers, argued for a new regime of accountability, transparency and fairness, which in fact have become foundational principles of modern privacy law. But given the power of technique, Ellul was ultimately "not very optimistic about our prospects in finding solutions," said Rotenberg.

A fourth major thinker with compelling insights is the German sociologist Max Weber, who is famous for his studies of modern bureaucracy, the rise of capitalism, and how social relations have changed as a result. Bureaucracy and capitalism elevated the "rational legal" construct as the essence of modern relationships, especially as played out in organizations. In contrast to premodern societies that were organized around traditional authority or charismatic leaders, modern societies seek the rationalization and routinization of human activity, said Weber.

This idea reaches its logical culmination with such techniques as the FAST computer program [Future Attribute Screening Technology] introduced by US Homeland Security in 2011. By compiling sufficient personal attributes about an individual, the FAST system purported to predict with some degree of probability the likelihood that someone would commit a crime. In this, FAST echoes the plotline of the dystopian

sci-fi film *Minority Report,* a Tom Cruise thriller about "a special police unit that is able to arrest murderers before they commit their crime."

Rotenberg believes that these four thinkers help us frame the primary questions we must ask: "How do we preserve autonomy (Weizenbaum) in a world of pervasive technique (Ellul) and continuing rationalization (Weber)? What distinguishes human beings from machines (Kasparov)? What makes us human? What does it mean to move to a point where machine intelligence, however we define it, exceeds human intelligence?" A memorable cautionary tale about such questions is the famous scene in the sci-fi film *2001: A Space Odyssey.* When the astronaut Dave commands the spaceship's computer, "Open the pod bay doors, HAL," to de-activate the AI system, the computer, replies: "I'm sorry, Dave. I'm afraid I can't do that."

Does AI Enhance or Diminish Human Beings?

A previous Aspen Institute report outlined the many benefits that AI is likely to bring in developing autonomous vehicles, improving healthcare, and introducing new reporting and analytic techniques to journalism.[3] In addition to offering new capabilities, AI is seen as automating work that is repetitive, arduous or dangerous while improving efficiencies and customizing goods and services. While acknowledging these many benefits, conference participants challenged some of these conventional ambitions and suggested that there are larger, deeper questions that need to be asked. As one computer scientist put it, "The real societal question is 'What's the end game? Is the ultimate goal consciousness in a machine?'"

The current utility function of AI systems, replied Jean-François Gagné of Element AI, "is to maximize the efficiency of a specific task — to maximize profit/efficiency. But replacing humans having full perspectives with tools that are narrow, but so much more efficient, is creating distortions and introducing tons of fragilities. Given the power of these tools, we now need to question what exactly it is we are shooting for. It cannot be as simple as just profits or efficiency. The power of these tools is now getting us to question what exactly it is we are shooting for." Father E. Salobir, a Roman Catholic priest and Founder and President of OPTIC, a network that promotes research and innovation

in the digital humanities, believes the motivations behind AI systems are critical, "That's my question: Who is designing and training the machine, and are those things in accordance with our values?"

> ## "Who is designing and training the machine, and are those things in accordance with our values?"
> *– Fr. Eric Salobir*

The pitting of machine intelligence against human capabilities sets up an invidious comparison that some find troubling. John C. Havens of the Institute of Electrical and Electronics Engineers (IEEE) believes that AI as currently cast "begins this larger narrative that positions humans as 'worse' than machines, or somehow deficient. There is a real risk that AI will cast humans as 'flawed' and 'in need of improvement,' which then puts discussions about humanity into a whole new paradigm," he said. "In a sense, humans have already 'lost' because we're saying that 'machines will be better than us.'" Havens argued that it would be better to see AI as *complementing* humans, and to avoid conceiving AI design "as if human beings are broken."

Joi Ito, Director of the MIT Media Lab, agreed with the earlier suggestion that perhaps AI systems really are disrupting our Enlightenment faith in the individual and rationality: "Most of our problems today are problems where having 'more' doesn't make them better. Throwing more resources at problems, as in rebuilding Europe and Japan after World War II, or improving productivity, is not necessarily the solution any more. We have run the course of 'more is better.'"

Although many computer scientists regard enhancing autonomy as an ideal goal for AI, Ito argued that autonomy is really illusory: "We are constantly involved in relationships with each other and the Earth, and machines mediate and rearrange those relationships. From a systems dynamics perspective, there is no such thing as 'autonomy.' We are embedded in a world that is a complex, self-adaptive system; we are not 'autonomous.'" Terrence Southern, Founder of Illuminate STEM and Global Lead Robotics and Automation Engineer for GE Global Research, expanded, "The technologies we're using are mimick-

ing social connectivism and simultaneously making us still feel more isolated from the world. We are reducing our social and emotional interdependence, ultimately reducing how we value each other.”

“We are constantly involved in relationships with each other and the Earth, and machines mediate and rearrange those relationships. From a systems dynamics perspective, there is no such thing as ‘autonomy.’ We are embedded in a world that is a complex, self-adaptive system; we are not ‘autonomous.’” *- Joi Ito*

For Ito, “This makes the basic question of ‘What is good?’ really interesting. What does ‘flourishing in nature’ means in this context? What does it mean to be happy?” Instead of focusing on traditional notions of “liberty,” “autonomy,” “control” and “growth,” Ito suggested that “these kinds of Western paradigms are kind of outdated.”

Given the scope and power of emerging AI innovations, there was broad agreement that AI will likely change our ideas about what it means to be human. But are current trends in AI development encouraging or alarming?

Naveen Rao of Intel’s AI products group has little doubt that “the notion of what it means to be human will change and evolve” as a result of new AI systems. He suggested that the changes wrought by smartphones today will simply be extended by neuroprosthetics and other AI systems in the future. Are these changes all that different?” The brain has always filtered out “noise” from our environment; AI will simply augment that function in the future, Rao predicted, to the extent that it may “literally be merged into our conscious experience at some point.”

Rao quickly added, “I don’t see that as a horrible thing. This process has been going on for a long time. I’m a neuroscientist, and I can tell you with a very high degree of certainty that your brain is not the same as it was ten years ago. We are different human beings today than we

were one hundred years ago, because of technology. Is it really such a horrible moral problem that machines change what it means to be human?"

> ## . . . "the notion of what it means to be human will change and evolve" as a result of new AI systems.
> ### – *Naveen Rao*

Paul Blase, Managing Partner of AI & Data Solutions at tronc, Inc., sees AI as a valuable tool to sift through large, diverse datasets to identify problems and model improvements to our collective social and economic systems. He cited a PricewaterhouseCoopers (PwC)[4] study that used advanced analytics to assess barriers preventing the advancement of women in India. By drawing on data about domestic violence, family structures, education, women in the workplace, and so on, the model yielded new, more holistic insights into what might be done at the policy level to yield greater benefits, than discrete interventions. Blase concluded, "AI can be a force for good by helping us assemble and model data that provides a more accurate representation of the way the world works to help solve complex problems like this."

