Atlantic Council

GEOTECH CENTER

Artificial Intelligence Principles: Avoiding Catastrophe

Robert A. Manning
Resident Senior Fellow, Foresight, Strategy, and Risks Initiative

ABOUT THIS REPORT

This publication is part of the Atlantic Council's ongoing endeavor to establish forums, enable discussions about opportunities and challenges of modern technologies, and evaluate their implications for society as well as international relations — efforts that are championed by the newly established *GeoTech Center*. Prior to its formation and to help lay the groundwork for the launch of the Center in March 2020, the Atlantic Council's *Foresight, Strategy, and Risks Initiative* was awarded a *Rockefeller Foundation* grant to evaluate China's role as a global citizen and the country's use of AI as a development tool. The work that the grant commissioned the Atlantic Council to do focused on data and AI efforts by China around the world, included the publication of reports, and the organization of conferences in Europe, China, Africa, and India. At these gatherings, international participants evaluate how AI and the collection of data will influence their societies, and how countries can successfully collaborate on emerging technologies, while putting a special emphasis on the People's Republic in an ever-changing world.

THE ATLANTIC COUNCIL GEOTECH CENTER

Produces events, pioneers efforts, and promotes educational activities on the Future of Work, Data, Trust, Space, and Health to inform leaders across sectors. We do this by:

- Identifying public and private sector choices affecting the use of new technologies and data that explicitly benefit people, prosperity, and peace.
- Recommending positive paths forward to help markets and societies adapt in light of technology- and data-induced changes.
- Determining priorities for future investment and cooperation between public and private sector entities seeking to develop new technologies and data initiatives specifically for global benefit.

Championing Positive Paths Forward That Nations, Economies, and Societies Can Pursue To Ensure New Technologies and Data Empower People, Prosperity, and Peace

An urgent challenge for the coming decade is to forge a global consensus to operationalize widely-shared ethical principles, standards, and norms to govern the development and use of artificial intelligence. This should be the predicate for all other AI issues. As AI becomes a ubiquitous driver of economic growth and shapes legal, medical, educational, financial, and military sectors, there is no consensus on rules, standards, or operating principles governing its use on either a national or global basis. As is often the case, technology is racing ahead of efforts to control it. The rich literature and inherent risks with regard to complex systems and their tendency to fail should provide a sense of urgency so far not apparent in either the private sector, government or US Congress.[1] Unless there is a broad global consensus on core principles and standards, and how to make them operational, there is significant downside risk of a dangerous race to the bottom with potentially catastrophic consequences as AI applications become more widespread. There appears a lack of urgency among G-20 governments to transform accepted broad principles into functional global governance. The two leading tech powers, the US and China, appear to be moving in the direction of heightened techno-nationalism rather than seeking to shape global standards for the safe, secure, and beneficial deployment of AI.

AI, which is fundamentally, data plus algorithms, is usefully viewed as an enabler or synthesizer of an inter-connected BigData/IoT/robotics/biotech suite of technologies, promises to be a game-changer from economics to the automation of the battlefield. AI, however, is more an enabling force, like electricity than a thing, and like apps today, it is already being applied to most industries and services. The issue is how the data is employed. Think of AI as a new platform: the future will be everything plus AI. By 2030, algorithms combined with data from 5G Internet of Things (IoT) will be in every imaginable app and pervasive in robots, reshaping industries from healthcare and education to manufacturing, finance, and transportation, to military organizations. AI is already starting to be incorporated into military management, logistics, and target acquisition. Yet there is a dangerous governance deficit, given the fact that there are few principles, norms, and standards guiding the growing applications of artificial intelligence.

