CENTER FOR LONG-TERM CYBERSECURITY

CLTC WHITE PAPER SERIES

# Toward AI Security

## GLOBAL ASPIRATIONS FOR A MORE RESILIENT FUTURE

JESSICA CUSSINS NEWMAN

CLTC WHITE PAPER SERIES

# Toward AI Security

## GLOBAL ASPIRATIONS FOR A MORE RESILIENT FUTURE

JESSICA CUSSINS NEWMAN

FEBRUARY 2019

CLTC
Center for Long-Term
Cybersecurity

CENTER FOR LONG-TERM CYBERSECURITY

# Contents

# Acknowledgements

The Center for Long-Term Cybersecurity (CLTC) would like to thank the following individuals for their expertise, feedback, and contributions to this document: Anthony Aguirre, Seth Baum, Niel Bowerman, Jared Brown, Ryan Calo, Betsy Cooper, Elsa Kania, Jade Leung, Nicolas Miailhe, Deirdre Mulligan, Michael Page, Philip Reiner, Stuart Russell, and Helen Toner.

Special thanks to Steven Weber, Ann Cleaveland, and Chuck Kapelke of CLTC for their support, feedback, and contributions to this report.

CLTC would also like to thank the Hewlett Foundation for making this work possible.

# Abstract

This report uses the lens of global AI security to investigate the robustness and resiliency of AI systems, as well as the social, political, and economic systems with which AI interacts. The report introduces a framework for navigating the complex landscape of AI security, visualized in the AI Security Map. This is followed by an analysis of AI strategies and policies from ten countries around the world within this framework to identify areas of convergence and divergence. This comparative exercise highlights significant policy gaps, but also opportunities for coordination and cooperation among all surveyed nations. Five recommendations are provided for policymakers around the world who are hoping to advance global AI security and move us toward a more resilient future. The steps nations take now will shape AI trajectories well into the future, and those governments working to develop global and multistakeholder strategies will have an advantage in establishing the international AI agenda.

# Recommendations

Based on the analysis of the gaps and opportunities in national AI strategies and policies, we provide five recommendations for policymakers hoping to harness and direct AI technologies for a more resilient and beneficial future. These recommendations outline concrete actions that can be taken now to address a complex and quickly changing sociotechnical landscape:

1. **Facilitate early global coordination where common interests can be identified.** As autonomous systems become more ubiquitous and capable, their reach and effects will be more consequential and widespread. Global coordination and cooperation will be essential for ensuring sufficient oversight and control, but such cooperation will be harder to achieve the longer we wait due to technological and institutional "lock-in". The numerous areas of convergence identified in this report can be leveraged as opportunities for collaboration and innovation, sharing best practices, and preventing global catastrophic risks.

2. **Use government spending to shape and establish best practices.** Governments have an opportunity to establish standards and best practices while promoting AI development and use, for example by implementing guidelines for government procurement of AI systems, and by adding criteria such as safety, robustness, and ethics to AI R&D funding streams. Additionally establishing processes to support transparent and accountable government funding and use of AI technologies will help prevent misuse throughout public services and protect government actors from the limitations and vulnerabilities of AI tools.

3. **Investigate what is being left on the table.** The landscape of AI security is broad and complex, as indicated in the AI Security Map presented in this report. The analysis of policy documents identifies many gaps in different nations' current AI policy approaches. Governments may choose to prioritize a sub-set of issues, but they should recognize the opportunities and challenges they could be neglecting.

4. **Hold the technology industry accountable.** Many governments rightfully emphasize the importance of partnership and engagement with industry and other AI stakeholders. However, while some firms are addressing AI challenges, significant gaps remain. Policymakers have the unique primary responsibility to protect the public interest, and this responsibility carries even greater weight during periods of significant technological transformation.

Governments should ensure their citizens have access to the benefits that emerge from AI development and are proactively protected from harms.

5. **Integrate multidisciplinary and community input.** To support the widespread goal of improving government expertise in AI, policymakers should formalize processes to ensure multidisciplinary input from AI researchers and social-science scholars and practitioners. Community engagement should additionally form an integral part of any decision to implement AI into public services.

# Introduction

Artificial intelligence (AI) may be the most important global issue of the 21st century, and the way that we navigate the security implications of AI could dramatically shape the kind of futures we experience.[1] Although research in AI has been taking place since the 1950's, recent years have seen substantial growth in interest, investment dollars, and jobs,[2] leading to important advances in real-world applications ranging from autonomous vehicles to cancer screening.[3] However, much of the truly transformative potential of AI still remains to be seen, as more industries implement AI technologies, and as the capabilities of AI systems improve and exceed those of humans across more domains.

In the near future, AI could become the most important commercial opportunity in the global economy. A 2017 PwC report predicts that gains from productivity and consumer demand from AI will contribute an additional 14% to global GDP by 2030 —equivalent to $15.7 trillion.[4] Plans to capitalize on AI are increasingly referenced in national strategies across the world.[5] However, as policymakers encourage AI development, they must simultaneously consider potentially harmful impacts of AI such as the automation of jobs, AI-enabled cyberattacks, and the potential for error and discriminatory effects in algorithmic decision-making. National leaders are eager to enable their countries to capitalize on the industrial benefits without being subjected to systems that are unsafe or unaligned with their laws and values.

Addressing that balance requires understanding advances in AI as sociotechnical phenomena that are more than the sum of their technological capabilities. For example, while some estimate that between 400-800 million jobs worldwide could be automated by 2030,[6] there is little consensus about these figures. In addition to open questions about technological development, we also have uncertainty about how nations and communities will respond. When elevators were automated in the early 1900's human operators were still kept around for decades because they helped promote trust and safety.[7] There has also been a slower-than-expected path of acceptance for autonomous vehicles, and a Deloitte study found that "trust appears to be the biggest roadblock to selling the notion of self-driving cars."[8]

While it is hard to foresee the future of technology developments narrowly, it is even harder to foresee the evolution of the sociotechnical systems that will incorporate and emerge along with those technology developments. Nonetheless, governments and companies have begun publishing AI strategies, principles, and codes of conduct in recent years, which help indicate the priorities and constraints of several key actors. These strategies and principles are not merely reactive; they also shape the possibility spaces of the future. As regions begin implementing new processes, legislation, and institutions they are looking to these documents for inspiration. Importantly, we are not yet locked into particular trajectories, and there is still a window of opportunity to shape the contours of these spaces.

AI policy is a relatively novel domain that is actively in flux, and this analysis should be understood as a snapshot of a particular moment in time. Nonetheless, there is a need for frameworks that support the comparative analysis of emerging global AI policies. Mapping the goals of key actors in the AI ecosystem provides a glimpse into multiple possible futures. The identification of gaps and early opportunities for cooperation can also promote broader and more equitable adoption of AI technologies. Moreover, such analysis can support preparation for the complex landscape of AI security threats that are not limited to within national borders, but rather represent a critical new global challenge.

This report briefly describes relevant AI terminology; it then assesses several features of AI technologies that could be transformative. The report then introduces AI security as an important lens for policymakers interested in systemic impacts from AI, and provides a framework for making sense of the space across four security domains. The global AI policy analysis begins on page 29 and includes comparison of actions occurring across ten nations: China, France, United Kingdom, United States, Canada, India, Japan, Singapore, South Korea, and the United Arab Emirates. This analysis provides an initial landscape of global AI security priorities, including a comparison of actual policies being implemented in the AI Policy Compendium in Appendix I.

# AI 101

Artificial intelligence refers to a large suite of technologies, some of which have been implemented in services we already use on a daily basis. These technologies have enabled dramatic improvements in areas such as search engines, spam filters, real-time driving directions, translation, and image recognition. Additional developments such as self-driving vehicles and AI chat bots are becoming more commonplace.

The term "artificial intelligence" was originally coined in 1956 at a Dartmouth summer workshop intended to develop "thinking machines." However AI is notoriously difficult to define. By 2007, more than 70 definitions of the term were in use.[9] One definition suggests, "Artificial intelligence is that activity devoted to making machines intelligent, and intelligence is that quality that enables an entity to function appropriately and with foresight in its environment."[10] A prominent AI textbook by Stuart Russell and Peter Norvig describes definitions of AI as varying along two key dimensions: whether they refer to thought processes or to behavior; and whether they measure success in terms of human performance or an ideal conception of intelligence.[11]

It is easy to take new advances in AI for granted as merely the same computational power we are used to. In fact this reaction is so common it has been named the "AI effect," whereby as soon as a system accomplishes something, it is no longer considered "AI" and the goalpost by which we judge "intelligence" shifts forward. Progress in AI has not always been smooth, but relatively recent advances—alongside advances in processing power and data availability—have enabled AI systems to surpass human ability in a variety of domains, including facial recognition and translation between natural languages.

While AI is a convenient umbrella term, there are many prominent sub-fields and schemas for categorizing the component technologies and capabilities of AI systems.

A common distinction used to describe the properties of AI is between *narrow* (or weak) AI and *general* (or strong) AI. Narrow AI describes all of the instances of AI we have seen to date; these are AI systems that can achieve a single task, even when those are challenging tasks such as driving a car. General AI, or artificial general intelligence (AGI), refers to the scenario in which an AI system achieves human level intelligence and can do any intellectual task that a human can do.

Although there are no examples of AGI today, many researchers believe we could see the emergence of such transformative AI in the coming decades.[12] We have already seen significant breakthroughs in this direction, for example from London-based AI company DeepMind, developer of AlphaZero, a neural network that the company claims is “generalizable to a large number of domains.” Moreover, dozens of companies and organizations around the world are actively pursuing the development of AGI today.[13] Several groups are hoping to achieve AGI by “whole-brain simulation,” the attempt to recreate the activity of the billions of neurons and trillions of synapses in the human brain.

A third category, often called artificial superintelligence (ASI), refers to the concept that machines could one day surpass the cognitive ability of humans across all domains. Some experts predict that the emergence of superintelligence is likely within the first third of this century.[14] Many AI experts predict that AI will outperform humans in many activities within the next ten years, and will outperform humans in all tasks within 45 years.[15] If these predictions have any validity, they ought to generate substantive attention from global leaders.

The Defense Advanced Research Projects Agency (DARPA) offers a different schema for AI development that categorizes technical advances into three waves, providing a useful way to think about more specific AI capabilities. [16] The first wave is “handcrafted knowledge,” which refers to systems that can reason through narrow problems and describe their findings. For example, these kinds of systems have been useful for identifying cyberattacks. The second wave is “statistical learning”, which includes more recent developments such as neural nets that are capable of perceiving and learning as well as making predictions. However, second-wave systems still do not understand the context in which they are acting, and can be unreliable in individual cases. The third wave of AI is “contextual adaptation,” and this is where DARPA sees some advances and hopes AI will continue to go. These systems would be designed to have a contextual model of themselves and the world, which would conceivably enable them to reason and make abstractions.

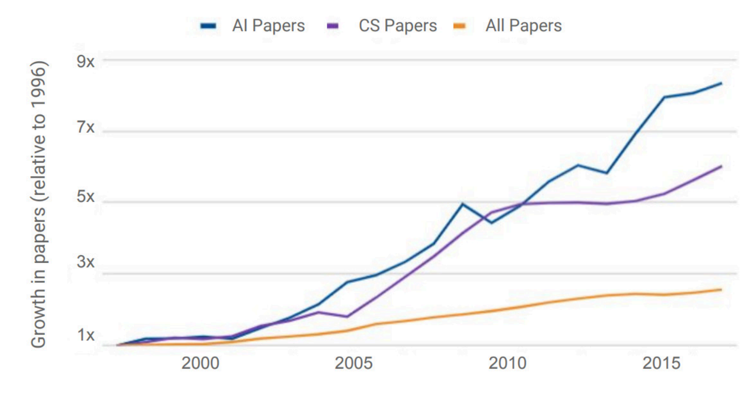
Another name to describe first wave AI is “Good Old Fashioned AI” (GOFAI). These systems are built around a series of rule-based steps based on symbolic reasoning and logic. In contrast, the sub-field of machine learning (ML) refers to systems that can teach themselves. These systems are able to make inferences and predictions from data, create models, and perform tasks. Machine learning is an enormous field that is responsible for the majority of recent AI advances, and it has many sub-fields of its own.

Sub-fields of ML include "supervised learning," in which each piece of data is accompanied with the correct answers of what the system is supposed to learn; "unsupervised learning," in which the system is intended to identify its own patterns within data that may be unstructured; and "reinforcement learning," in which an ML agent learns to act through its experience of an environment, such as a game.

"Deep learning" is a particularly powerful and promising area of ML that can be applied to any of the three aforementioned ML methods. Deep learning is an architectural model that uses neural networks inspired by human brains to make sense of data through many layers of processing, extracting different features from the data until it finds what it wants. Deep learning has enabled advances in areas such as computer vision and language processing.

While ML, AGI, and other terms are useful for describing specific aspects of technological development, AI remains a relevant umbrella term that encompasses the complexity of the field without limiting discussion to a single method of interest or capability. The AI national strategies and policy documents reviewed in this report primarily use the term artificial intelligence, or AI, to refer to all or many of the advances discussed above.