The Hidden Biases Embedded in AI

Other participants raised yellow flags about the hidden biases sometimes embedded in AI, most notably the quest for greater efficiency and uniformity among human beings. "The pressure that will drive co-evolution of humans and AI as complex adaptive systems is efficiency," said Louis Rosenberg, Founder and CEO of Unanimous A.I. "If you're an AI system," he said, "the more uniform that humans are in the system, the better. AI would love to get rid of outliers and have us all be uniform. That is not a prescription for autonomy, or for what's best for humanity." Rosenberg also noted that AI routinely makes pivotal, undisclosed decisions about what is "noise" in datasets — meaning information that we can safely ignore. "Well, that noise might be important to us humans," said Rosenberg. "Yet AI will decide what is the 'thought drudgery' that we don't have to concern ourselves with."

Rosenberg argued that human societies, like any biological systems, thrive on a swarm intelligence based on a wide diversity of opinion. "If a group is too monolithic," he said, "then it loses something. It gets dumber. And so we have this tension, which is that the human part of the system benefits from diversity, but AI systems are intensifying uniformity and efficiency. As AI becomes smarter, it could make people dumber and humanity more uniform."

There may be a certain hubris in thinking that AI evaluations of data are more insightful and reliable than analog methods, noted J. Nathan Matias, a postdoctoral researcher at Princeton University and Aspen Institute Guest Scholar. Just as the behavioral economist Daniel Kahneman has shown that humans are not as "rational" as economists like to think they are, so AI "may be over-optimistic about its ability to influence and change behavior," said Matias.

He cited the case of Instagram adjusting its search algorithms in an attempt to reduce users' interest in self-harm. "Instagram had this great idea that if they made self-harm information harder to search for, maybe it would produce better mental health outcomes for Instagram users. But it turns out that people who support and organize around self-harm are part of a distinct culture, and they found ways to circumvent the search barriers," he said. When researchers came back four years later, they discovered that Instagram's changes to its algorithms had actually caused self-harm material to become *more popular* on Instagram. In other words, AI is no magic bullet; the law of unintended consequences still applies.

Applications of AI that presume cause-and-effect relationships reflect a simple-minded notion of what human beings are, said Wendell Wallach, the Yale University policy expert and ethicist, "We have a scientific model of what humans are right now that misses the point. I don't think we even have the science to begin to talk about who we are collectively." Wallach faults, among other things, the mind/body dualism inaugurated by Rene Descartes that persists to this day, and mechanistic worldviews about how the world works. While AI machines are making some tremendous advances in calculative rationality and efficiency, they are less capable of recognizing and respecting some core aspects of our human consciousness and behavior.

Joi Ito countered that "machines don't necessarily have to optimize for efficiency." He argued that AI systems can model complex,

self-adaptive systems, and in so doing, build systems that are resilient, adaptive and capable of healing themselves. Ito pointed out that neural cognitive science, which increasingly informs the design of many AI projects, "is not about the most firepower. It's about how do we create things that are more interestingly complex." While AI in the short run may indulge its "efficiency addiction" and drive for "economic growth," it is fully capable of moving in more positive directions, said Ito.

> **"One of the greatest risks of machine learning and automation is that we will leave huge segments of the population behind."**
> *– Douglas Frantz*

But this will require a greater respect for human agency and more appropriate AI/human interfaces, a.k.a "augmented intelligence" strategies, instead of AI infrastructures designed to minimize meaningful human agency. This raises the discomfiting question, Can AI be designed to accommodate human self-determination and diverse interests open-source style? Kate Crawford, Co-Founder of the AI Now Institute, Distinguished Research Professor at NYU, and Principal Researcher at Microsoft Research is doubtful. "One of the more pernicious myths that we need to explore is the idea that autonomy [in AI design] is evenly shared," she said. "There are very few people who can really create AI at scale. We also have very strong empirical evidence that these tools may be accelerating inequality. So when we use the term 'we', I want us to remember that there isn't a shared 'we' here. Instead, let's consider who is being included and who is not included."

Douglas Frantz, who was Deputy Secretary-General of the Organisation for Economic Co-operation and Development (OECD) at the time of this conference, agreed, "One of the greatest risks of machine learning and automation is that we will leave huge segments of the population behind." Frantz worries that the largest tech companies and nation-states, especially China and the US, could "turn everyone else into client states" because of their vastly disproportionate control over resources, talent, data, and machine-learning technology. Michael

Chui, Partner with the McKinsey Global Institute, reported that the latest McKinsey research shows that "of all experimental investment in AI, 66% is in the United States, 13% is in China, with the remainder split among many countries."

> **"At the moment, it looks like there will be an AI oligopoly that will be much more powerful than any oil company ever was."** – *Wendell Wallach*

"At the moment," said Wendell Wallach, "it looks like there will be an AI oligopoly that will be much more powerful than any oil company ever was. We are moving into a universe where multinationals may have unbelievable amounts of power, all of it entwined with how technology is deployed, the control of data, and the capture of profits from productivity gains. In previous administrations, the State Department actually had an 'ambassador to Silicon Valley' because the Government's relationships with companies there are more complex than its relationships with many states."

The Perils of Predictive Analytics in Criminal Justice

The unacknowledged biases of AI systems are perhaps most pointedly evident in the criminal justice system. Here AI is often used to make data-based generalizations about people, which are then used in policing on the street, criminal sentencing, parole decisions and much else.

It is all part of "the coded gaze," as Joy Buolamwini of the Algorithmic Justice League puts it. The coded gaze is about the power to create and use AI technologies to evaluate other people and make decisions about their fates. Structural inequalities such as historic racial, ethnic and gender prejudices, and social and wealth inequalities feed into this process, with the result that AI amplifies past inequities. For example, said Buolamwini, facial recognition software may rely on datasets that are easily accessible, which may mean that ethnic minorities with under-represented faces, may be excluded from the system in the first place.[5]

New York Times reporter Adam Litwak in 2017 asked of Chief Justice John Roberts, Jr., "Can you foresee a day when smart machines, driven

with artificial intelligences, will assist with courtroom fact-finding or, more controversially even, judicial decision-making?" Roberts' answer stunned everyone: "It's a day that's here, and it's putting a significant strain on how the judiciary goes about doing things."[6]

Court observers speculated that Roberts was talking about the case of a Wisconsin prisoner sentenced to six years in prison based in part on a software program produced by Northpointe, Inc., called COMPAS. A prosecutor used the program to persuade a trial judge that the prisoner showed "a high risk of violence, high risk of recidivism, high pretrial risk." Meanwhile, the news organization Pro Publica published an article about the COMPAS secret algorithms in 2016 that concluded that black defendants in Broward County, Florida "were far more likely than white defendants to be incorrectly judged to be at a higher rate of recidivism."[7]

When closed, proprietary algorithms become so deeply implicated in the justice system, "we move from a justice of consequences to a justice of correlation." *– Fr. Eric Salobir*

Joshua Browder, Founder and CEO of DoNotPay Robot Lawyer, a chat-bot lawyer that helps people with their legal issues, reported that "about 20% of all cases in the law now have automatic sentencing programs using AI technology that applies some sort of sentencing recommendation to the judge. There's been this huge backlash because none of the algorithms are open despite sentencing being part of a public process." (Browder had created his software chatbot to turn the tables on the formulaic and unfair nature of law: by enabling motorists to appeal parking tickets, over 175,000 succeeded in getting their tickets thrown out, saving them an estimated $5 million.)