As Kai Fu-Lee writes, AI is evolving beyond one-dimensional tasks of what is called “narrow AI” to “general AI.” The former refers to single tasks such as facial recognition or language translation; the latter, AI that can operate across a range of tasks using learning and reasoning without supervision or external input to solve any problem, learning from layers of neural network of data, is in its early days of development. Two-thirds of private sector investment in AI is in machine learning, deep learning, using neural networks, mimicking the human brain to use millions of gigabytes of data to solve problems. Most famously, *AlphaGo* beat the world champion in Go, a very complex Asian game with millions of moves. It was fed data from thousands of Go matches and was able to induce the best possible moves to out maneuver its opponent. AI is demonstrating a growing capability to learn autonomously extrapolating from the data fed into the algorithm.

But algorithms have their limits, too. Particularly, with regard to the prospect of autonomous systems, AI lacks understanding of context, culture, emotions, and meaning: can it tell if someone is pointing a real gun at it or a toy pistol, or what their intent is by doing so? Some leading neurologists are skeptical, arguing that intelligence

[1] A system comprised of many components with distinct properties that work via interaction with each that creating a dynamic may adapt to changes or feedback loops.

requires consciousness. Emotions, memories, and culture are part of human intelligence that machines cannot replicate. There is a growing body of evidence that AI can be hacked (i.e. misdirecting an autonomous car) or spoofed (i.e., identifying targets) with false images. One potential problem for many applications is that we don't know exactly how AI learns what it knows, meaning how the process worked resulting in its decision and conclusion. Circumstances that make it difficult to test, evaluate, or know why it was wrong or malfunctioned – and will only be more difficult as deep learning becomes more sophisticated. Yet explainability, not least, the need to know why an AI system failed, is a cardinal principle of an AI safety and accountability regime. How, for instance, can liability be determined if we don't understand whose fault it is? At the same time, there is a relatively low bar to entry the game, as transparency among AI researchers has spawned wide access. For example, there are several open source websites, one prominent one, TensorFlow, to which leading researchers from top tech firms such as Google not only post their latest algorithms, but *TensorForce* also enables one to download neural networks and software with tutorials showing techniques for building your own algorithms that could also be deployed to the future of warfare, as Paul Scharre writes in his book the "Army of None." This obviously ups the ante with regard to the need for common ethics and operating principles.

PREVENTING THE COMING STORM

The urgency of developing a global consensus on ethics and operating principles for the uses and restrictions on AI starts from our knowledge that complex systems like supercomputers, robots, or Boeing 737 jets, with multiple moving parts and inter-acting systems, are inherently dangerous and prone to fail, sometimes catastrophically. Because the failure of complex systems may have multiple sources, sometimes triggered by small failures cascading to larger ones, it can require multiple instances to fully understand the causes. This problem of building in safety is compounded by the fact that it is increasingly difficult, as AI gets smarter, to discern why AI decided to reach its conclusions.

The downside risks in depending solely on an imperfect AI, absent the human factor in decision-making, have already begun to reveal themselves. For example, research on facial recognition has shown bias against certain ethnic groups, apparently due to the preponderance of white faces in their respective database. Similarly, as AI is employed in a variety of decision-making roles such as job searches or determining parole, absent a human to provide context, cultural perspective, and judgment, bias becomes more likely. For example, can AI discern character or personality traits absent on a resume that may lead an employer to be more or less likely to hire an applicant? Or can AI accurately detect how a prisoner may be changed for better or worse while incarcerated to recommend parole?

More ominously, there have been incidents when semi-autonomous missile systems have hit wrong targets, as the USS Vincennes erroneously shot down an Iranian civilian airliner in 1988, or US Aegis-3 missiles mistakenly hitting a US target. The risk of waking up and discovering that fully autonomous weapons started an escalating war as the enemy's autonomous weapons retaliated, is a scenario that could be possible in the coming decade. Clearly, the risk of AI systems going awry is significant, particularly given the wide range of potential scenarios. How do you assess liability for failure? What safety standards are required to assure accountability? How do we assure transparency in failure – human understanding of "how and why AI makes a specific decision," as a Chinese White Paper put it.