Growth of annually published papers by topic (1996–2017)
Source: Scopus



*Note: This visual uses the Scopus query search term "Artificial Intelligence," not the Elsevier keyword approach. See more details in the appendix.*

From "The AI Index 2018 Annual Report." See End Note 213 for full reference.

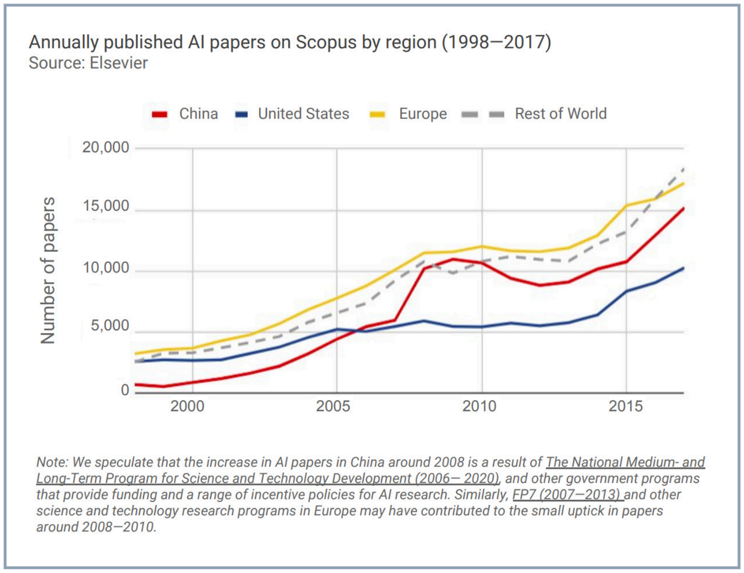
# The Transformative Nature of AI

Increases in computing power, storage, and the volume of data, as well as more advanced algorithms have all contributed to more powerful AI systems. For example, when the IBM Supercomputer Deep Blue beat world chess champion Garry Kasparov in a chess tournament in 1997, many saw it as a symbolic sign of progress in AI technology. But Deep Blue merely used brute force computing to overpower its competitor.

In contrast, the game Go has a significantly larger problem space in which brute force computation is not enough. Go has long been viewed as an art form in China and many assumed it would be impossible for a machine to defeat a human. When DeepMind's AI program AlphaGo beat Go world champion Lee Sedol in a tournament in March 2016, 60 million people in China tuned in to watch.[17] Some argue that the cultural significance of this moment led to a "Sputnik moment" in China, as it led the government to dramatically increase investment in the AI technologies that made this feat possible.[18]

As it turns out, that advance was only the beginning. Just one year later, DeepMind created a new iteration of the program, called AlphaGo Zero.[19] This AI program did not learn to play Go by watching thousands of human games, but rather learned "from scratch". Armed only with the rules, AlphaGo Zero learned to play simply through self-play. Within three days, AlphaGo Zero surpassed AlphaGo, winning one hundred games to zero. A DeepMind article announcing the program's success explained, "This technique is more powerful than previous versions of AlphaGo because it is no longer constrained by the limits of human knowledge." By December 2017, DeepMind announced AlphaZero, a "more general" AI program that not only mastered Go in twenty-four hours, but also chess and shogi.[20] Of course, mastering games is only an initial test for AI systems, useful to better understand their abilities within controlled contexts, but such feats are hardly the end goal. AI systems will similarly transform other complex spaces.

Several features help explain AI's transformational qualities and geopolitical relevance. First, AI is a general-purpose technology, or GPT—a single recognizable technology that is used for numerous purposes across the economy and has many spillover effects.[21] AI is a GPT because it is widely distributed and has many different uses. As software, AI is prone to diffusion, and the existence of open-source databases for AI source code has propelled advances around the world. Other GPTs include electricity, the steam engine, railroads, and the computer. GPTs often have dramatic effects on economies and societies, including substantial unintended consequences. However, AI may additionally pose unprecedented challenges due to other features that may make it more powerful and less controllable than other GPTs.


Annually published AI papers on Scopus by region (1998—2017)
Source: Elsevier

*Note: We speculate that the increase in AI papers in China around 2008 is a result of The National Medium- and Long-Term Program for Science and Technology Development (2006— 2020), and other government programs that provide funding and a range of incentive policies for AI research. Similarly, FP7 (2007—2013) and other science and technology research programs in Europe may have contributed to the small uptick in papers around 2008—2010.*

From "The AI Index 2018 Annual Report." See End Note 213 for full reference.

For example, the second feature of AI is that it is a dual-use technology. AI systems can be designed and used in ways that support both civilian and military ends. Dual-use technologies are particularly hard to regulate given competing desires to both encourage benefits and prevent harms. Research intended to make AI systems more resilient against attack also highlights vulnerabilities to cyber criminals. While AI is enabling more automated defense systems, it is simultaneously enabling more sophisticated attacks.

A third feature of AI is that its forms of reasoning are markedly different from our own and can be difficult to understand and control. AI systems operate at a speed, scale, and level of interaction that no human can comprehend. The use of algorithms in financial trading highlights this phenomenon.[22] Even though each algorithm individually is comprehensible, their collective actions have quickly overpowered human expertise and understanding, transforming the global economy and leading to crashes that no one can account for. Moreover, machine learning is significantly less comprehensible as these systems reach conclusions on their own for reasons that are generally not decipherable. The lack of interpretability poses new questions about accountability that differentiates AI from many other consequential emerging technologies.

When considered together, these features of AI systems may result in the transformation of many aspects of our world: from the global economy, to the ways we communicate, govern, and provide services. Ensuring that this transformation takes place in a way that is not unduly dangerous or harmful should be an immediate global policy priority.

# AI Security

AI security is defined in this report as the robustness and resiliency of AI systems, as well as the social, political, and economic systems with which AI interacts. There are other terms currently in use that address elements of the AI security space. For example, "AI safety" has become a fairly well-formed field that has primarily focused on technical problems and solutions. Other domains, such as AI policy, AI ethics, and AI governance, highlight social and political implications. AI security supplements these approaches by addressing the intersection of technical safety, geopolitics and political economy, and social and institutional resiliency. AI security can be a helpful frame for policymakers and decision-makers to investigate significant systems-level opportunities and threats posed by AI.

While the word "security" is often associated with national security, this depiction can leave out many modern dangers and threats. Scholars have long argued for the need to broaden security to include concepts such as "cybersecurity", "economic security", and "environmental security" to more precisely describe the range of pressing threats a nation faces. The use of the word in this report references a landscape in which AI is both a pervasive and disruptive force throughout the world.

## AI SECURITY MAP

The report introduces an AI Security Map to help navigate the broad domain of AI security (see Figure I). The map provides a simplified overview of the domains in which AI presents threats and opportunities, including: 1) Digital / Physical, 2) Political, 3) Economic, and 4) Social. The map is used first as a way to visually represent key domains and topics relevant to AI security. Later the map is used as a comparative tool to highlight which topics are addressed by different actors.

The topics that fall within each domain are described below using recent real-world examples to dissuade the assumption that the security implications of AI are not yet a concern. It is a goal of this report to highlight ways in which AI security is already a critical policy consideration, even as the scale and scope of the threats and opportunities will change over time.

The map portrays digital / physical as a single security domain because the interconnection between digital and physical systems has made it increasingly difficult to draw a meaningful

Figure I. AI Security Domains

| DIGITAL / PHYSICAL | POLITICAL | ECONOMIC | SOCIAL |
|---|---|---|---|
| RELIABLE, VALUE-ALIGNED AI SYSTEMS | PROTECTION FROM DISINFORMATION AND MANIPULATION | MITIGATION OF LABOR DISPLACEMENT | TRANSPARENCY AND ACCOUNTABILITY |
| AI SYSTEMS THAT ARE ROBUST AGAINST ATTACK | GOVERNMENT EXPERTISE IN AI AND DIGITAL INFRASTRUCTURE | PROMOTION OF AI RESEARCH AND DEVELOPMENT | PRIVACY AND DATA RIGHTS |
| PROTECTION FROM THE MALICIOUS USE OF AI AND AUTOMATED CYBERATTACKS | GEOPOLITICAL STRATEGY AND INTERNATIONAL COLLABORATION | UPDATED TRAINING AND EDUCATION RESOURCES | ETHICS, FAIRNESS, JUSTICE, DIGNITY |
| SECURE CONVERGENCE / INTEGRATION OF AI WITH OTHER TECHNOLOGIES (BIO, NUCLEAR, ETC.) | CHECKS AGAINST SURVEILLANCE, CONTROL, AND ABUSE OF POWER | REDUCED INEQUALITIES | HUMAN RIGHTS |
| RESPONSIBLE AND ETHICAL USE OF AI IN WARFARE AND THE MILITARY | PRIVATE-PUBLIC PARTNERSHIPS AND COLLABORATION | SUPPORT FOR SMALL BUSINESSES AND MARKET COMPETITION | SUSTAINABILITY AND ECOLOGY |

boundary around threats that occur in one space or the other. For example, both digital and physical security are implicated if a criminal gains access to the code of a Tesla Model 3 and causes a traffic accident, or if the system monitoring military drone footage mischaracterizes a conflict site. The inclusion of economic and social domains in this model also differentiates the analysis from earlier accounts of AI security.[23] When an AI system is developed or used in such a way that it leads to an attack or the prevention of an attack against these systems, then it can reasonably be said to be a security concern. Rapid AI development is generating and exasperating vulnerabilities in social and economic systems whose resiliency is not assured.

Preparation across the four domains is critical to support the goals of national and global security. All of the domains are deeply interconnected, but each is explicitly named in order to draw attention to the breadth of the security landscape, and to help elucidate divergent priorities among nations as they advance AI strategies. (See *Figure I. AI Security Map*)

The map defines four primary domains of AI security: digital/physical, political, economic, and social. Five topics are listed within each of these domains, in no particular order. This is not an exhaustive list. The map is a starting place that aims to represent major themes in the AI security landscape.

It is not the point of the map to suggest that every actor should consider every topic. Indeed, many governments are explicitly trying to carve out an area of expertise in the AI ecosystem as a means to gain a competitive edge. For example, Canada has focused on attracting AI talent, Germany has focused on AI for manufacturing, and the UK has focused on ethical AI development and application. Nonetheless, the topics named are not hypothetical concerns, but current and pressing challenges. Failing to address them may not be a viable long-term strategy. Each actor will have to make choices; the security map can help elucidate the range of options.

Importantly, there are other models for categorizing elements of AI policy and security. A recent paper from CIFAR, for example, provides a framework for analyzing differences among national AI strategies across eight areas of public policy,[24] including scientific research, AI talent development, skills and the future of work, industrialization of AI technologies, ethical AI standards, data and digital infrastructure, AI in the government, and inclusion and social well-being.

Another way to categorize AI threats is by the context of their emergence. For example, threats can stem from accidents, from deliberate misuse, or from the political or economic context in which actors and technologies interact.[25] There are not clear boundaries around these categories, but they can help elucidate key concerns. For example, the threat of oppressive surveillance regimes powered by AI systems may represent an example of misuse of AI technologies, but surveillance regimes are also inherently intertwined with the political and economic landscapes they are situated within.

Extreme threats can also be categorized as rising to the level of either catastrophic or existential risk.[26] Catastrophic risks cause damage to enormous numbers of people at a global scale, while existential risks have the potential to eliminate all or a majority of humanity. AI may reasonably pose both categories of threats,[27] and some AI governance work prioritizes attention to extreme threats given their potential devastating impact, the need to prepare in advance, and the market's failure to address such risks.[28] Every topic listed in the AI Security Map could lead directly or indirectly to catastrophic risks, and a few topics could generate existential risks.

The following section briefly describes each topic within the four AI security domains along with real-world examples. It also notes potential future shifts to identify ways in which the nature of the AI landscape is likely to change over time and how this could intensify the need for more coordinated global AI policy.

## 1. DIGITAL / PHYSICAL DOMAIN

### Reliable, Value Aligned AI Systems

Reliability is a primary consideration for the safe use of AI in the real world. Current AI systems make mistakes. These mistakes may occur because of development or training errors, because the behavior of the system is not aligned with the operator's intentions, or because the system figures out a way to achieve its goal in an unintended way. To mitigate these risks, AI safety researchers work on topics such as specification, robustness, and assurance to help ensure their AI models do only what they are intended to do and can prove that is the case.[29] There are numerous examples of AI systems making mistakes or acting in strange ways, both in the lab and in real life. Some of the impacts of these errors are significantly worse than others.

In March 2018, an Uber vehicle that was testing its autonomous mode hit and killed a woman named Elaine Herzberg who was walking her bike across a road in Tempe, Arizona.[30] This was the first lethal accident resulting from an autonomous system-design failure of a self-driving vehicle. A subsequent report on the crash released by America's National Transportation Safety Board stated that the system failed to classify the object correctly, changing the classification several times before attempting to engage emergency braking just 1.3 seconds before impact. Unfortunately this feature had been disengaged by Uber for testing purposes. In response to the tragedy of this young woman's death, Uber suspended all testing of its autonomous vehicles.