For the most part, said one participant, judges are not unhappy to be able to shift responsibility for verdicts to AI systems and their supposedly more rigorous risk assessment scores. This has serious risks for the morality and legitimacy of the system, said Father Eric Salobir of OPTIC: When closed, proprietary algorithms become so deeply implicated in the justice system, "we move from a justice of consequences to

a justice of correlation," undermining the integrity of justice and the process. Explainability is a critical element of the judicial system.[8]

Secret algorithms rendering risk assessments are not confined to criminal sentencing. They are also being used to try to predict criminal activity, which in turn is influencing policing priorities. "More and more police departments are adopting AI systems to predict where crime will happen," said David Copps, Founder and CEO of Brainspace, a Dallas-based firm that does "investigative analytics." Copps was once shown a video about an AI system in which police arrested a "suspicious" person in a neighborhood that was supposedly more crime-prone. It turned out that there was an arrest warrant out on the person, enabling the police to make an arrest. Copps said: "All I could think is that that person didn't do anything. He had not committed a crime." Even though predictions may be unreliable, said Copps, the danger is that "intention creates reality. It almost creates a false-positive."

Kate Crawford cited a multi-year RAND study of predictive policing in Chicago. "It showed that the system was completely ineffective at reducing crime," she said. "Absolutely net zero. It did have one significant net impact, however — increasing the harassment of people on the 'heat list', according to RAND." Crawford also cited a research paper about an "automatic criminality detector" that purports to use machine-learning to discriminate between criminals and non-criminals based on their assessment of people's faces. The system analyzed the faces of 1,586 real persons, nearly half of whom were convicted criminals, and concluded that the AI could detect "structural features for predicting criminality, such as lip curvature, eye inner corner distance, and the so-called nose-mouth angle."[9]

Crawford said that the authors of the piece, responding to great criticism, claimed that any resemblance to the [discredited] use of physiognomy and phrenology on their part is purely accidental because, in their view, machine learning is neutral. This is one of Crawford's central concerns about AI technologies: they are often portrayed as "neutral tools." Some users presume that by pushing lots of (potentially skewed) data through an opaque deep learning algorithm, the results are somehow neutral and reliable. "We need to think a lot more about what design specs are built into a system, and who has the power

to shape those specifications," Crawford said. Commissioner Terrell McSweeny of the Federal Trade Commission (speaking for herself and not the FTC) agreed, "These are completely opaque systems that are held by a very few powerful entities, and we have almost no access to determine when bias is even occurring. We are relying completely on companies' own internal testing and control. One must question whether that is sustainable or desirable."

Algorithmic Accountability

The defense of AI risk assessment and prediction systems is that they provide a more factual basis for decision-making and can therefore actually *reduce* biases and mistakes. "If AI is used to supplement human decisions in complex circumstances, shedding light on very, very complex situations with more variables than the human brain can hold, I think the technology opens to us the possibility for more fairness," said Jean-François Gagné of Element AI. He cited progress on Generative Adversarial Networks, or GANs, which attempt to extract and explain biases in algorithms in datasets.[10]

FTC Commissioner Terrell McSweeny is troubled by this kind of automated justice, however: "You are eliminating human accountability and the role that humans play in sitting in judgment on other human beings." That's what the mandatory minimum federal drug sentencing guidelines for judges attempted to do, but failed, she said. "We have lost an entire generation of mostly men of color to a system that was meant to be weeding out bias in individual judges," said McSweeny. "When you eliminate human discretion in decision making, you can potentially cause much greater problems."

Another factor at play here is our varying levels of tolerance for error, she said. With self-driving cars, there is likely to be a very low willingness for mistakes. This does not appear to be the case with AI-assisted systems for assessing job performance and meting out criminal sentences.

A print cartoon about this theme has a defendant asking a computer, "How can you judge me?" The computer replies, "You wouldn't understand." In that sardonic humor lies a deeper philosophical debate about the need for transparency and accountability in jurisprudence. "I

would argue that as we move into the AI realm, there is a new problem, the opacity of decisions," said Marc Rotenberg of the Electronic Privacy Information Center. "Data can be enormously helpful in helping us to extract bias or reveal bias, but when we embed a rule, we are encoding whatever normative value we think produces necessary outcomes."

This dynamic was dramatically on display in a federal case brought by the Houston Federation of Teachers Local 2415 against the Houston school district in April 2014. The teachers claimed that a proprietary algorithm that measures teacher performance based on student test scores could violate their civil rights because of a lack of due process to publicly evaluate evidence used to fire them. The system gave teachers a raw performance rating relative to the state average, with no recognition that some school districts may have larger populations, disadvantaged cohorts of students, or other relevant factors. In May 2017, the Fifth Circuit Court of Appeals ruled in *Houston Federation of Teachers et al. v. Houston Independent School District* that the teachers had legitimate arguments that the assessment program may violate their Fourteenth Amendment due process protections.[11]

There is a similar lack of transparency involving the use of body cameras that an AI company, Axon (formerly Taser), is giving away to police departments across the US. Axon has added real-time facial recognition software to the cameras, which means that the video feeds are being held by a private company whose data models are unavailable to public authorities. "Currently, there are no public means of accountability," said Kate Crawford, adding, "There should be a much higher bar for how these devices work, how they are being used, and how they are being audited. Ultimately, we need a commitment to forms of due process when high-stakes decision-making is involved, such as criminal justice, welfare and education."

Now AI systems are being built into the inner workings of government itself, said Crawford, referring to Palantir, a company that has a large-scale contract with the Trump administration to build a new machine-learning platform for the US Immigration and Customs Enforcement agency. "What we've learned about this platform is that it has enormously sensitive data. It is bringing together datasets from many different databases that have not been combined before. It could indeed be an engine for very large levels of deportation in this country.

The question that I would put to us is, What sort of procedural due process rights would anyone have against a system like this?"

Crawford recommended reading Hannah Arendt's classic essay, *The Origins of Totalitarianism* (1951), which she sees as presaging the surveillance-and-control ambitions of the Palantir platform. Arendt writes:

> Now the police dreams that one would look at a gigantic map on the office wall, and that should suffice at any moment to establish who is related to whom, and in what degree of intimacy. This dream is not unrealizable, although its technical execution is difficult. If this map really did exist, not even memory would stand in the way of the totalitarian claim to domination. Such a map might make it possible to obliterate people without any trace, as if they had never existed at all.