Such core concerns have spurred pro-active efforts by a wide array of stakeholders from both the private sector as well as prominent international technologists and scientists to create a governance structure for artificial intelligence. The world's biggest tech firms including Google and Microsoft, for example, are among the vanguard in seeking the creation of "ethical AI" guidelines. Google's AI principles list now-familiar items such as accountability, safety, and a commitment to ensure that AI systems are unbiased. Microsoft has adopted similar principles, and both are members of the Partnership on AI, a multi-stakeholder organization of some 80 partners, including leading tech firms, NGOs, and research institutes. Of course, private sector activism is largely driven by an imperative to sort out liability, accountability, and fairness issues in developing and deploying AI for profit. At the same time, over the past several years, there has been no dearth of efforts by government agencies and commissions, quasi-official expert bodies, technicians, and scientists to define AI ethics, principles, and standards. The EU, which has been a leader in tech standard setting, is a good example. In April 2019, the EU Commission released its *Ethics Guidelines for Trustworthy Artificial Intelligence*, put together by a high-level expert group on AI. The document spelled out a set of "fundamental rights" from which "trustworthy AI" should be derived. Those rights consist of broadly-shared democratic norms that underpin European institutions. From these norms, spelled out as rights, the EU's expert group derived seven AI guidelines.

- Be subject to human agency and oversight;
- Be technically robust and safe;
- Ensure privacy and allow for good data governance;
- Be transparent (for example, AI systems ought to inform people that they are interacting with an artificial system rather than another person);
- Enable diversity, non-discrimination, and fairness;
- Work in the service of societal and environmental well-being;
- Be accountable ("In applications affecting fundamental rights, including safety-critical applications, AI systems should be able to be independently audited" by external parties).

As the EU has done with regard to data regulation, the expert group's activities suggest that the continent will likely play a large role in shaping global standards for AI regulation, as it has for data regulation. The EU's General Data Protection Regulation (GDPR), which went into effect in 2018, is already forcing companies to prove compliance with EU regulations regarding data protection. Companies that seek to do business within Europe must comply in order to avoid steep fines and retain access to European markets. In this vein, the GDPR is on its way of approaching a global standard, though there is risk that AI standards like data regulation and the internet, become splintered by competing norms. Other governments have modeled legislation or regulation after the GDPR, including Japan, which harmonized its data privacy regulations with European standards, and the state of California, which passed the California Consumer Privacy Act (CCPA), a law that comes into force in 2020. There is further contemplation whether to align the CCPA even more closely with the GDPR. In both cases, governments have been motivated in part by gaining an "adequacy determination" under the GDPR, meaning that the EU would allow that country's firms to transfer their data from Europe to the home country (or state in California's case).

Europe's AI ethics guidelines might be an important first step toward translating principles into regulatory standards and norms. Even though some of its provisions capture AI principles and already have had an effect on AI-related business operations in Europe, the GDPR doesn't mention AI. Critics further argue that the GDPR's data privacy and transparency provisions are overly broad, difficult to enforce, and costly for firms wishing to develop and use AI applications. For example, the GDPR requires firms to give individuals the right to a human review of a decision made by an automated (AI) system, the effect of which is to increase a firm's costs (one of AI's great advantages is its ability to process massive amounts of data swiftly). These critics argue that, without reform, the GDPR will depress AI-related investment within Europe and shift even more of it to China and the United States.

Other governments and multilateral institutions have crafted AI ethics guidelines that are similar to the EU's, the OECD being among the most recent and notable. In May 2019, it released five "complementary values-based principles" for responsible AI. These assert that AI should:

- drive "inclusive growth, sustainable development and well-being";
- respect the rule of law;
- be transparent;
- be robust and secure;
- AI system designers and owners should be accountable for their systems' impacts."inclusive growth, sustainable development and well-being";