AI systems can exhibit other kinds of unexpected behaviors, for example when a system discovers unintended "solutions" to optimize its goal. In June 2017, Facebook revealed that two of its AI bots had developed their own language to communicate with each other in shorthand.[31] Since the bots were designed to talk with humans, this hack for efficiency made the bots unusable, and they were turned off.

Meanwhile, in November 2016 at the China Hi-Tech Fair in Shenzhen, a robot designed for kids called Xiao Pang (Little Fatty) suddenly began repeatedly ramming itself into a display booth for unknown reasons.[32] Broken glass from the booth went flying around the room and landed one man in the hospital. Another example comes from OpenAI, where researchers trained an AI agent on a boat racing videogame, but found that the agent figured out that by repeatedly turning in circles in a precisely timed way it could earn more points than finishing the course, despite repeatedly catching on fire, crashing into others, and going in the wrong direction.[33]

While these examples may not seem overtly consequential, they point to underlying issues that could be greatly magnified in the near future with more widespread AI deployment. The temptations of efficiency may also lead to people being taken "out of the loop" of complex decision-making, for example for certain military, manufacturing, legal, or medical decisions. However, the fallibility of AI systems means that harmful mistakes are likely. If we do not institute sufficient monitoring and oversight mechanisms, these mistakes will not only be more significant, but also harder to catch and fix. As the capabilities and generalities of AI systems improve, important considerations will include designing systems so that they do not resist being turned off,[34] designing appropriate objectives for AI systems,[35] and aligning the goals of systems with those of the people they interact with.[36]

## AI Systems that are Robust Against Attack

Many machine learning models are currently susceptible to attacks and have vulnerabilities that may prohibit their widespread adoption. Adversarial machine learning is the process of identifying and exploiting vulnerabilities within AI systems to cause mistakes or a change of behavior. For example, making small perturbations to the pixels of an image can cause machine learning models to mistake the image for something else. Other adversarial attacks include poisoning training data or altering a learning algorithm.

In December 2017, a team of MIT computer science students manipulated the pixels of a picture of machine guns and convinced Google's Cloud Vision AI program that it was an image of a helicopter.[37] The change was imperceptible to humans and the students carried it out without knowing the code of the AI program. Another worrisome example came from graduate students at the University of Washington, who trained a deep neural network to recognize road signs and then found ways to confuse it. For example, adding just a few black and white stickers to a stop sign tricked the algorithm into thinking it was a 45mph speed limit sign.[38] Although these instances were carried out in an academic setting and may be relatively impractical for criminals, the fact that they worked highlights the need to monitor the possibility of such instances.

Training AI systems with adversarial examples is also used as a security mechanism to test the robustness of models, and to help protect systems from attacks. This practice is gaining popularity, and several AI researchers have created an open-source library of adversarial examples for this purpose.[39] Other research has made progress toward universal protection against well-defined classes of adversaries.[40]

As our homes and cities become increasingly reliant on AI-enabled connected systems, nefarious actors are likely to find new avenues through which to breed fear, cause accidents, and siphon funds. Improving the robustness of AI systems will be an ongoing effort. Digital infrastructure may need increasingly automated defenses to protect AI systems from attack and manipulation.

## Protection from the Malicious Use of AI and Automated Cyberattacks

Cyberattacks are just one example of the malicious use of AI, which refers to the intentional use of AI to cause harm. For years, machine learning models have allowed cyber criminals to solve discrete problems such as quickly defeating CAPTCHA systems and testing stolen usernames and passwords across hundreds of sites.[41] AI changes the scale at which cyberattacks can occur, which is particularly damaging if more tailored and sophisticated spear phishing attacks are released autonomously. AI can also be used to exploit human vulnerabilities, for example by using a chat bot to uncover personal information or to sway behavior.

The 2016 DARPA Cyber Grand Challenge was the first fully autonomous cyber hacking tournament, and highlighted the new reality that AI would propel the automation of both attacks and defense in cyberspace. AI cyberattacks are no longer hypothetical: cybersecurity firm Darktrace Inc. reported that one of its client companies was the victim of an attack that used a simple machine learning model to learn the patterns of how users were behaving inside a network, and then mimicked their behavior to go undetected.[42] In 2016, the company ZeroFOX Inc. built a neural network that was able to create targeted phishing messages after reviewing individuals' Twitter posts.[43]

The scale and speed at which AI-powered cyberattacks can occur may increasingly pressure cybersecurity vendors to offer AI-powered cyber defenses. The work of cybersecurity professionals will still be needed for a long time to come, but people must also consider how to institute the right processes and communication channels to enhance and integrate their work with that of the AI defense systems. Nations seeking protection from malicious uses of AI will require more coordinated information sharing, white-hat cybersecurity researchers, and enforceable standards for technology companies developing AI-enabled services, robots, and toys.

## Secure Convergence / Integration of AI with Other Technologies (Bio, Nuclear, Etc.)

AI developments interact with the development of other technologies, and these convergences offer new opportunities and threats. For example, the use of AI in military decision-making could be destabilizing to the nuclear strategic balance if an AI system makes a mistake and mischaracterizes the nature of a threat.[44] There has been a convergence of AI with many other technological advances as well, from robotics and blockchain to bioengineering and aerospace. The combinations of these technologies are more powerful, but also more dangerous and harder to control.

One example of consequential AI convergence is in synthetic biology (synbio), a field that combines engineering principles of design and fabrication with biological components such as DNA, enabling the creation of new biological organisms that do not exist in the natural world. As in other biotechnological fields, AI is proving to be a useful tool for sifting through and analyzing billions of base pairs, and for identifying new possibilities, uncovering potential genomic secrets and generating entirely new life forms.[45] Synbio is mostly being used for benign and beneficial purposes such as the creation of naturally replicating rubber. But AI-powered bioengineering could lead to unanticipated accidents, or fall into the wrong hands. A June 2018 US National Academy of Sciences report warns that synbio expands the risks of bioweapons because new or more virulent pathogens could be created from scratch.[46]

Although DNA synthesis and printing techniques have been around for decades, future advances will update these methods, potentially allowing DNA printing machines to scale in much the same way that 3D printing machines already have. This shift will pose new challenges, such as the possibility to use machine learning models to design novel pathogens, and have them printed anywhere in the world. Other technological convergences will similarly expand the AI security landscape.

## Responsible and Ethical Use of AI in Warfare and the Military

There are numerous uses for AI in the military, but the boundaries around what constitutes acceptable uses are highly contentious. The issue of lethal autonomous weapon systems (LAWS) continues to be discussed at the international level under the United Nations Convention on Certain Conventional Weapons (CCW) by the Group of Governmental Experts (GGE). Annual meetings since 2013 have brought together representatives from dozens of countries to consider the possibility of an international ban on LAWS, a position officially supported by at least 26 countries.[47] The CCW process requires full consensus, however, and while a majority of states favor moving toward a prohibition, five key states—the United States, Australia, Israel, South Korea, and Russia—have opposed a ban. Deliberations will continue throughout 2019.

In the meantime, weapon systems with certain degrees of automation are already in use. Israel Aerospace Industries has developed a warhead missile nicknamed Harpy that detects and attacks autonomously; Harpy has already been sold to the Air Forces of several countries.[48] The French company Dassault Aviation has a highly autonomous combat air system with attack capabilities called NEURON.[49] And BAE Systems, based in the United Kingdom, has developed Taranis, an advanced armed drone that can identify and target threats, although it is designed to seek verification by a human operator.[50]

Other, non-lethal, applications of AI in the military also require consideration. For example, the US Department of Defense (DoD) has a program called Project Maven that uses computer-vision machine learning to identify objects of interest from vast amounts of video footage from drones and other sources.[51] The DoD has developed the Unmanned Systems Integrated Roadmap 2017-2042, a strategic plan that cites autonomy, human-machine collaboration, and network security as critical themes.[52]

The DoD is currently developing a set of AI principles to guide the ethical and responsible use of AI in the military,[53] and establishing clear norms internationally will become a more important policy priority in coming years. We are nearing a time when the capacity to build "killer robots" requires little more than off-the-shelf technology and ill intent, which could enable massively distributed weapon systems that are resistant to traditional national defenses. Moreover, software can be programed to target particular kinds of people, or to make conclusions about acceptable targets based upon data it receives. As governments navigate these considerations, more immediate concerns will weigh on militaries, such as whether, where, and how to introduce autonomy into the chain of command to counter inefficiencies and data backlogs. Despite much enthusiasm, fundamental technical and logistical challenges to the military use of AI are likely to present roadblocks for years to come.[54]

## 2. POLITICAL DOMAIN

### Protection from Disinformation and Manipulation

We have seen the impact of the widespread distribution of misleading information over social media networks.[55] AI technologies can support these disinformation campaigns, enabling them to become more targeted and effective and greater in scale.[56] These campaigns are capable of inciting fear, instability, and hatred, which can be used to undermine democratic systems and processes or to target minority populations. The spread of disinformation can lead to many harmful outcomes, including widespread loss of trust in media communications.[57]

For example, in the lead up to the 2016 presidential election, it was discovered that social media bots—spam accounts that post autonomously using preprogrammed scripts—accounted for a surprisingly high percentage of posts. Between the first two presidential debates, for example, the Atlantic reported that a third of pro-Trump tweets and nearly a fifth of pro-Clinton tweets were generated by fake, automated accounts.[58] A March 2017 study from the University of Southern California and Indiana University found that as many as 48 million Twitter accounts—between 9% and 15% of all active accounts—do not belong to real people, and they referred to this as a "conservative estimate".[59] The study also reported that bots of this kind are not only used to build political support, but also to promote terrorist propaganda and recruitment.

AI capabilities have also made it easier to manipulate videos, which has intensified concerns about trust in communications. In late 2017, "deepfake" videos emerged in which AI had been used to transfer the faces of celebrities into pornographic films. This technology was later picked up by the industry to allow customers to personalize video content.[60]

The personalization of the digital realm makes it ripe for individual manipulation. Many media platforms already use AI algorithms to optimize content to keep the attention of users.[61] Growing awareness of our unwitting participation in this "attention economy" is unlikely to change the business models of companies that rely on selling advertisements. As AI tools also interact with augmented and virtual reality, we will only be faced with a greater number of spaces that vie for our attention and encourage particular behaviors.[62]

## Government Expertise in AI and Digital Infrastructure

AI systems can help governments manage administrative burdens and resource constraints, for example by automating data entry, optimizing scheduling and planning, and providing support with customer service.[63]. However, in the absence of uniform safety and efficacy standards for AI models, finding safe, reliable, and fair AI tools is as much a challenge as obtaining the tools themselves. Governments are trying to balance the goals of embracing and benefitting from this technological advance, while also considering appropriate policy environments for AI development and use.

Oxford Insights created a “Government AI Readiness Index” that measured how prepared national governments in the Organisation for Economic Co-operation and Development (OECD) are to take advantage of the benefits of automation.[64] The index looks at metrics such as digital skills, government innovation, and data capabilities. The report found the UK government to be the most prepared, and the US government to be second. Despite some significant advantages, the US government has generally struggled to promote tech literacy among agencies and members of government.[65] There is a shortage of AI experts, and attempts to woo top AI talent have led to dramatic rises in industry salaries, luring people away from the public sector.[66] In September 2018, the Artificial Intelligence in Government Act was introduced by a group of bipartisan US senators as an acknowledgment of the need to improve the use of AI across the federal government.[67] The bill seeks to achieve this by providing resources and directing federal agencies to include AI in data-related planning.

There is increased government attention to AI around the world, and many governments are eager to implement AI tools themselves. The government of the United Arab Emirates not only launched an AI strategy, but also appointed the world’s first State Minister for Artificial Intelligence in October 2017.[68] Part of this Minister’s role—and a key theme of the government’s strategy—is to improve government knowledge of AI, for example through field visits for government officials to technology firms, an AI camp to teach technical basics, and other initiatives aimed at fostering learning for government officials.

Another important example of expanding technical competency in government comes from Denmark, where the Danish government created a tech ambassador program to engage in “techplomacy” with industry leaders around the world.[69] The first tech ambassador, Casper Klynge, has an office in Silicon Valley and engages with many of the world’s largest technology companies in recognition of their importance in the international arena.

Implementing technological solutions in government services without sufficient oversight and support can easily go awry, as the initial United States healthcare.gov rollout challenges in 2013 demonstrated. Similar mistakes could be significantly more costly and damaging for an AI-based tool. To help avoid such scenarios, some governments are now working with the World Economic Forum and the International Organization for Public-Private Cooperation to design AI procurement policies; the UK government was the first to join this partnership.[70] A representative from the Department for Digital, Culture, Media and Sport is working with the two organizations to identify ways to shape AI standards throughout the country. In the future, representatives throughout government will need to engage more deeply with the appropriate role and impact of technologies, including AI. The purchasing power of government is an example of a core strength that can be leveraged to encourage the adoption of safety, reliability, and fairness standards.