As Crawford concluded: "Arendt's fear was that the reason totalitarianism failed in the 20th century was because it simply didn't have access to sufficiently powerful technologies."

Can Insurgent Data Systems Neutralize Bias?

If powerful players can use algorithmic bias to affect or control people's lives, the question arises: Can countervailing AI or data systems be built to act as correctives? Could open-source artificial intelligence neutralize or overcome AI bias by providing the transparency and diversity of perspectives that are needed?

J. Nathan Matias, a postdoctoral researcher at Princeton and Aspen Institute Guest Scholar believes that "there is an opportunity to use databases and AI systems to rethink, overturn or optimize an unjust system." He cited projects by online feminist groups that have amassed data to protect themselves against online harassment, for example, and informal citizen communities that are using machine-learning systems to try to assure due process in court settings. Some intriguing innovations in citizen-based data projects include:

- The Algorithmic Justice League, founded by Joy Buolamwini, a student at MIT Media Lab, is a collective that is dedicated to "highlighting algorithmic bias through media, art and sci-

ence; providing space for people to voice their concerns and experiences with the 'coded gaze'; and developing practices for accountability during the design, development and deployment of coded systems." The project has sponsored research initiatives, reports, videos and art-driven protests, among other strategies to draw attention to the "coded gaze."

- Three enterprising data activists created an app called White Collar Crime Risk Zones[12] that "uses machine learning to predict where financial crimes are most likely to occur throughout the US." Upon entering a zip code into the search box, the app produces a map showing pushpins indicating where a documented crime occurred, the approximate "crime severity" (in US dollars), and the "top risk likelihoods" in that area (such as "breach of fiduciary duty" and "employment discrimination based on age").

- A project at the MIT Media Lab is building a new sort of human-machine interface that explicitly tries to identify and model bias as a way to root it out, said Joi Ito. The interface does this by interacting with a person to identify and highlight his or her own biases while also revealing how the machine itself is making judgments. The idea is to make visible the algorithmic assumptions of the machine and the biases of the human decision maker. The tool itself is positioned as augmented intelligence, not a substitute for human decision-making.

- Gliimpse is a personal health data platform that lets people aggregate their health data from dozens (or more) data sources, and then collect and personalize the use of the data. A startup acquired by Apple in 2017, Gliimpse helps individuals take charge of their data in assessing their personal health, making healthcare decisions and sharing the data with trusted third parties.

Is Open Source AI Possible?

If the goal is to increase transparency and control of data for users, one of the most obvious strategic approaches is open source development. But is it really possible to create an "open AI" and could it be effective?

Naveen Rao of Intel said that "AI as a research and development field is actually one of the most open ones around," noting that methodologies and code are generally published and not held as trade secrets. However, this may not be significant enough, replied Joy Buolamwini of MIT Media Lab, "because these are data-centric technologies. If you have access to the models and learn the predictions, but you don't have access to the data, you're missing half of the picture. The data itself needs to be part of any process to increase transparency within AI."

For Jean-François Gagné of Element AI, the idea of open AI "kind of misses the point. When you think about it, AI is transforming the way we code, so we no longer prescribe and encapsulate insights into the code. We now build models that are tools, and then train them. So providing access to methods just doesn't move the needle at all — and it's not a matter of not having access to datasets. It's much broader than that, much more fundamental," he stressed.

What makes AI different from open source software, Gagné said, is that AI tools using enormous pools of data can self-learn and improve their methodologies faster and better than anyone else. Eventually, the tools will "reach a point of escape velocity" that insulate the big AI players from competition and AI tools themselves from scrutiny. "The volume of search queries and other data that Google can access is so big that its algorithms can perform way better than others," said Gagné. The AI tools controlled by the big players are becoming more powerful for another reason, he said, they can convert what used to be considered "noise" — meaningless data-points buried within unfathomably vast datasets — into useable information. Meanwhile, said Gagné, "the social contract" that purportedly applies to this transfer of data is a relic from another time.

Kate Crawford argued that the problems go beyond access to AI methodologies and data, to the ability to access "infrastructure of scale." Only a few companies — like Amazon Web Services, Google, Baidu and Microsoft — currently have the types of massively parallel

computing infrastructures to produce the computations at competitive speeds and efficiencies, she said. "It's a matter of data plus infrastructure plus large amounts of capital," said Crawford. "There is already a profound concentration of power in the AI industry that they have left behind large swaths of the world. We are really talking about a tiny group of global players doing this."

Jean-Francois Gagné added, "Startups are being totally crushed by a Google or Amazon, leveraging an absolutely unfair amount of datasets and algorithms to lower the price-point and be so much more efficient than anyone else, not to mention [saving money by] reporting tax in different countries. As we think about AI, these phenomena are going to get amplified. For a lot of countries that are out of this race, this is going to be a huge issue," he predicted. "It is almost literally impossible to compete."[13]

AI Governance and Its Future

As it becomes clear that AI could "change every facet of life," as Wendell Wallach put it, it becomes equally clear that focused, intelligent forms of governance are needed to address the disruptive economic, social and political impacts. This is a daunting challenge not only because the questions are so large and complex, but because the vehicles for political choice and thoughtful policymaking are so fragmented and inadequate.

"If we had to write down the five rules that we want some government agency to impose on, say, Internet of Things networks," said Reed Hundt, the former FCC Commissioner, "we wouldn't even know what to say right now." There is a void in government in terms of addressing AI governance, whether it is automation, robotics, medical diagnostics or consumer marketing. As AI technologies commoditize human judgment and intelligence in dozens of jobs and even in white-collar professions, Hundt concluded that "there is no future for work as we know it today; only a future for a different kind of work. But there is basically no transition plan in the world of politics at all."

The void in AI governance may have many explanations: the sheer speed of AI innovation and the uncertain pathways it will take; the disruptive and complicated ramifications that elected officials might rather avoid; the chronic difficulty in coordinating diverse laws and

federal agencies; and general industry resistance to the very idea of government regulation.

"The political structure of the United States is not well-suited to the social and technological challenges posed by AI," said Hundt dryly. Wendell Wallach added, "Industry doesn't want to be regulated because it feels government doesn't know how to regulate it. Yet it wants the public to see it as being responsible. Ultimately, there is a need to demand that industry take some responsibility here." The government, for its part, has a keen interest in developing some sort of AI governance, if only to track the international implications of AI for national security, cyberwarfare and progress toward Sustainable Development Goals.

> **The three basic challenges that any governance regime must meet are to evaluate AI innovations according to ethical criteria, determine if these ethical standards remain relevant, and use governance to "nudge AI toward a better path"….**
> *– Wendell Wallach*

A primary task, many conference participants agreed, is to figure out the right institutional structures and terms of governance for AI. What sort of institutions are needed, and which could be effective? How can the benefits of AI be maximized while the social harms mitigated or stopped?

The three basic challenges that any governance regime must meet, said Wallach, are to evaluate AI innovations according to ethical criteria, determine if these ethical standards remain relevant, and use governance to "nudge AI toward a better path" that promotes the benefits and mitigates risks. Wallach added that conversations about these topics will inescapably raise deep philosophical questions about individual autonomy, collective action and our vision of humanity.