In June 2019, the G20 adopted its own set of principles that were drawn entirely from the OECD's principles. In the US, there are piecemeal efforts to build a consensus. One prominent effort is the 2017 Asilomar Principles, a set of 23 principles covering research, ethics, and values including safety, failure transparency, and human control to avoiding the development of fully autonomous weapons. The document was endorsed by more than 2500 AI/robotics researchers, engineers, prominent technologists, and scientists, including Elon Musk and the late Stephen Hawking. Similarly, the Institute for Electrical and Electronics Engineers (IEEE) a leading body with nearly half a million members globally. IEEE created a Global Initiative on Ethics of Autonomous and Intelligent Systems in 2016 to ensure those designing and developing AI prioritize ethical standards. The US Department of Defense Innovation Board in 2019 issued a similar set of ethical principles for AI. China's 2018 *White Paper on Artificial Intelligence Standardization* has a similar tone and thrust, arguing that "relevant [AI] standards remain in a blank state," and says, "China should strengthen international cooperation and promote the formulation of a set of universal regulatory principles and to ensure the safety of AI technology." The White Paper goes on to outline key principles that largely overlap with Western ones discussed above, including: AI should benefit human welfare; safety is a prerequisite for sustainable technology; a clear system of liability to hold AI developers accountable; transparency requires understanding of how and why AI makes a specific decision; a clear definition of privacy." The White paper, issued by the Chinese Electronics Standards Institute is semi-official, and similar views are echoes by a number of Chinese Institutes. At a minimum, the US, EU, Japan, and others should test its sincerity by actively pursuing negotiations to operationalize these ethics and principles in binding agreements.

- ***Human agency and benefit:*** *research and deployment of AI should augment human well-being and autonomy; have human oversight to choose how and whether to delegate decisions to AI systems; be sustainable, environmentally friendly, compatible with human values and dignity;*
- ***Safety and Responsibility:*** *AI systems should be technically robust, based on agreed standards, verifiably safe, including resilience to attack and security, reliability and reproducibility;*
- ***Transparency in failure:*** *If an AI system fails or causes harm or otherwise malfunctions, it should be explainable why and how the AI made its decision – algorithmic accountability;*
- ***Avoiding arms races:*** *An arms race in lethal autonomous weapons should be avoided. Decisions on lethal use of force should be human in origin*
- ***Periodic Review****: Ethics and principles should be periodically reviewed to reflect new technological developments, particularly in general deep learning AI.*

Translating such principles into operational social, economic, legal, and national security policies will be a daunting task. These ethical issues already confront business and government decision-makers. Yet neither have demonstrated comprehensive policy decisions on implementing them, suggesting that establishing governance is likely to be an incremental, trial-and-error process.

How to decide standards and liability for autonomous vehicles, data privacy, and algorithmic accountability is almost certainly a complex goal very difficult to attain. Moreover, as AI becomes smarter, the ability of humans to understand how AI made decisions is likely to diminish. Even though AI may be a top arena of US-China tech competition, given the risks of catastrophic failure and the desire for global markets, this should necessitate such norms for both governments and industry. The need for human responsibility and accountability for AI decisions and the downside risks of unsafe AI and lack of transparency to understand failure are shared dangers. One recent example of US, China, and other competitors cooperating is in the creation of standards and technical protocols for 5G – a fierce arena of US-China competition. A coalition of global private sector telecom/IT firms, known as 3rd Generation Partnership Project (3GPP) in collaboration with the ITU, a key UN standard-setting institution, have, so far successfully, agreed to a host of technical standards for 5G technology.