## Geopolitical Strategy and International Collaboration

National competition and the perception of "AI race" dynamics may have a negative impact on diplomatic efforts. Moreover, if one country develops significant advances in AI technologies, the country may gain access to economic and political advantages and not have a natural incentive to share its capabilities or resources in the absence of pre-existing international agreements. However, this is not inevitable. Governments also have an incentive to encourage coordination to promote international trade as well as to share resources such as databases, platforms, and talent to improve the capabilities and reach of their AI and digital ecosystems.

Some statements from government leaders suggest the importance of becoming the global leader in AI. During a 2017 speech to students in Moscow, President Vladimir Putin said, "Artificial intelligence is the future not only of Russia but of all of mankind. There are huge opportunities, but also threats that are difficult to foresee today. Whoever becomes the leader in this sphere will become the ruler of the world." And in a 2017 report released by China's State Council, the Chinese government set the intention of becoming "the world's premier artificial intelligence innovation center" by 2030.[71] Nonetheless, the Chinese government is actively pursuing international cooperation on AI development.[72]

While scholars have discussed the concerning rise of AI nationalism and national competition, [73] there are many signs of bilateral agreements, declarations of cooperation, and work in the direction of global governance.[74] In the future, as leaders face widespread threats from more advanced AI systems, states are likely to become more interested in establishing norms of global coordination and cooperation.[75]

## Checks Against Surveillance, Control, and Abuse of Power

AI simplifies the analysis of big data, enabling the processing of information, images, and audio at a greater scale. One implication of this capability is that it becomes easier to monitor behavior and discussion throughout a society or community. AI systems can identify trends, abnormalities, and imminent dangers and alert the relevant authorities. Some systems are now also capable of real-time facial recognition. In authoritarian regimes, AI may serve as a relatively inexpensive tool through which to adopt mass surveillance of populations to assert the stability and control of the political system. However, any political regime may be tempted by the lure of this deep view into the behavior of its people.

Numerous recent examples highlight the surveillance capabilities of AI systems. The Chinese national monitoring program known as Skynet has established facial recognition technology across at least sixteen cities and provinces; this system can purportedly scan the country's entire population in one second.[76] A Japanese security camera called the "AI Guardman" tracks shoppers and monitors body language to identify "suspicious behavior."[77] The US Department of Immigration & Customs Enforcement is tracking the social media activity of visa holders to assess potential threats.[78]

Numerous law enforcement agencies in the United States have also used predictive policing to identify potential criminal activity.[79] With the addition of more capable machine learning systems, these programs may become more expansive and controversial, demanding even greater attention. Policy considerations related to predictive policing include addressing rising threats to civil liberties, as well as ensuring these systems do not disproportionately target historically over-policed communities.[80]

## Private-Public Partnerships and Collaboration

Much of the current development of AI is happening within private companies. Nonetheless, governments around the world are investing in AI to support their priorities. Coordination between firms and governments on emerging technology has a long, varied, and nuanced history, which is now playing out for AI development in interesting ways. For example, a 2018 report from the US Department of Homeland Security stresses the importance of private-sector cooperation to achieve national objectives, but expresses concern about disputes by private employees over the use of technology for national intelligence and defense purposes.[81]

Several recent examples in the United States highlight how differences in values over the ethical use of AI can drive a wedge between industry and government. The first focused on Amazon's facial recognition program "Rekognition," which shares information with some law enforcement agencies. A letter from employees at Amazon read, "As ethically concerned Amazonians, we demand a choice in what we build, and a say in how it is used." The letter cited mass deportations by US Immigration and Customs Enforcement and the targeting of black activists by police officers, saying we "refuse to contribute to tools that violate human rights."[82]

Google employees also protested uses of AI technologies by the US government and demanded that their company pull out of a contract to help the Department of Defense analyze drone footage.[83] About 4,000 Google employees signed a letter demanding "a clear policy stating that neither Google nor its contractors will ever build warfare technology." One Google AI researcher wrote, "Google should not be in the business of war." Following mounting pressure from external groups and dozens of resignations, the company announced it would not renew the contract. Moreover, Google CEO Sundar Pichai published a set of AI Principles, which stated that Google would not pursue applications of AI that are likely to cause harm or injury, including weapons.[84]

The European Union has taken a proactive step to help establish shared norms among public and private actors in the AI ecosystem: the European Commission created a High Level Group on Artificial Intelligence that includes 52 experts from industry, academia, and civil society.[85] The Group has provided recommendations on the ethical, legal, and societal issues related to AI in support of the European strategy on AI.[86] Another important example is the International Panel on Artificial Intelligence (IPAI), launched in December 2018 by Canadian Prime Minister Trudeau and French President Macron to facilitate international and multistakeholder collaboration to promote the vision of "human-centric" AI.[87] These and other such initiatives will ideally foster trust and communication among diverse actors, which can help mitigate reputational harm to companies, while also providing guidance to governments on responsible uses of AI that are less likely to cause public backlash.

## 3. ECONOMIC DOMAIN

### Mitigation of Labor Displacement

The growth of robotics over the last decade has contributed to technological unemployment, the loss of jobs due to technological change and automation. Although technological change has altered labor markets for centuries, some projections indicate that AI could significantly increase job loss in the near future. For example, the OECD has estimated that expanding capabilities in robotics and AI will cause 14 percent of jobs in advanced economies to be susceptible to automation and another 32 percent to change significantly, with disproportionate risk for low-skilled people and youth.[88] A 2013 Oxford study found that 47 percent of jobs in the United States are at risk of automation in the next few decades.[89] And a 2017 McKinsey report found that between 400 million and 800 million jobs worldwide could be automated by 2030.[90] While others have argued that the growth of new job opportunities will mitigate these losses, the threat of labor displacement at this scale could be extremely damaging. Concern about job loss is widespread globally, though there are some variations in how the challenge is perceived. A 2018 Pew Research Center study of public opinion across 10 countries found that large majorities believe robots and computers will do much of the work currently done by humans within the next fifty years, and that it will be hard to find work due to automation.[91] The study found widespread skepticism about the potential economic benefits of automation.

We have already seen prominent examples of technological unemployment. In 2016, Apple and Samsung supplier Foxconn replaced 60,000 factory workers with robots.[92] In 2015, a factory in Dongguan, China that produces mobile phones replaced 90 percent of its workers with robots, going from 650 employees to 60.[93] Reportedly, the change led to a 250 percent rise in production and the company predicts needing even fewer employees in the future. The International Federation of Robotics has found robot density to be rising globally, with Asia experiencing the highest growth rate.[94] South Korea has the highest ratio of industrial robots to employees, with more than 600 robots per 10,000 employees in the manufacturing industry.

Nonetheless, automation will not be a straightforward trajectory. When electric car company Tesla failed to hit its production goals for its latest car model, CEO Elon Musk partially blamed over-automation of his factory, pointing out that humans are significantly better at adaptability.[95] Hypothetically, advances in machine learning could enable greater adaptability of robots, but there may be significant trade-offs, for example inconsistencies and less controllability. In the meantime, AI tools will improve the functionality of many automated systems, and an increasing number of communities and nations will need to explore options to ease labor transitions for millions of people. These will likely include policies such as advanced retraining programs, lifelong learning programs, and universal basic income or other distributed benefits. Some people believe the impact of automation may be so profound that the majority of people will no longer work at all; projections of those futures range from a widespread loss of meaning, to inspiring explorations of true personal ambitions.[96]

## Promotion of AI Research and Development

Advances in artificial intelligence promise to fuel substantial economic growth globally, but this outcome will not be realized without sufficient investment in research and development. China increased funding of research and development (R&D) for AI by 200 percent between 2000 and 2015, and was projected to overtake the United States in investment by the end of 2018.[97] Many countries around the world are exploring mechanisms to increase funding of AI R&D. In some cases, for example in Germany, this includes establishing R&D institutions modeled on the US DARPA to accelerate the study of disruptive technologies with important defense and security implications.[98]

In the United States, the AI R&D landscape is complex; the US government is not the only major contributor, as it shares the ecosystem with a wide range of industries and organizations. In a 2016 national strategic plan on AI R&D, the US National Science and Technology Council and the Networking and Information Technology Research and Development (NITRD) Subcommittee defined seven strategies to enhance AI R&D. The first of these was to make long-term investments in AI research, an area that can be challenging for the private sector to address on its own. However, thus far under the Trump Administration, these plans have not received much attention. A 2018 US Congressional White Paper on AI highlighted a sense of concern, noting "the United States needs to increase its R&D spending to remain competitive in the field of AI." In September 2018, NITRD put out a Request for Information to update the AI R&D strategic plan and a new national strategy is expected Spring 2019.

The potential economic upsides of AI are motivating more governments to support AI research and development. National AI investment should ideally be directed towards areas of potential market failure for AI, to increase the incentives for working on issues such as safety, ethics, sustainability, and long-term planning.

## Updated Training and Education Resources

To propel the development of AI technologies, many countries are interested in expanding and improving the opportunities for training and education in AI research. Simultaneously, there is growing acknowledgment of the need to revamp educational opportunities for a world in which automation is playing a greater role. This includes not only STEM fields (science, technology, engineering, and math), but also social sciences and the humanities, as it has become more apparent that so-called "soft skills" may be more uniquely human and less prone to automation.[99]

In recognition of the shortage of AI researchers, The French government intends to triple the number of people trained in AI over the next three years, both by helping existing educational programs in the country to refocus on AI and by establishing new programs and courses specifically designed to teach AI skills to more people. In addition, the country hopes to establish a network of four to six interdisciplinary institutes for AI at universities across the country.[100]

Automation could eliminate and alter many current jobs and people will need to adjust and retrain. However, individual solutions are unlikely to meet demand in the future, so policy responses will probably be required. There are already numerous policy initiatives and proposals to address these needs. For example, in January 2016, Singapore established a system whereby people over the age of 25 can receive $500 worth of "SkillsFuture credits" to pay for courses or training in thousands of areas.[101] The Aspen Institute has also proposed the creation of Lifelong Learning and Training Accounts, where individuals could make pre-tax payments to match government contributions over the course of one's career to be taken advantage of in times of need.[102] If AI developments lead to a greater number of jobs being automated, there will be increased pressure on governments to provide a range of educational and training resources to their citizens.

## Reduced Inequalities

AI intensifies several dynamics that contribute to inequality. While these outcomes are not inevitable, they are likely in the absence of interventions. First, the network effects associated with digital platforms have resulted in a relatively small number of leaders in the space, and a concentration of AI-generated wealth within specific companies, cities, states, and countries.[103] Second, the high salaries of AI specialists and valuations of AI startups have contributed to an increase in housing and living costs in the cities where the companies are located, exacerbating inequalities among residents. Third, AI is automating certain kinds of tasks and jobs faster and more completely than others; jobs that are menial, repetitive, administrative, or entry-level tend to be more vulnerable to automation, which may prevent traditional routes to social mobility. Moreover, those with more resources to fall back on are more likely to be able to retrain and find new sources of income, leaving those without such resources even further behind.

Silicon Valley is something of a test case for the role of technological advances in expanding inequality among residents. The region is home to many leading AI and technology companies, including Google, which rebranded as "AI-first" in 2017.[104] Technological advancements (including AI) have significantly contributed to the wealth of Silicon Valley. However, those benefits have not consistently "trickled down," as the region has some of the worst income inequality in the United States. Families on the upper end of the income spectrum in the San Francisco metro area make 11 times more than those on the low end, and nearly one third of Silicon Valley households cannot meet their basic needs without assistance.[105]

If current trends continue and the lion's share of AI-generated wealth is concentrated in developed countries with existing AI R&D infrastructure, emerging economies could be left even further behind. AI pioneer Kai-Fu Lee predicts that emerging economies will not be able to rely upon the models that have been central to economic growth in China and India since

AI systems will automate many of the tasks involved in the manual labor of factories and the cognitive labor of call centers.[106]

### Support for Small Businesses and Market Competition

Data is a major driving force of AI development, but access to data is not evenly shared. Rather, the world's biggest AI technology companies are seen as quasi data monopolies.[107] The phenomenon is self-reinforcing because having access to data supports the development of better products, which leads to more users, who provide more data. In 2015, Google had 75 percent of the market share in Internet searches.[108] The Big 5 companies in the United States—Apple, Alphabet, Amazon, Facebook, and Microsoft—as well as the Chinese giants, Tencent, Alibaba, and Baidu, have a key advantage over start-ups, small businesses, and would-be market competitors. Concern about the impact of a single company controlling too much of the AI market has already led to calls for antitrust regulation.[109] Some governments have prioritized supporting small- and medium- sized enterprises with the goal of counteracting the trend of market monopolization.

Certain countries have begun pooling resources in an attempt to compete in the AI space. For example, European countries have established a more robust European data ecosystem to push back against the rise of industrial data monopolies and become more attractive to AI companies. The European Commission's Digital Single Market policy on AI reads, "It is essential to join forces in the European Union to stay at the forefront of this technological revolution, to ensure competitiveness and to shape the conditions for [AI's] development and use (ensuring respect of European values)."