Participants identified a number of other questions that must be asked: What unit of governance is appropriate? Should policy be

driven by regulations, best practices or consensus norms? Should the authority to intervene be based on existing laws or are new laws needed? Should the focus be on individual AI sectors or on certain types of machines and capabilities?

Kate Crawford of Microsoft Research believes that, given the way that AI generally works, governance should be focused on specific industry sectors — workplaces, healthcare, retail, etc. — which would allow policy to get very specific. She also believes that there are many laws on the books that could be appropriately extended to cover AI systems. Part of the challenge would be to harmonize the different statutory regimes. However, Joi Ito of the MIT Media Lab believes that "there's something fundamentally different about AI that requires new laws." He based this judgment on his belated recognition that new types of cyberlaw should have been enacted in the 1990s to take account of the special character of the World Wide Web.

Based on his experience in developing effective technology policies, Marc Rotenberg of EPIC urged that regulation be focused "on data, not devices." This helps keep any regulations technology-neutral and thus more innovation-friendly. Rotenberg noted, too, that "the application of rights and responsibilities are necessarily asymmetrical" — that is, the parties who are most able to reduce risks should shoulder greater responsibilities, and those who are more vulnerable (usually, the unorganized public, consumers or workers) should have greater rights.

At a more refined level, participants raised questions about which instrumentalities might be best for governing AI and Big Data. Should they use "hard" statutory law and regulation, or "soft law" that attempts to promote certain best practices and norms in industry? Perhaps government structures could have a looser framework than strict regulation, much as the United Nations' Sustainable Development Goals have sought to spur new types of business investment and practices.

Because the speed of change in AI is so great, there is always a question whether governance can act in a timely fashion, or even in proactive, anticipatory ways. For many participants, this problem suggests that "we will need AI to control AI." Governance that "builds in" design features into AI itself is more likely to be timely, focused and effective.

Existing AI Governance Initiatives

It turns out that there are quite a few initiatives already underway to study the future of AI, its ethical and social implications, and potential governance approaches. But there is little coordination among these projects, or even much mutual awareness of the landscape of players.

In terms of the federal government, potential authority over AI is, as noted, diffuse or at least uncertain. However, there are two science policy bodies — the National Science and Technology Council, within the White House, and the American Academy for the Advancement of Science, that could play important roles. There is also the White House Office of American Innovation. In Congress, Senator Maria Cantwell of Washington in July 2017 proposed creating an AI committee within the US Department of Commerce, with a special focus on how automation will affect the workforce.[14]

At the international level, the Organisation for Economic Co-operation and Development (OECD) in 2017 launched a major two-year project called "Going Digital" that will attempt to sort out the impacts of the coming technological revolution.[15] The project, which spans ten OECD directorates, is focused around "jobs and skills, privacy, security, and how to ensure that technological changes benefit society as a whole, among others." Concerned that some countries and sectors of society may be left behind, the Going Digital project is also addressing how to "build a coherent and comprehensive policy approach" to help assure "stronger and more inclusive growth."

Douglas Frantz, the Deputy Secretary-General of OECD, said that 35 OECD-member countries will rely on this project to help guide their own policies going forward. Marc Rotenberg of EPIC said that he has high hopes that OECD will indeed produce a new consensus framework for AI accountability because the OECD did just that in the 1980s in developing a "light-touch" policy framework for privacy that is still used today.

Meanwhile, there are several independent academic and industry-sponsored projects exploring various ways to manage AI technologies. These range in focus from industry best practices and technical standards, to state-based policy principles and standards, to open-ended research to make sense of the many AI developments now unfolding. Some of the more notable research efforts include:

- The MIT Media Lab and Berkman Klein Center for Internet and Society at Harvard University in January 2017 embarked upon a new $27 million initiative (Ethics and Governance of Artificial Intelligence Fund) to "bridge the gap between the humanities, the social sciences, and computing by addressing the global challenges of artificial intelligence from a multi-disciplinary perspective."[16]

- The Institute of Electrical and Electronics Engineers (IEEE) has eleven standards-working groups right now that are dealing with various aspects of AI. They draw upon hundreds of engineering experts to develop consensus technical standards that are interoperable and practical. "These are the first suite of standards directly addressing AI ethical issues," said John C. Havens. "In effect, they serve as a kind of 'soft governance,' even if not all of them say 'AI and ethics' explicitly," he said.

- The organization, Partnership on AI, was established by Amazon.com, Google, Facebook, Microsoft, Apple, IBM and DeepMind, among others, "to study and formulate best practices on AI technologies" and promote public discussion and understanding of AI.[17]

- The International Telecommunications Union has a standing AI Group. It also hosted an "AI for Good Global Summit" in Geneva to explore the issues in June 2017.[18]

- The United Nations Economic and Social Council hosted an event in October 2017 on how AI could help "achieve economic growth and reduce inequalities."[19]

- Through a three-year project, "Control and Responsible Innovation in the Development of Autonomous Machines,"[20] law professor Gary E. Marchant and ethics scholar Wendell Wallach explored governance options for AI. They have proposed a Governance Coordination Committee (GCC) to try "to harmonize and integrate the various governance approaches that have been implemented or proposed."[21]

It is unclear how these various projects will evolve or what impact they will have, noted Tim Hwang, Director of Ethics and Governance at the Artificial Intelligence Fund and former Global Public Policy Lead, AI/ML at Google, as each participant "tends to lean very heavily in one direction or another," based on their particular perspectives. What is significant is that "everybody is kind of putting their chip down on the table."

AI Governance by Design?

A presentation by FTC Commissioner Terrell McSweeny explained the shifts in regulatory approaches at the FTC in recent decades, which could inform future regulatory approaches to AI. (Again, McSweeny was speaking for herself, and not necessarily for the FTC or any other commissioners.) Based on its general authority to regulate unfair and deceptive trade practices, the FTC has long relied on "rational choice theory" in its regulatory interventions. "The idea is that if individual consumers can acquire accurate information, they will make rational choices in the marketplace that will prod acceptable balances between individual and commercial interests," said McSweeny. Of course, this framework assumes that consumers have accurate information in a transparent context.

Another approach that the FTC has relied upon, especially in the 1990s with respect to online privacy, is a "notice and choice" framework. The Commission has seen itself as a "norms entrepreneur" to prod websites to post their privacy policies online, and then consumers can click "I agree," or decline. The weakness of this approach has been acknowledged, however, spurring the FTC to develop a "context model" that tries to get companies to offer appropriate set of choices and information about privacy settings, at a suitable time and in a usage context. Any data disclosure must be appropriate to reasonable consumer expectations regarding the usage context, so that, for example, geolocation data are not collected for, say, a flashlight app without a clear and timely opt-in choice being provided to the user.