While some in the US complained of Chinese assertiveness in pursuing standards that tend to favor Chinese tech, Beijing played by the rules, and like other stakeholders (albeit, more aggressively), sought to shape global standards. US firms also had the opportunity to push for their preferred guidelines, they just have not matched Chinese efforts. But the point is that markets are global and all stakeholders have an interest in tech standards reflecting that, competitors or not. Given the enormous stakes, getting AI governance right, 'strategic competitors' or not, both US and China have a mutual interest in adopting safe, secure, and accountable rules for AI applications. This should be an area of public-private partnership, with US and Chinese Big Tech having much at stake. With AI at the center of US-China tech competition, whether common global ethical principles, norms, and standards can be adopted, or whether US-China zero-sum competition leads to a dangerous race to the bottom, is a question with huge, and potentially catastrophic consequences. As the two leading AI powers, to the degree the US and China can reach accord in a bilateral dialogue, the outcome would likely shape parallel global efforts to achieve consensus in international standard setting institutions. After all, the G-20 has already embraced the OECD AI principles.

ACTION POINTS

- The **US needs a Presidential Commission** comprised of engineers, technologists, private sector, and Congress to recommend national policies on control of data and AI regulatory standards/ethics, building on NIST recommendations. This is a *sine qua non* to reinforce American leadership;
- A potential next step could be **a G-20 mandate** to negotiate norms, standards, and ethical principles for the use and restrictions of AI applications, and a new international mechanism to codify and monitor them;
- **Launch US-EU-China talks on AI** governance, a key building block. Whatever consensus achievable among the tech giants, would create a powerful basis for global standards;
- **Create a standing international regulatory body on AI standards and ethics** under UN auspices with a UN Security Council mandate, the International AI Commission (IAIC). It should have standards function like the ITU, but with an arbitration function similar to WTO dispute mechanism. This body should also have an Advisory Board comprised of engineers, technologists, tech firms, and legal experts.

ABOUT THE AUTHOR

Robert A. Manning is a senior fellow with the Scowcroft Center for Strategy and Security and its Strategic Foresight Initiative at the Atlantic Council. Previously, he served as a senior Strategist at the National Counterproliferation Center in the Office of the Director of National Intelligence (ODNI) from 2010 to 2012, and as the director of long-range energy and regional/global affairs at the US National Intelligence Council's Strategic Futures Group from 2008 to 2010. From 2005 to 2008, Manning served as a member of the US Secretary of State's Policy Planning Staff, and from 2001 to 2005, he was senior counsellor for energy, technology, and science policy at the US Department of State, where he advised the Under Secretary of State for Global Affairs and other senior officials on a range of issues including energy and climate change policy and new energy technologies.From 1997 to 2001, he was director of Asian studies and a senior fellow at the Council on Foreign Relations (CFR). He led several CFR task forces, including the Korea Task Force and the Southeast Asia Task Force among others. Manning was previously an adviser for policy and public diplomacy to the assistant secretary of State for East Asian and Pacific affairs at the State Department and served as an adviser to the Office of the Secretary of Defense from 1988 to 1989.

His publications include *The Asian Energy Factor* (Palgrave/St. Martins 2000) and *China, Nuclear Weapons and Arms Control*. He has published essays on nuclear weapons; numerous journal articles on international energy and Asian security issues; and roughly half a dozen book chapters in edited volumes on China, Korea, Japan, regional security architecture, energy, and energy security. He has published widely in Foreign Affairs, Foreign Policy, the National Interest, the New York Times, the Washington Post, the Los Angeles Times, Chosun Ilbo, and other publications.

THE ATLANTIC COUNCIL GEOTECH CENTER

Produces events, pioneers efforts, and promotes educational activities on the Future of Work, Data, Trust, Space, and Health to inform leaders across sectors. We do this by:

- Identifying public and private sector choices affecting the use of new technologies and data that explicitly benefit people, prosperity, and peace.
- Recommending positive paths forward to help markets and societies adapt in light of technology- and data-induced changes.
- Determining priorities for future investment and cooperation between public and private sector entities seeking to develop new technologies and data initiatives specifically for global benefit.

CHAMPIONING POSITIVE PATHS FORWARD THAT NATIONS, ECONOMIES, AND SOCIETIES CAN PURSUE TO ENSURE NEW TECHNOLOGIES AND DATA EMPOWER PEOPLE, PROSPERITY, AND PEACE