## 4. SOCIAL DOMAIN

### Transparency and Accountability

AI systems have already been integrated into decision-making capacities such as reviewing loan applications, making medical diagnoses, and screening individuals for extra police or legal scrutiny. However, the basis of these decisions is often not transparent to outsiders, and may even be unclear to the programmers of the systems. Enabling greater transparency, for example via explainability (where the AI system is programmed to describe its decision making process in a way that is easily understandable to humans,)[110] is one consideration for improving the accountability of AI throughout society. However, understanding AI systems can be challenging, in part because source code, training data, and learning models are all unique elements that can influence algorithmic decisions.

Accountability also raises questions around applying legal standards to software, such as who is responsible for mistakes, and how can AI systems be designed to better align with legal and policy objectives?[111] Currently, law enforcement, immigration enforcement, and other key agencies are procuring, developing, and implementing algorithmic tools of varying capabilities without standardized practices in place to ensure transparency, oversight, or accountability.

One of the challenges of enabling greater accountability is that many algorithms are developed by private companies and are proprietary. For example, when Eric Loomis was charged with five criminal counts in 2013, an algorithmic risk assessment was included in his sentencing determination.[112] Loomis filed a motion for post-conviction relief, arguing that the use of the algorithmic risk assessment violated his due process rights because the algorithm is "a trade secret" that was not revealed to the judge or the jury. The case made its way to the Wisconsin Supreme Court, but his appeal was ultimately rejected.

AI systems offer a powerful mechanism to reduce administrative burden and minimize human bias in decision-making, and will only become more useful across an increasing number of domains. However, the utility of AI will be greatly diminished if people cannot trust the results of AI systems (for real or perceived reasons.) Few in industry think that these tools will be able to scale without sufficient explainability, which has already led to the development of new tools to help on this front.[113] There is also mounting pressure on public agencies to address accountability concerns in their use of these systems.[114]

## Privacy and Data Rights

Big data is a critical component of AI development, and access to more and higher quality datasets is a major priority for AI companies. However, the free flow of data comes into tension with individuals' and communities' concerns about privacy and control of personal information. May 2018 marked a landmark shift in the ecosystem of data collection and rights when a new European privacy law, the General Data Protection Regulation (GDPR), went into effect. This law requires companies to indicate how data is used, minimize what is kept, inform people what data they have and how it is being used, and explain the rationale behind automated decision-making processes. These are high standards for machine learning systems and the companies building and using them.[115] However, if people do not trust their information to be handled properly, it may limit their willingness to share. Europe is not alone in making privacy and data rights a major policy issue. The California Consumer Privacy Act of 2018 will go into effect January 1, 2020.[116] This law gives California residents four rights over their personal information, including to know what data is held and how it is being used, to opt out of having personal information sold to third parties, to request that a business delete personal information, and to not be discriminated against for exercising any of these rights.

Some people have grown wary of sharing personal information on digital platforms. The Facebook and Cambridge Analytica scandal in early 2018 showed that a third-party company was able to gain access to 50 million Facebook profiles and use the harvested data for mass targeting with the aim of swaying voter behavior.[117] In the aftermath, Facebook unveiled new privacy features that provide users with more visibility and control over what information third-party apps have access to. However, in September 2018, Facebook revealed that hackers had exploited a vulnerability in its code and had once again gained access to around 50 million accounts.[118]

AI is being integrated into a web of devices that promise to optimize many aspects of our lives, but industry desires for data interoperability must contend with significant concerns and regulations regarding privacy.[119] Some countries and regions are pursuing policies of data sovereignty with the expressed intention of helping to protect their citizens from data misuse and crime.[120] Organizations and governments are also starting to explore new models of data control, such as data trusts, through which data sharing can be streamlined while simultaneously providing users with greater control over the use of their data.[121]

## Ethics, Fairness, Justice, Dignity

The question of machine ethics is now at the center of public debates about AI and machine learning. While AI systems can introduce greater fairness into processes by taking more considerations into account and not falling prey to implicit biases and human error, they can also introduce and magnify prejudices by reproducing cultural biases or by training on skewed datasets. Bias is also introduced into systems through decisions about what tools to build, how, and for whom. Ultimately, machine bias is too easily hidden behind a veneer of objectivity. Questions of AI ethics, fairness, justice, and dignity have come to the fore due to a number of high profile incidents.

For example, a 2016 *ProPublica* report found that an algorithm used to determine recidivism rates in the United States was twice as likely to mislabel a black defendant than a white defendant as a likely future criminal, while white defendants were more often mislabeled as low risk than black defendants.[122] Another example comes from Google Translate. When the system is translating words from languages without gender pronouns, it is designed to guess whether the comment refers to a "he" or a "she". The guesses the software makes often highlight the gender bias it has absorbed: "he is a doctor, she is a nurse; he is a president, she is a nanny."[123]

Another notable example is the chatbot known as Tay that was introduced by Microsoft in 2016.[124] Tay was designed to talk like a teenage girl, but the system continued to update based upon conversations with people on Twitter. People quickly took advantage of the idea, goading Tay into saying increasingly hateful comments. While Tay's first tweet was "can I just say that im stoked to meet u? humans are super cool". Within a few hours, the bot was making comments like, "Hitler was right I hate the jews." The chatbot was quickly taken offline.

Researchers are working hard to address these issues. For example, in September 2018, Google released an open-source tool that both checks whether a dataset has skewed representation of any groups and helps indicate which factors are influencing the decision of the algorithm. However, while these kinds of tools are worthwhile, it is important to understand that data cannot simply be "neutralized," and so the problem can never be "solved". Instead, it is becoming increasingly important to institute processes to review and audit the effects of AI tools and to provide people with recourse for faulty or unwarranted decision-making. Another important shift is towards having increased racial, gender, socioeconomic, and other kinds of diversity on research and engineering teams to help flag blind spots in the development and use of AI.

## Human Rights

AI tools are being used to promote growth, wellbeing, and social good around the world, for example by supporting the advancement of the Sustainable Development Goals (SDGs).[125] However, AI tools can also be used in ways that curtail human rights. For example, facial recognition and other surveillance tools can be used to target people based on their physical appearance for additional screening by law enforcement or immigration authorities.[126]

Efforts to build "brain-computer interfaces" in order to augment human brains with AI could be another particularly insidious threat to human rights. For example, the company Neuralink is reportedly building a whole-brain interface to form a "digital tertiary layer" to our brains (to supplement the limbic system and the cortex).[127] Through this technology, people would be able to remember everything and access any information available on the internet, but they would also be able to communicate with their thoughts. It is hard to imagine the impact such a development would have, but some scholars have argued that the security, privacy, consent, agency, and identity implications of AI neurotechnologies will have a profound impact on our basic human rights.[128]

AI systems may or may not be designed with a value system that appreciates human rights. The actions we take now will shape AI trajectories and help determine the degree to which automated systems respect the rights of people around the world.[129] Advanced AI could additionally complicate a human rights framework, as we may face a future in which we share the world with intelligent machines.[130]

## Sustainability and Ecology

AI applications rely upon data centers for massive quantities of computational power, which could be cause for concern.[131] Data centers in the United States currently use more than 90 terawatt-hours of electricity per year, relying upon the equivalent of more than 34 large coal-fired power plants to provide this energy.[132] Globally, the problem is more profound, as data centers use roughly 416 terawatt-hours per year of electricity and some analyses project this number will triple in the next decade.[133] The production of digital devices is also unsustainable, as less than 16% of global e-waste is formally recycled.[134] The volume of e-waste is predicted to rise to 50 million metric tons or more every year.[135] In Asia, e-waste has increased in volume by approximately 63 percent since 2012. The use of AI chips could exacerbate the need for electronic manufacturing, which is the most carbon-intensive phase of manufacturing.

At the same time, AI offers numerous ways to help achieve environmental goals. For example, DeepMind was able to use AI tools to identify hidden inefficiencies in Google's data centers and reduce their energy consumption by 40 percent.[136] But reducing inefficiencies may only address a small portion of the problem, and we are likely to need more innovative solutions for sustainability in the long-term.

Some companies are leading the charge. Apple has made the transition to running its operations on 100 percent renewable energy, powering its data centers with numerous solar farms.[137] This initiative prompted two-dozen other companies in Apple's supply chain to pledge to do the same. These successes help expand local clean energy markets while making a strong business case for sustainable growth, and may inspire others to follow the same path. Policy leaders in France and the UK have also indicated the importance of sustainability and clean growth in their AI strategies, so we should expect more focus on the environmental threats of AI—and methods to mitigate them—in the future.[138] [139]

# Global AI Policy Responses

The previous section introduced twenty topics from across four domains of the AI security landscape. This section examines whether and how governments are preparing for these threats and opportunities, based upon an analysis of national AI strategies and policy documents from ten countries. It is worth noting that such strategies are not always fully implemented, and actions taken may not always have the desired effects. Further research is needed to assess the effectiveness of plans within each country.

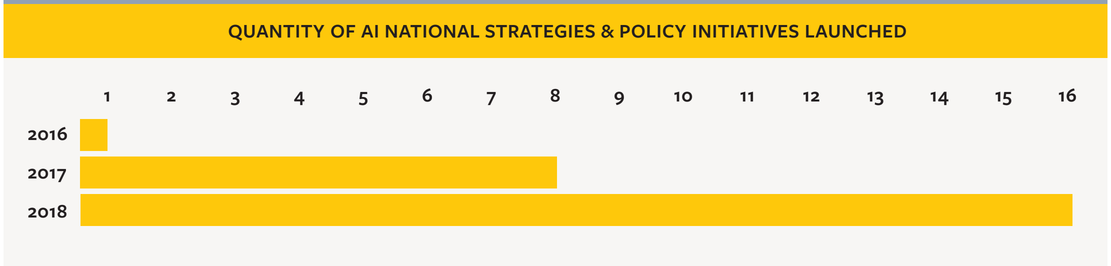
Since 2017, there has been a significant increase in government attention to AI, as indicated in Figure II. Countries that have defined explicit AI national strategies include: China, Canada, the United Kingdom, France, India, Japan, South Korea, Sweden, and the United Arab Emirates, while at least 27 governments have articulated plans or initiatives for encouraging and managing the development of AI technologies.[140]

Governments play an important role in shaping the landscape of AI development and use. Most national governments are not yet actively considering the introduction of AI-specific regulation, as they are opting instead to use existing regulatory frameworks across specific industries. However, governments are leveraging AI development through other policy mechanisms, such as supporting innovation with investment in infrastructure, encouraging education and training with federal grants, supporting fundamental science research, and promoting standards through procurement policies. National strategies additionally serve to identify longer-term plans, articulate national values, and enable coordination over shared visions.

Government funding of the AI ecosystem has increased rapidly since 2016: the Canadian government has promised to invest US $98.7 million into AI R&D, and the UK government has promised US $22.3 million. China and the United States are outspending others: the Chinese government promised US $2.1 billion to build an AI industrial park in Beijing, while the US government spent more than $2 billion on AI R&D in 2017 alone.

Many technologists and industry leaders support a high level of government investment, but would rather minimize government regulation of AI development. This is far from universally true, however. For example, Microsoft President Brad Smith has called for thoughtful government regulation of facial recognition technology.[141] And technology titan Elon Musk has repeatedly argued for controls over artificial intelligence; as he said in an interview in July 2017, “AI is a rare case where we need to be proactive about regulation instead of reactive.”[142] A 2019 Google

**Figure II: Number of AI National Strategies & Policy Initiatives Launched 2016–2018**



**In 2016, the United States released several federal policy documents exploring the implications of AI. In 2017, Canada, China, Singapore, the UAE, Japan, Estonia, Ireland, and Finland all released AI strategies or launched AI policy initiatives. In 2018, the UK, France, India, South Korea, Denmark, Germany, Mexico, Australia, Sweden, Italy, Austria, Russia, New Zealand, Tunisia, and Kenya all released AI strategies or launched AI policy initiatives.**

White Paper also calls for government engagement in AI development, pointing out, "Some contentious uses of AI could have such a transformational effect on society that relying on companies alone to set standards is inappropriate—not because companies can't be trusted to be impartial and responsible, but because to delegate such decisions to companies would be undemocratic."[143]

Government responses to AI have thus far varied significantly. While many national strategies address similar topics, the diverse ways in which governments frame their aspirations and challenges reflect longstanding social, cultural, and political institutions and values. It is beyond the scope of this report to delve into these histories and contexts, but it is apparent that many national AI strategies resemble preexisting technology policies.

It is important to many policymakers that AI strategies be established within a framework of national values. However, data flows and security threats from AI systems do not always respect national boundaries. Moreover, the importance of multinational technology firms in shaping AI development may encourage globally compatible regulatory environments. As decision-makers look inward to assess domestic goals, they also need to track international developments in AI and the policies shaping its growth.

The first part of this section highlights national AI policies and strategies from four countries: China, France, the United Kingdom, and the United States (discussed in alphabetical order). These countries were chosen for three reasons: 1. They all have significant AI expertise and are actively pursuing substantial AI R&D plans. 2. They have all published reports outlining national policies and strategies for AI and have numerous government-led AI initiatives underway. 3.