About seven years ago, the FTC began advocating moving to a "privacy by design" approach that attempts to prod industry players to build in privacy or security features into their products at the outset.

McSweeny suggested that perhaps governance by design might be an effective approach toward the regulation of AI technologies. This approach would attempt to create governance frameworks to assure accountability for AI performance and decision-making. In creating a usable framework, stakeholders and policy makers would need to ask such questions as:

Who?

- o Where within organizations does accountability lie for performance of, or decisions made by, AI?
- o Should humans and/or their organizations be held accountable for actions taken by their AI?
- o Are there decisions that should remain human?
- o What is the appropriate role for government regulation vs. self-regulation?

What?

- o What are the key components for AI governance? What would reasonable governance by design include?
- o Is the concept of "compliance" sufficient?
- o What cultural norms, governance models and laws can we draw on to inform governance frameworks? What's different and requires specific response?

When?

- o Do we have the right knowledge-base to draw on for governance frameworks?
- o Are there sufficient incentives in the marketplace for adoption of governance frameworks, or is a stronger government response needed?
- o Even if we come up with the right frameworks, can they keep up with AI?

For McSweeny, the goal is to come up with the "Goldilocks zone" in which the intensity and scope of governance is "just right"—a "habitable zone of AI governance."

Participants noted specific challenges in devising effective forms of governance for AI. There is, first of all, a cross-disciplinary challenge in drawing insight from a wide array of scientific and academic disciplines. There is an analogous coordination problem in orchestrating many different government officials and agencies, not just at the federal level, but at the state and local level. What should be the role for governors and mayors, for example, and what role for existing federal agencies?

> **…the goal is to come up with the "Goldilocks zone" in which the intensity and scope of governance is "just right"—a "habitable zone of AI governance."** – *Terrell McSweeny*

At each level, there are likely to be gaps in technical expertise that are not easily filled, if only because so much AI research and development is at the cutting edge, leaving only a limited pool of expertise. In addition, as Alberto Ibargüen, President and CEO of the Knight Foundation noted, "The speed of innovation makes it mind-bogglingly difficult to deal with governance when we have institutions that are basically backward-looking. How to create new structures or systems for changes that happen tomorrow is another discussion entirely."

Another problem might be called the "ontological mismatch" of law and AI: Most existing laws are based on human intention, but the behaviors and impacts of machine-learning are likely to be unpredictable or unknowable as they evolve.

For all of these reasons, transparency in the process is key, said Marc Rotenberg. Following a series of privacy complaints that his group EPIC had brought against various search engines and websites, he concluded that "we cannot rely on companies' representations about what they have done." Rotenberg suggested that ultimately government must have the clear authority and willingness to "pull the plug" on AI projects that are incompatible with core societal values. This is not unthinkable, he said, citing Facebook's own cancellation of an AI project in which machines reportedly had developed a language that its overseers did not understand. Kate Crawford reported that regulators

at the AI Now symposium at MIT in 2017[22] specifically discussed the possible need for moratoria on "certain domains of algorithmic determinations until they can be shown to be far more fair than they are right now." The unresolved issue in such cases is who shall shoulder the burden of proof to show harm — those questioning its fairness or the owners of the AI technology?

A Future Research Agenda for AI

Based on the remarkable potential of new AI systems, it is clear that much more research needs to be done — about the technology itself and its human interfaces, but also of the economic, social, civic and political implications. This constitutes a rather sizeable frontier.

France A. Córdova, Director of the National Science Foundation, noted that in a review of ten big ideas for future investments, machine learning was important in each area — so there really is a very broad agenda for AI in the future. Several participants pointed to quantum computing as a field likely to yield the next fundamental advances in AI. Another field of interest is "artificial general intelligence," AGI, which focuses on how devices can think about what other devices are thinking.

An important focus for research is learning the limits of AI and how it blends with existing social institutions and dynamics. It is not always clear, for example, when the technology is reliable enough to replace humans. "Studies on deep neural nets in medical contexts raise serious concerns, as we've seen from Rich Caruana's research,"[23] said Kate Crawford. "We are not at a stage where we can know for sure why a model produced a particular result." Similar concerns seem to apply to autonomous vehicles and other deep neural nets in open social contexts.

Crawford warned that we need to be mindful of the errors retroactively discovered in technologies once thought to be utterly reliable, such as MRI scans (which had a software error). "We have a lot more work to do on the socio-technical, legal and fairness frameworks" before AI can take over, she said. We do not really have any metrics for understanding the social impacts of AI or how biases come to be embedded in AI, socially, contextually and technically. More thought is needed for how to implement accountability mechanisms, and who should oversee them. Jean-François Gagné of Element AI generally agreed, "The level of maturity of the technology and our understanding of it are still very, very, very low."

But Gagné added, uncertainties give us the time and opportunities to deal with fairness issues, which is a positive thing. If AI is developed as augmented intelligence and not a replacement for people, that also gives humans the chance to intervene in AI processes. Such an approach may be the better part of wisdom, in any case, because our understanding of user/AI interfaces remains very rudimentary if not naive.

Ruchir Puri, Chief Architect of IBM Watson, pointed out that many AI systems remain fairly rudimentary and fragile, "failing to recognize things that are obvious to human eyes. Just changing a couple of pixels in a photo of a school bus in the ImageNet database, for example, can make it register as a pizza. AI technology relies heavily on massive amounts of 'labeled data,'" he said. "This makes it prone to cyberattacks." Another major problem, Puri said, is the enormous increases of power needed by AI systems if they are to emulate the capabilities of the human brain (which runs on a meager 20 watts of electricity).

The challenges are not just technical, however, but ontological. "We are dealing with a poverty of understanding about human cognition and how effectively humans and machines will interface with each other," said Wendell Wallach. "AI systems will need something like emotions and emotional intelligence, a theory of mind, consciousness and other supra-rational faculties such as empathy in order to make appropriate decisions in morally significant situations. Furthermore, we humans are dynamically embodied and embedded in the socio-technical environments in which we dynamically interact with other humans and other agents. The complex adaptive behaviors represented by these qualities and capabilities aren't fully captured by reason alone," said Wallach. "The idea that we are going to be living in a constantly interfaced world comes out of a rather simplistic notion of human cognition." He said that it would be useful for AI to develop a "theory of mind and emotions" that takes account of supra-rational faculties and morality, and recognizes the dynamics of agents being embedded in complex adaptive systems. "These are things that aren't fully captured by reasoning," said Wallach.

There was broad agreement that future AI research must be interdisciplinary, precisely because the ramifications of the technology reach into so many different corners of life. "AI cannot be solely a technical field," said Crawford, "It has to be a 'sociotechnical field.'" This suggests the need for training graduates to speak across disciplines; to con-

vene more diverse sets of researchers; and to establish the right norms in legitimizing what sets of problems shall be studied, and how.