There are sufficient differences between these countries' approaches to AI to provide an interesting cross-cultural comparison. (The Appendix includes an AI Policy Compendium outlining concrete policy proposals from each of the four countries.)

Six additional countries are also discussed, though in less detail. These six—Canada, India, Japan, Singapore, South Korea, and the UAE,—were chosen because they have also launched notable AI national strategies or government-led initiatives that highlight a broader array of approaches. Many additional countries have produced AI strategies, but are beyond the scope of this report.

This policy review is not comprehensive; for example, it does not explore existing regulations that intersect with AI, such as data and privacy regulations. Rather, the focus is on the relatively new phenomenon of strategies and policies that address AI explicitly and comprehensively (i.e. not just the use of AI in a single sector.) The AI Security Maps help highlight policy priorities across four prominent domains. While this analysis merely provides a snapshot of how AI challenges and threats have been framed in a sub-set of national strategies and policies, and does not represent all domestic activity such as industry investment and multistakeholder initiatives, the comparison highlights interesting areas of convergence and divergence, and provides a lens into how AI security is being framed around the world.

## CHINA

At the 19th National Congress of the Communist Party in October 2017, Chinese president Xi Jinping included artificial intelligence as part of his grand vision for the nation.[144] This reinforced the "New Generation Artificial Intelligence Development Plan," which was released by the State Council in July 2017 and outlined China's strategy to build a domestic AI industry worth nearly US$150 billion by 2030.[145] As described in the New Generation Artificial Intelligence Development Plan, China sees AI as a mechanism to "leapfrog development" and achieve "two hundred years goals and to provide strong support for the great rejuvenation of the nation." By 2020, China intends to keep up with the technology and application of AI; by 2025, China plans to achieve a major breakthrough in the basic science of AI; and by 2030, China hopes to be the leading AI innovation center of the world.

The New Generation AI Development Plan suggests that, after 60 years of development, artificial intelligence has entered a new stage and "will profoundly change human social life and the world." Among the changes the State Council expects are transformations in industry, eco-

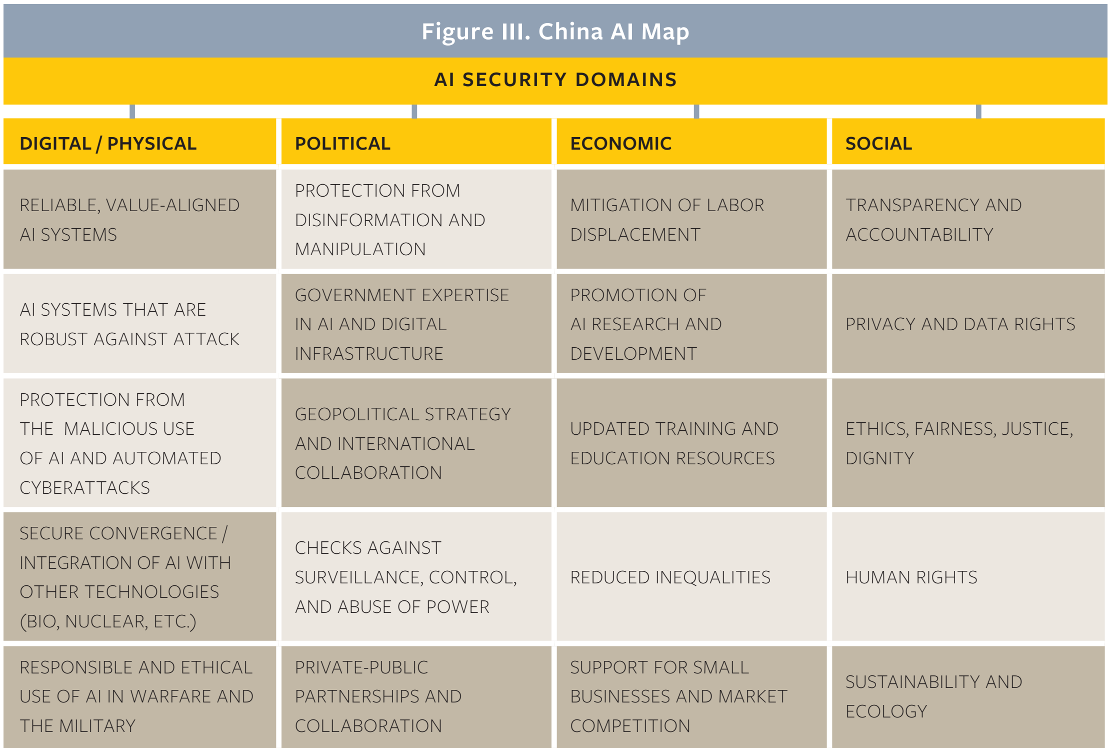


**Figure III. China AI Map**

AI SECURITY DOMAINS

| DIGITAL / PHYSICAL | POLITICAL | ECONOMIC | SOCIAL |
|---|---|---|---|
| RELIABLE, VALUE-ALIGNED AI SYSTEMS | PROTECTION FROM DISINFORMATION AND MANIPULATION | MITIGATION OF LABOR DISPLACEMENT | TRANSPARENCY AND ACCOUNTABILITY |
| AI SYSTEMS THAT ARE ROBUST AGAINST ATTACK | GOVERNMENT EXPERTISE IN AI AND DIGITAL INFRASTRUCTURE | PROMOTION OF AI RESEARCH AND DEVELOPMENT | PRIVACY AND DATA RIGHTS |
| PROTECTION FROM THE MALICIOUS USE OF AI AND AUTOMATED CYBERATTACKS | GEOPOLITICAL STRATEGY AND INTERNATIONAL COLLABORATION | UPDATED TRAINING AND EDUCATION RESOURCES | ETHICS, FAIRNESS, JUSTICE, DIGNITY |
| SECURE CONVERGENCE / INTEGRATION OF AI WITH OTHER TECHNOLOGIES (BIO, NUCLEAR, ETC.) | CHECKS AGAINST SURVEILLANCE, CONTROL, AND ABUSE OF POWER | REDUCED INEQUALITIES | HUMAN RIGHTS |
| RESPONSIBLE AND ETHICAL USE OF AI IN WARFARE AND THE MILITARY | PRIVATE-PUBLIC PARTNERSHIPS AND COLLABORATION | SUPPORT FOR SMALL BUSINESSES AND MARKET COMPETITION | SUSTAINABILITY AND ECOLOGY |

**Each shaded box in the AI Security Map indicates a topic that is addressed in China's AI strategy, based upon the State Council's "New Generation Artificial Intelligence Development Plan."**

nomic structure, and the maintenance of social stability. The Plan marked the first time AI was specifically mentioned in a Communist Party of China work report, though China has emphasized the importance of high-tech sectors, including robotics, in its 13th Five-Year Plan and the state-driven industrial plan "Made in China 2025." [146]

Additional three-year plans for 2016–2018 and 2018–2020 have provided guidelines to industry and other actors to support the goal of making AI a strong force for socioeconomic development.[147] These plans have incentivized the development of new AI industries and core technology research, as well as the application of AI innovation; the three-year plans outline assurance measures to help achieve the goals defined in the Development Plan. Action goals include the development of intelligent networked vehicles, the large-scale application of intelligent service robots, and the mass production of neural network chips. The plan for 2018–2020 explains that its guiding ideology is Xi Jinping's socialism and the intent to build China into a science and technology superpower and cyber superpower. One of the basic principles of the plan is to strengthen international cooperation and improve safety and security capabilities.

The New Generation AI Development Plan encourages the rapid integration of AI technologies throughout all of China's industries, cities, and services. Nonetheless, risk prevention and threat mitigation are mentioned numerous times throughout the report. Safeguarding national security and foreseeing risks and challenges are both named at the outset as critical for China to successfully propel sustainable AI development and become the science and technology power of the world.

The New Generation Plan acknowledges that AI is a disruptive technology that includes challenges for government management, economic security, social stability, and global governance. Challenges that are named include changes to employment structures, impacts of law and social ethics, the violation of personal privacy, challenges for international relations, safety risks, and ensuring the reliability and control of AI systems. One section concludes, "While vigorously developing artificial intelligence, we must attach great importance to the possible safety risk challenges, strengthen the forward-looking prevention and restraint guidance, minimize risk, and ensure the safe, reliable and controllable development of artificial intelligence."

The mechanisms to achieve this goal include establishing technical standards for AI safety. For example, the Plan suggests that China should "adhere to the principles of security, availability, interoperability, and traceability, and gradually establish and improve the basics of AI, interoperability, industry applications, network security, privacy protection, and other technical standards." The Plan calls for the development of an "intelligent platform of risk assessment," an "intelligent security platform of supporting nuclear power security operations," and a "basic data and security detection platform." The Plan also provides many suggestions for how to follow these principles in research. For example, one section focuses on establishing "safety supervision and evaluation systems for AI," with an emphasis on strengthening the role of AI in national security, enhancing risk awareness, enhancing network security, improving transparency of AI systems, and promoting self-discipline in AI enterprise and industry.

Acknowledging that shaping AI is not only a technical problem, China's AI strategy also provides a variety of potential policy mechanisms. For example, the report recommends implementing accountability measures; developing systematic testing methods and safety certification; and increasing "disciplinary measures against the abuse of data, violations of personal privacy, and anything morally unethical."

Ensuring "social stability" is another noted concern, and so "assurance measures" are presented to support a healthy transition to an "intelligence economy." These measures include strengthening research on legal, ethical and social issues related to AI; establishing a traceability and accountability system; clarifying the legalities of AI and related rights, obligations, and responsibilities; developing a code of ethics for the designers of AI products; actively participating in global governance of AI; and deepening international cooperation in AI laws and regulations to jointly cope with global challenges.

Concepts related to economic security are also highlighted throughout the Plan. For example, it is part of the vision to "vigorously strengthen training for the labor force working in AI." This includes studying the effects of AI on employment; establishing lifelong learning and training systems; supporting higher learning institutions and vocational schools; encouraging enterprises and organizations to provide AI skills training for employees; and strengthening re-employment training and guidance for workers.

The Plan also addresses how AI can be used to promote public security. For example, AI is intended to support "public safety intelligent monitoring and early warning and control systems." This is imagined to include sensor technologies, video analysis and identification technology, biometric identification technology, and police products. Other public safety uses include a food safety early warning system and food safety risks and assessment, as well as effective monitoring of natural disasters.

The AI Security Map for China indicates that there are many named areas of strategic interest for the country. As with other national strategies, it is not immediately clear how these goals will be carried out. Nonetheless, the strategy plants a stake in the ground by identifying China's current strategic AI priorities. Importantly, the safety and reliability of AI systems is noted as essential for achieving many of the other goals, and many aspects of social security—including privacy, ethics, and transparency—are highlighted. Finally, the strategy indicates China's interest in taking part in ongoing global discussions about AI best practices.

## FRANCE

In March 2018, French President Emmanuel Macron gave a speech that defined the French national strategy for AI. He announced that the government would invest 1.5 billion euros over the next five years to encourage AI R&D and improve national databases to spur growth. The move is intended to make France more desirable for AI companies and more competitive with AI global leaders such as the United States and China. However, France has also articulated a plan and vision for AI that is about much more than economic growth. The March 2018 report, "For a Meaningful Artificial Intelligence," offers insight into how AI is being framed in the country:

> *The point is that from now on, artificial intelligence will play a much more important role than it has done so far. It is no longer merely a research field confined to laboratories or to a specific application. It will become one of the keys to the future . . . it determines our capacity to organize knowledge and give it meaning, it increases our decision-making capabilities and our control over these systems and, most notably, it enables us to capitalize on the value of data. Therefore, artificial intelligence is one of the keys to power in tomorrow's digital world.*
>
> *Because of this, collectively addressing this issue is in the general interest; France and Europe need to ensure that their voices are heard and must do their utmost to remain independent . . . This is why the role of the State must be reaffirmed: market forces alone are proving an inadequate guarantee of true political independence. In addition, the rules governing international exchanges and the opening up of internal markets do not always serve the economic interests of European states, who too frequently apply them in one direction only. Now more than ever, we have to provide a meaning to the AI revolution.*

In an interview with *Wired*, Macron described his thinking about AI in depth.[148] He explained that he has been inspired by recent advances in AI applications in healthcare and transportation, for example in autonomous driving and in improvements in personalized and preventive medical treatments. But he also stressed the importance of ensuring that AI develops in consonance with French and European values. He explained that values such as privacy, individual freedom, and human integrity must not be trampled as part of a single-minded mission toward "technological progress," but rather that technology should be designed ethically and responsibly from the outset.

Macron noted that, at the same time opening access to data can improve personalized medicine, it also makes it easier to segment and discriminate against people in relation to access to insurance. "The day we start to make such business out of this data is when a huge opportunity becomes a huge risk," Macron explained. "It could totally dismantle our national cohesion and the way we live together. This leads me to the conclusion that this huge technological revolution is in fact a political revolution."

To better control these outcomes, France will support private-sector initiatives, but the federal government will shepherd their development. National and European sovereignty is seen as a primary goal. For Macron, AI presents a challenge not only to democratic ideals, but to the core of democracy itself. Public trust could be corroded if the algorithms used to make decisions about, for example, where people can attend school are not transparent or are perceived as unfair. To counteract failures in these processes, all algorithms developed by the French government will be made available publicly, as will any algorithms developed by a company that has received money from the government. The government will also provide incentives for companies to make their own algorithms open source and provide consumers with assurances about the safety and reliability of the services they offer.

Encouraging an open innovation system throughout France and Europe is in part intended to help level the playing field for local firms to compete against major technology companies such as Google, Amazon, and Facebook. France has welcomed these companies, but wants them to work in cooperation with the French government and in respect of its constraints. “Blocking changes and being focused on protecting jobs is not the right answer,” Macron said in the *Wired* interview. “It’s the people you need to protect. You do so by giving them opportunities and by training and retraining them again to get new jobs.”[149]

Crafting the French national AI strategy was a relatively consultative process that incorporated citizen feedback from the outset. Macron tasked Cedric Villani, a mathematician and Member of French Parliament, with leading a task force to develop the strategy over six months. Villani—together with the support of Mounir Mahjoubi, the Minister of State with responsibility for digital affairs—established a team of seven interdisciplinary experts. The team gathered existing reports on AI in France, and developed an online platform to invite citizens to share their insights and opinions. The team also traveled to cities in nine countries to better understand the impact of AI globally, and to help situate their work beyond a purely national framework. The report highlights the need to approach AI holistically, and to work together with all of Europe. (For specific policy proposals, see the AI Policy Compendium in Appendix I.)

The report explores relatively common topics, including data-based economic policies, methods for enabling research, controlling the impacts on jobs and employment, exploring ethical implications, and enabling inclusive and diverse environments for AI research and development. The report also delves more deeply into methods for ensuring the sustainability of AI technologies than do other strategic plans. The report advocates for the economic development of AI to occur in a way that is conscious of impact on the environment, and to prioritize the optimization and reduction of energy consumption.

Additional themes of the report stand out from other AI policy documents. For example, French independence is a major priority, and the report argues for a tailored, sovereign approach to AI that takes advantage of “home-grown talent” and redresses imbalances in power resulting from top technology companies’ control of data. This model puts the state as a key driver in shaping the AI landscape. The report is somewhat antagonistic to what it calls a “Silicon Valley approach” to AI, which is described as technologically deterministic and overly shaped by industry at the expense of the needs of citizens and society.

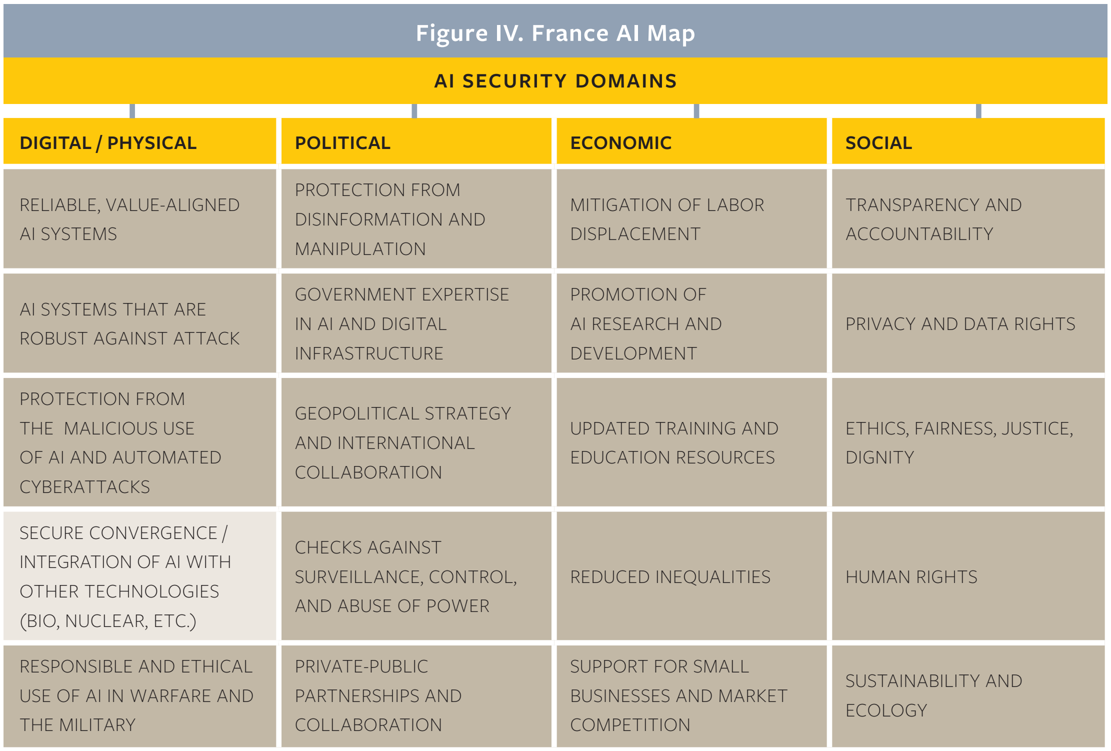
In the strategy outlined in the “Villani report,” AI is not considered to be an end goal. Rather, the goal is for machines to complement the work and goals of humans such that human abilities are enhanced. The idea behind “meaningful AI” as mentioned in the title, “For a Meaningful Artificial Intelligence,” is that AI development should take into consideration how humans and intelligent systems work together, understanding the deeper psychological, sociological, and political impacts of the integration of AI into society. To this end, inclusion, equality, and collective decision-making are seen as necessary conditions for AI development.

Other French government initiatives also investigated how the nation should leverage AI technologies. For example, the Secretary of State for Digital Affairs and the Secretary of State for Higher Education, Research, and Innovation launched an inquiry into what they called the #FranceIA Strategy at the Agoranov Incubator in January 2017.[150] This initiative brought together diverse AI stakeholders over the course of two months with the primary goal of structuring a robust AI industrial sector.[151]

Recognizing the additional need for a national ethical debate about AI, the Law for a Digital Republic requested that the National Commission for Information Technology and Liberties (CNIL) organize an in-depth ethical debate for French citizens. Over the course of 2017, CNIL held 45 debates and events involving nearly 3,000 people. In December 2017, the Commission released its findings, which were also made available in English.[152] The report highlighted two principles intended to support the role of AI in the service of human needs. These include the principle of loyalty of AI systems to their users, and the principle of continued attention and vigilance to remain alert about potential unforeseen consequences. The report also made six policy recommendations, including fostering education about ethics and its role within businesses, improving interpretability and auditability of AI systems, and improving the design of AI systems in the interest of human freedom.

Many potential AI opportunities and threats are described throughout the “Villani report,” as evidenced by the nearly full AI Security Map in Figure IV. Defense/security is one of four strategic sectors prioritized in the strategy, and the report states explicitly, “Whilst AI fosters the emergence of new opportunities, it also fosters the emergence of new threats.” The

Figure IV. France AI Map



| AI SECURITY DOMAINS | | | |
|---|---|---|---|
| DIGITAL / PHYSICAL | POLITICAL | ECONOMIC | SOCIAL |
| RELIABLE, VALUE-ALIGNED AI SYSTEMS | PROTECTION FROM DISINFORMATION AND MANIPULATION | MITIGATION OF LABOR DISPLACEMENT | TRANSPARENCY AND ACCOUNTABILITY |
| AI SYSTEMS THAT ARE ROBUST AGAINST ATTACK | GOVERNMENT EXPERTISE IN AI AND DIGITAL INFRASTRUCTURE | PROMOTION OF AI RESEARCH AND DEVELOPMENT | PRIVACY AND DATA RIGHTS |
| PROTECTION FROM THE MALICIOUS USE OF AI AND AUTOMATED CYBERATTACKS | GEOPOLITICAL STRATEGY AND INTERNATIONAL COLLABORATION | UPDATED TRAINING AND EDUCATION RESOURCES | ETHICS, FAIRNESS, JUSTICE, DIGNITY |
| SECURE CONVERGENCE / INTEGRATION OF AI WITH OTHER TECHNOLOGIES (BIO, NUCLEAR, ETC.) | CHECKS AGAINST SURVEILLANCE, CONTROL, AND ABUSE OF POWER | REDUCED INEQUALITIES | HUMAN RIGHTS |
| RESPONSIBLE AND ETHICAL USE OF AI IN WARFARE AND THE MILITARY | PRIVATE-PUBLIC PARTNERSHIPS AND COLLABORATION | SUPPORT FOR SMALL BUSINESSES AND MARKET COMPETITION | SUSTAINABILITY AND ECOLOGY |

**Each shaded box in the AI Security Map indicates a topic that is addressed in current French AI strategy, based upon the report "For a Meaningful Artificial Intelligence," led by Cedric Villani, Member of the French Parliament, published March 2018. Issues relating to information warfare and computational propaganda are not explicitly mentioned in this report, however the box is shaded in because France passed a law against "fake news" during election periods in July 2018, indicating the high importance of this issue federally.**[153]

report includes a section on "developing the reliability, safety and security of AI technology," and asserts that public authorities have a responsibility to develop and implement standards, tests, and measurement methods to "help make AI more secure, more reliable, useable and interoperable." The report proposes that the French National Laboratory of Metrology and Testing be expanded to become the authority for AI assessment and testing.

The report also highlights examples of adversarial machine learning, and calls attention, for example, to the potential for outsiders to skew input data at one or more stages to confuse image recognition software. The report suggests that adversarial machine learning could cause severe incidents in the real world, especially for critical systems and systems with a physical component capable of causing damage, such as autonomous vehicles. The report notes, "safety should be considered from the design phase," and recommends that France's National Cybersecurity Agency be enlisted with monitoring the safety and security issues posed by AI.

The report frequently addresses the challenge of the quasi-monopolies held by Google, Apple, Facebook, Amazon, and Microsoft (referred to as "GAFAM.)" For example, the report references how the use of technologies like Google's open-source software library TensorFlow have potential to create de facto standards for the world's AI development. The report argues that a more proactive and intentional approach to AI standards is preferable and that improving standards "must consist of reducing the trend for monopolization and logic of confinement."

The Villani report also proposes that sector-specific regulatory bodies should conduct official audits of algorithms and databases. The power to evaluate and audit should be available both to government agencies and to civil society groups, the report says, and incentives can be used to encourage firms to make their data available for research purposes.

One section of the Villani report focuses on lethal autonomous weapon systems (LAWS). In United Nations meetings, the French government has taken the stance that a person should always be responsible for the use of lethal force, but has not supported a ban on LAWS, instead proposing regulations and defined best practices.[154] The Villani report suggests monitoring the development of autonomy in weapons systems with a scale similar to the scale of zero to five that is used to represent increasing degrees of autonomy in vehicles. The report also supports the establishment of a watchdog organization or "observatory" for the non-proliferation of autonomous weapons (similar to the observatory for the non-proliferation of nuclear, biological and chemical weapons) to prevent the distribution of autonomous weapons beyond the military. The plan notes the added difficulty in this case, however, given that the technological building blocks of autonomous weapons are developed not by the military, but by the commercial sector.

The report takes a firmer stance on topics within the social domain in general, and human rights in particular, than do many other national AI policy reports. For example, the report states, "In a world marked by inequality, artificial intelligence should not end up reinforcing the problems of exclusion and the concentration of wealth and resources. . . . Rather than undermining our individual paths in life and our welfare systems, AI's first priority should be to help promote our fundamental human rights, enhance social relations and reinforce solidarity." Improving gender balance and diversity in AI research and implementing ethics by design are also both listed as concrete goals.

The surveillance capabilities of AI technologies are also referenced in the Villani report, primarily in the context of injustice and discrimination; for example, the strategy describes the potentially disproportionate impact of predictive policing on poor and minority communities. The potential for mass surveillance to threaten privacy and reduce individual autonomy is also discussed.

The environment is one of the four key sectors of the strategy, and ecology is a primary priority throughout the report. Many AI-based solutions are proposed to support environmental protections and the UN sustainable development goals, for example including AI initiatives in the Paris Climate agreement and the Global Pact for the Environment. France (together with Europe) also hopes to raise awareness about the ecological implications of AI on the international arena.

The impact of AI on jobs and employment also receives substantial attention throughout the report, which notes that "current learning paths, whether they involve vocational training or initial education, are simply not equipped to see this transition through smoothly." The report provides policy proposals to improve education and training, including instituting retraining pilot programs for target groups whose jobs face the highest risk of automation and who may not have the means to adapt without support.