It seems that academics are likely to take some different approaches to research than corporate researchers, many participants agreed. But this may be less a matter of research priorities than "the power and capacity to instrumentalize research," said Kate Crawford. "Universities commonly have less data, less infrastructure, and far less capacity compared to the private sector." She told of a researcher who left a large company to work in academia, and realized that he could no longer ask the same questions. Not surprisingly, academics often lag behind some of the questions that corporate researchers are addressing.

But Ruchir Pur of IBM said that academics may not be as handicapped as they might think because they tend to be very resourceful and come up with different perspectives that do not emerge in corporate settings. While academics may have fewer resources, said Naveen Rao of Intel, "I would argue that academic research is much broader and more open. Academics have access to philosophers and social studies, and a cross-pollination of ideas, which you don't have as much of in a corporate setting."

The AI Now Institute at New York University, now in its third year, is a new effort to bridge a lot of topics related to AI. It is focused on four major areas of study: bias and inclusion, specifically in machine learning, labor and automation; personal autonomy, including work issues; basic rights and liberties; and critical infrastructures such as power grids, hospitals and education. Another focal point for AI research is an annual conference known as FAT/ML — Fairness, Accountability and Transparency in Machine Learning — which is now entering its fifth year.

Participants noted activists and journalists are a rich source of research and new ideas because they are often closer to problems on the ground than professional researchers. J. Nathan Matias recalled that the founders of the medical journal *Lancet* were doctors concerned about food safety who went out to do their own firsthand research on the streets of London. It seems likely that many insights into the problems of AI will emerge from such practices, said Natalie Bruss, who focuses on Special Projects for tronc, Inc., the media company. "Many issues only manifest and get talked about when they are used in narrative storytelling."

Conclusion

As the astonishing functionalities of AI burst forth and proliferate, it is becoming clearer that these hyper-rational instruments do not just perform certain intended tasks amazingly well. They radically change the context in which traditional institutions, the economy and everyday life operate. Artificial intelligence is not just about a certain type of robotics or predictive analytics or on-the-fly marketing. It is about a much larger landscape—major industrial and commercial sectors, the character of democratic governance, the sovereignty of people in making their own choices, and the lens through which we will view human purpose and meaning. No wonder the questions and quandaries provoked by AI are multiplying!

Artificial intelligence is about a much larger landscape—major industrial and commercial sectors, the character of democratic governance, the sovereignty of people in making their own choices, and the lens through which we will view human purpose and meaning.

Fortunately, as the conference confirmed, there are many thoughtful initiatives underway to try to make sense of a rapidly approaching future. There are concerned players attempting to figure out suitable forms of governance for this very disruptive class of technologies. The disruptions will yield many important benefits, but they could well be eclipsed if trusted, transparent and responsive governance does not also materialize. We hope that this report will help catalyze deeper discussions about this urgent challenge.

Endnotes

1. Gary Kasparov. *Deep Thinking: Where Machine Intelligence Ends and Human Creativity Begins.* Public Affairs Books. (2017), p. 255.

2. Jacques Ellul. *The Technological Society.* (n.d.). Available online: https://ratical.org/ratville/AoS/TheTechnologicalSociety.pdf.

3. David Bollier. "Artificial Intelligence Comes of Age: The Promise and Challenge of Integrating AI into Cars, Healthcare and Journalism." The Aspen Institute Communications and Society Program. Washington, D.C. (2017). Available online: http://csreports.aspeninstitute.org/Roundtable-on-Artificial-Intelligence/2016/report.

4. PwC is an audit and assurance, tax and consulting services. For more information see https://www.pwc.com/.

5. For more on bias in face recognition technologies, see a report by Clare Garvie, Alvaro Bedoya and Jonathan Franle, "The Perpetual Lineup: Unregulated Police Face Recognition in America." Georgetown Law Center on Privacy & Technology. (October 18, 2016). Available online: https://www.perpetuallineup.org.

6. Adam Liptak. "Sent to Prison by a Software Program's Secret Algorithms." *The New York Times.* (May 1, 2017). Available online: https://www.nytimes.com/2017/05/01/us/politics/sent-to-prison-by-a-software-programs-secret-algorithms.html.

7. Julia Angwin et al. "Machine Bias." *Pro Publica.* (May 23, 2016). Available online: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

8. The need for AI systems to "explain themselves" so that human institutions can justifiably rely on them has given rise to a new field of research, "Explainable A.I.," or X.A.I. See Cliff Kuang. "Can A.I. Be Taught to Explain Itself?" *The New York Times.* (November 21, 2017). Available online: https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html.

9. Xiaolin Wu and Xi Zhang. "Automated Inference on Criminality Using Face Images." ArXiv. (2016). Available online: https://www.semanticscholar.org/paper/Automated-Inference-on-Criminality-using-Face-Imag-Wu-Zhang/1cd357b675a659413e8abf2eafad2a463272a85f.

10. For more on GANs, see the Wikipedia entry. "Generative Adversarial Network." (n.d.) Available online: https://en.wikipedia.org/wiki/Generative_adversarial_network.

11. *Houston Federation of Teachers et al. v. Houston ISD.* 14-1189 - Houston Federation of Teachers Local 2415 et al v. Houston Independent School District.

12. White Collar Crime Risk. Available online: https://thenewinquiry.com/white-collar-crime-risk-zones. The methodology for the project is laid out in a white paper by Brian Clifton, Sam Lavigne and Francis Tseng. "Abolish." *The New Inquiry Magazine.* Vol. 59. (April 27, 2017). Available online: https://whitecollar.thenewinquiry.com/static/whitepaper.pdf.

13. Or as a *The New York Times* headline put it, "If a wildly creative company with an app used by 178 million people every day can still be crushed by Facebook, how is anyone supposed to succeed?" The article about Snap, the parent company of Snapchat, was written by Kevin Roose. *The New York Times.* (November 16, 2017). Available online: https://www.nytimes.com/2017/11/15/business/snapchats-new-test-grow-like-facebook-without-the-baggage.html.

14. Tom Krazit. "Updated: Washington's Sen. Cantwell prepping bill calling for AI committee." *GeekWire*. (July 20, 2017). Available online: https://www.geekwire.com/2017/washingtons-sen-cantwell-reportedly-prepping-bill-calling-ai-committee.

15. "Going Digital." OECD. (n.d.). Available online: http://www.oecd.org/going-digital.

16. "MIT Media Lab to Participate in AI Ethics and Governance Initiative." MIT News press release. (January 10, 2017). Available online: http://news.mit.edu/2017/mit-media-lab-to-participate-in-ai-ethics-and-governance-initiative-0110.

17. Partnership on AI. (n.d.). Available online: https://www.partnershiponai.org.

18. International Telecommunications Union. (n.d.). Available online: http://www.itu.int/en/ITU-T/AI/Pages/default.aspx.

19. "The Future of Everything: Sustainable Development in the Age of Rapid Technological Change." United Nations. (October 11, 2017). Available online: http://www.un.org/sustainabledevelopment/blog/2017/10/looking-to-future-un-to-consider-how-artificial-intelligence-could-help-achieve-economic-growth-and-reduce-inequalities.

20. Wendell Wallach and The Hastings Center carried out the project under a three-year Future of Life Institute grant awarded in 2015. Available online: https://futureoflife.org/first-ai-grant-recipients/#Wallach.

21. Gary E. Marchant and Wendell Wallach. "Coordinating Technology Governance." *Issues in Science and Technology*. 31(4). (Summer 2015). Available online: http://issues.org/31-4/coordinating-technology-governance.

22. AI Now Public Symposium. (July 2017). Available online: https://www.youtube.com/watch?v=npL_UsK_npE.

23. For example, Rich Caruana. "Intelligible Machine Learning for Critical Applications Such as Health Care." AAAS 2017 Annual Meeting. (2017). Available online: https://aaas.confex.com/aaas/2017/webprogram/Paper19142.html.

APPENDIX

Artificial Intelligence, The Great Disruptor

Aspen, Colorado
August 6-8, 2017

Roundtable Participants

Paul Blase
Managing Partner,
 AI & Data Solutions
tronc, Inc.

David Bollier
Rapporteur

Joshua Browder
Founder and
 Chief Executive Officer
DoNotPay Robot Lawyer

Natalie Bruss
Special Projects
tronc, Inc.

Joy Buolamwini
Founder,
 The Algorithmic Justice League
Graduate Researcher,
 MIT Media Lab
Aspen Institute Guest Scholar

Michael Chui
Partner
McKinsey Global Institute

David Copps
Founder and
 Chief Executive Officer
Brainspace

France A. Córdova
Director
National Science Foundation

Kate Crawford
Co-Founder, AINow
Principal Researcher,
 Microsoft Research
Distinguised Professor,
 New York University

Charlie Firestone
Executive Director
Communications and Society
 Program
The Aspen Institute

Douglas Frantz
Deputy Secretary-General
Organisation for Economic
 Co-operation and Development

Note: Titles and affiliations are as of the date of the conference.

Jean-François Gagné
Co-Founder and
  Chief Executive Officer
Element AI

John C. Havens
Executive Director
The IEEE Global Initiative on
  Ethics of Autonomous and
  Intelligent Systems

Reed Hundt
Chief Executive Officer
Coalition for Green Capital

Tim Hwang
Director of Ethics and
  Governance
Artificial Intelligence Fund

Alberto Ibargüen
President and
  Chief Executive Officer
Knight Foundation

Joi Ito
Director
MIT Media Lab

Mathew Lira
Special Assistant to the
  President for Innovation
  Policy and Initiatives
Office of American Innovation
White House

J. Nathan Matias
Postdoctoral Researcher,
  Princeton
Aspen Institute Guest Scholar

Terrell McSweeny
Commisioner
United States Federal Trade
  Commission

Ruchir Puri
Chief Architect, IBM Watson
IBM

Naveen Rao
Corporate Vice President and
  General Manager
Artificial Intelligence Products
  Group
Intel Corporation

Louis Rosenberg
Founder and
  Chief Executive Officer
Unanimous A.I.

Marc Rotenberg
President and Executive Director
Electronic Privacy Information
  Center

Fr. Eric Salobir
Founder and President
OPTIC

Terrence Southern
Founder, Illuminatestem.org
Global Lead Robotics and
  Automation Engineer
GE Global Research

Wendell Wallach
Author, Scholar
Yale University

Note: Titles and affiliations are as of the date of the conference.

Staff:

Kristine Gloria
Project Manager
Communications and Society
 Program
The Aspen Institute

Note: Titles and affiliations are as of the date of the conference.

About the Author

David Bollier is an author, activist, independent scholar and blogger well-known for his work on the commons as a new paradigm of economics, politics and culture. He pursues this scholarship and activism as co-founder of the Commons Strategies Group, an advocacy/consulting project that assists the international commons movement. Bollier has written or edited seven books on the commons, including *Patterns of Commoning* (2015), co-edited with Silke Helfrich; *Think Like a Commoner: A Short Introduction to the Life of the Commons* (2014); *Green Governance: Ecological Survival, Human Rights and the Commons* (2013), co-authored with Burns Weston; and an anthology of essays, *The Wealth of the Commons: A World Beyond Market and State* (2012), co-edited with Silke Helfrich.

Bollier spent many years in various policy jobs in Washington, D.C. (in Congress, the auto-safety agency, with Ralph Nader and others) in the 1970s and 1980s. In 2001 Bollier co-founded Public Knowledge, a Washington advocacy organization for the public's stake in the Internet, telecom and copyright policy. For twenty-five years, until 2010, Bollier collaborated with television producer, writer and activist Norman Lear on a wide variety of non-television public affairs and political projects. Bollier blogs at Bollier.org and lives in Amherst, Massachusetts.

About the
Communications and Society Program

www.aspeninstitute.org/c&s

The Communications and Society Program is an active venue for framing policies and developing recommendations in the information and communications fields. We provide a multi-disciplinary space where veteran and emerging decision-makers can develop new approaches and suggestions for communications policy. The Program enables global leaders and experts to explore new concepts, exchange insights, develop meaningful networks, and find personal growth, all for the betterment of society.

The Program's projects range across many areas of information, communications and media policy. Our activities focus on a broad spectrum of ICT issues such as artificial intelligence, broadband and spectrum policy, racial inclusion in communications, institutional innovation, and diplomacy and technology. Its major activity for 2017-18 is conducting the Knight Commission on Trust, Media and Democracy. The program also runs a project on the future of public libraries.

Most conferences employ the signature Aspen Institute seminar format: approximately 25 leaders from diverse disciplines and perspectives engaged in roundtable dialogue, moderated with the goal of driving the agenda to specific conclusions and recommendations. The program distributes our conference reports and other materials to key policymakers, opinion leaders and the public in the United States and around the world. We also use the internet and social media to inform and ignite broader conversations that foster greater participation in the democratic process.

The Program's Executive Director is Charles M. Firestone. He has served in this capacity since 1989 and is also a Vice President of the Aspen Institute. Prior to joining the Institute, Mr. Firestone was a communications attorney and law professor who has argued two cases before the United States Supreme Court and many in the courts of appeals. He is a former director of the UCLA Communications Law Program, first president of the Los Angeles Board of Telecommunications Commissioners, and an appellate attorney for the U.S. Federal Communications Commission.

Previous Publications from the Aspen Institute Roundtable on Artificial Intelligence

Artificial Intelligence Comes of Age: The Promise and Challenge of Integrating AI into Cars, Healthcare and Journalism, by David Bollier

The Aspen Institute Roundtable on Artificial Intelligence (AI) is an annual private seminar of business executives, government leaders, and visionaries. It examines the implications of artificially intelligent technologies on societies, governments, communities and individuals, and the new leadership roles that are required. As its inaugural year (2016), the report concentrated on three artificial intelligence sectors: self-driving cars, AI and Medicine, and AI and the Media.

Reports can be ordered online at *www.aspeninstitute.org/publications* or by sending an email request to *publications@aspeninstitute.org.*