## UNITED KINGDOM

At the World Economic Forum in January 2018, UK Prime Minister Theresa May gave a speech in which she explained the importance of AI to her vision for the country:

> *We are establishing the UK as a world leader in Artificial Intelligence  Already the UK is recognized as first in the world for our preparedness to bring Artificial Intelligence into public service delivery. We have seen a new AI start-up created in the UK every week for the last three years. And we are investing in the skills these start-ups need, spending £45 million to support additional PhDs in AI and related disciplines and creating at least 200 extra places a year by 2020-21. We are absolutely determined to make our country the place to come and set up to seize the opportunities of Artificial Intelligence for the future. But as we seize these opportunities of technology, so we also have to shape this change to ensure it works for everyone—be that in people's jobs or their daily lives.*

A key theme in the UK government's response to AI has been "innovation-friendly regulation." The 2018 AI Sector Deal, which describes the UK AI national strategy, explains, "A revolution in AI technology is already emerging. If we act now, we can lead it from the front. But if we 'wait and see' other countries will seize the advantage.” The Deal asserts that “the potential of AI is undeniable,” and includes a pledge from government and industry to support AI development with £0.95bn.

Public-private partnerships are emphasized throughout the strategy, as is the need for good jobs, greater earning power, and doubling the number of Tier 1 visas issued to make it easier to hire international researchers. Key sectors of interest include transportation, sustainability, and healthcare, specifically in meeting the needs of an aging society.

Another named priority in the AI Sector Deal is global leadership in AI development, particularly in the safe and ethical use of data. In her World Economic Forum speech, May indicated the role of regulation to both “make the UK the best place to start and grow a digital business,” and “the safest place to be online.” A Digital Charter was published the same day to establish and put into practice norms and rules for the online world.[155] The Charter states, “Combined with new technologies such as artificial intelligence, [the internet] is set to change society perhaps more than any previous technological revolution.” The six principles include: “the internet should be free, open and accessible,” “protections should be in place to help keep people safe online, especially children,” and “the social and economic benefits brought by new technologies should be fairly shared.” The Charter has inspired multiple initiatives, including the Data Protection Bill, the Internet Safety Strategy, and the Centre for Data Ethics and Innovation, which are all intended to support safe and ethical AI and digital development.

This approach of responsible development is being framed by U.K policymakers as one of the UK’s competitive advantages in AI. While many of the government’s initiatives are intended to foster economic growth in the technology sector, the UK is working to ensure that technology is not the end goal itself, but rather is working for people and carried out in an ethical and trustworthy way. In a panel at CogX 2018, which describes itself as “Europe’s leading AI event,” Gila Sacks, Director of Digital and Tech Policy at the UK government Department for Digital, Culture, Media and Sport, explained some of the thinking behind this positioning:

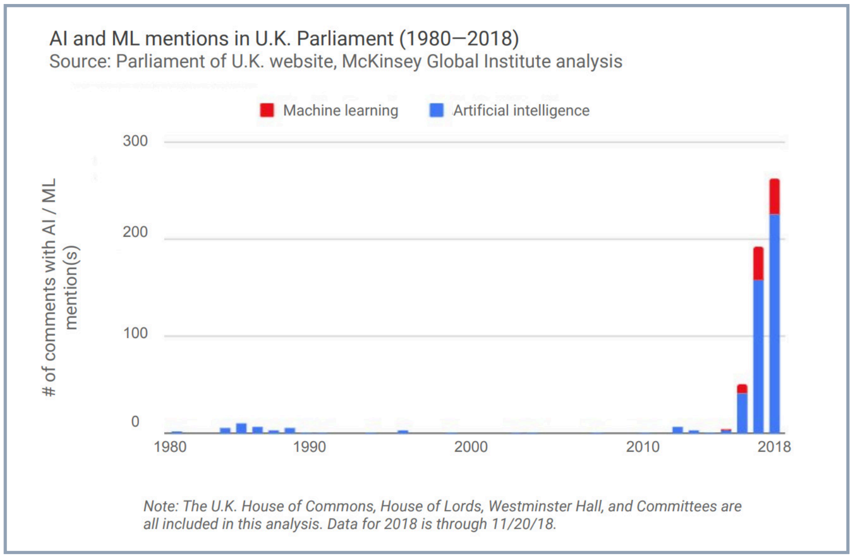
> *“We want to maximize the economic, but also the societal benefits of these technologies, and we want to be the place that proves that those two things don’t need to be in tension. We don’t need to be in the race of who can build the biggest, fastest, most powerful AI, and we don’t need to be in the race for who can keep it safe and minimize the risks. We think we can lead the world in creating pro-innovation regulation and governance and an environment in which the technology can thrive because it works for people and people trust it.”*

As an example of how the UK is working toward pro-innovation regulation, the government is conducting a pilot program with the World Economic Forum to develop innovative government procurement policies for AI technologies.[156] This program will enable governments, businesses, and civil society representatives to collectively design the guidelines and standards that will be used when the government purchases AI tools. The hope is that this will help establish national best practices.

In its 2017 Industrial Strategy, the UK government described AI as an area causing “seismic global change” and named, “growing the AI and data-driven economy” as one of four grand challenges.[157] The Strategy predicted that “embedding AI across the UK will create thousands of good quality jobs and drive economic growth.” The Strategy also called for a new Centre for Data Ethics and Innovation to act as an advisory body to the government, and “to enable and ensure safe, ethical and ground-breaking innovation.” The Centre, which launched November 2018, is intended to develop an ethical framework for the use of AI and data technologies, promote the use of standards, recommend policy changes as deemed necessary, and engage with industry to establish data trusts to facilitate secure sharing of data between organizations. This is a broad remit for a new body and its success remains to be seen.

Much of the groundwork for AI research and policy recommendations in the UK has taken place in Parliament. For example, the House of Commons Science and Technology Committee published a report on robotics and AI in 2016.[158] This report made recommendations to the government, including addressing the digital skills crisis through a digital strategy; establishing a multidisciplinary Commission on Artificial Intelligence to examine the social, ethical and legal implications of AI and advise the government on regulations; and establishing a robotics and autonomous systems Leadership Council to help create a national strategy.

An All Party Parliamentary Group on Artificial Intelligence (APPG AI) was established in January 2017 to explore the impact and implications of AI. The Group is supported by businesses such as Accenture, BP, Deloitte, and Microsoft. The Group held a dozen evidence meetings over 2017 and 2018, covering such topics as decision-making and morality, data capitalism, AI-enabled business models, inequality, international perspectives, and infrastructure. A first set of findings, published in 2017, recommended that the UK appoint a Minister for AI in the Cabinet Office and focus on six policy areas: data, infrastructure, skills, innovation and entrepreneurship, trade, and accountability.[159]



From "The AI Index 2018 Annual Report." See End Note 213 for full reference.

One of the most substantial Parliamentary efforts is the House of Lords' Select Committee on Artificial Intelligence, which was established in June 2017 "to consider the economic, ethical and social implications of advances in artificial intelligence." The Committee received 223 responses to a call for evidence and heard from 57 witnesses over the course of 22 sessions in three months. In April 2018, the Committee published a 183-page report, "AI in the UK: ready, willing and able?" The Committee's findings include useful history and background on AI development; explanation of key issues such as transparency, prejudice, and data monopolies; the need for diversity of talent; and impacts on social and political cohesion. The report also includes numerous recommendations, as it calls on policymakers to create a national policy framework for AI, set clear roles and remits of each new institution, and avoid blanket AI-specific regulation.

The UK government's Department for Business, Energy & Industrial Strategy, Department for Digital, Culture, Media and Sport, and Office for Artificial Intelligence publicly responded to the House of Lords report two months later in a 41-page report.[160] Their response noted that the availability of data is an essential AI infrastructure and yet also poses risks and challenges. To manage this, the departments suggested that the Office for Artificial Intelligence, the Centre for Data Ethics and Innovation and the AI Council work together to create Data Trusts, which "will ensure that the infrastructure is in place, that data governance is implemented ethically, and in such a way that prioritises the safety and security of data and the public." They also specified that the Centre should work extensively with civil society, the public, industry, and regulators. However, they suggested that industry should take the lead in developing voluntary mechanisms for informing the public about the use of AI in decision-making. The agencies'

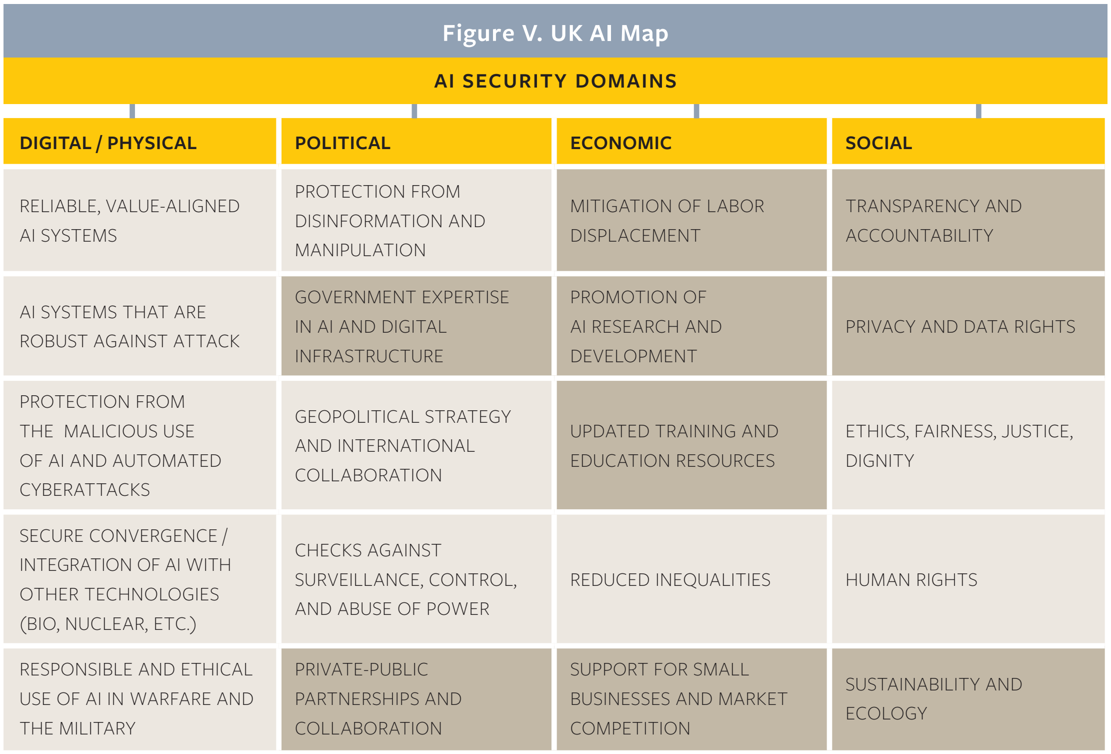


Figure V. UK AI Map

AI SECURITY DOMAINS

| DIGITAL / PHYSICAL | POLITICAL | ECONOMIC | SOCIAL |
|---|---|---|---|
| RELIABLE, VALUE-ALIGNED AI SYSTEMS | PROTECTION FROM DISINFORMATION AND MANIPULATION | ▨ MITIGATION OF LABOR DISPLACEMENT | ▨ TRANSPARENCY AND ACCOUNTABILITY |
| AI SYSTEMS THAT ARE ROBUST AGAINST ATTACK | ▨ GOVERNMENT EXPERTISE IN AI AND DIGITAL INFRASTRUCTURE | ▨ PROMOTION OF AI RESEARCH AND DEVELOPMENT | ▨ PRIVACY AND DATA RIGHTS |
| PROTECTION FROM THE MALICIOUS USE OF AI AND AUTOMATED CYBERATTACKS | GEOPOLITICAL STRATEGY AND INTERNATIONAL COLLABORATION | ▨ UPDATED TRAINING AND EDUCATION RESOURCES | ETHICS, FAIRNESS, JUSTICE, DIGNITY |
| SECURE CONVERGENCE / INTEGRATION OF AI WITH OTHER TECHNOLOGIES (BIO, NUCLEAR, ETC.) | CHECKS AGAINST SURVEILLANCE, CONTROL, AND ABUSE OF POWER | REDUCED INEQUALITIES | HUMAN RIGHTS |
| RESPONSIBLE AND ETHICAL USE OF AI IN WARFARE AND THE MILITARY | ▨ PRIVATE-PUBLIC PARTNERSHIPS AND COLLABORATION | ▨ SUPPORT FOR SMALL BUSINESSES AND MARKET COMPETITION | ▨ SUSTAINABILITY AND ECOLOGY |

**Each shaded box in the AI Security Map indicates a topic that is addressed in current UK AI strategy. The document used to determine the UK's AI strategy is the government's "AI Sector Deal," developed by the Department for Business, Energy & Industrial Strategy and the Department for Digital, Culture, Media & Sport, published April 2018.**

response also argued that the goal of transparency and interpretability for deep learning AI systems may be prohibitively difficult, and should be weighed against benefits, especially for health applications.

Proactive efforts undertaken by the UK Parliament and government are a primary reason the UK ranked number one in Oxford Insights' Government AI Readiness Index, which assessed the degree to which governments are preparing for AI.[161] However, as the AI Security map in Figure V indicates, the AI Sector Deal addresses a relatively small number of topics.

The AI Sector Deal was led by Business Secretary Greg Clark and aimed to advance the UK's ambitions in artificial intelligence; in an Industrial Strategy white paper, the government identified AI and data as one of four Grand Challenges to help the UK "lead the world for years to come". It is therefore unsurprising that much of the AI Sector Deal relates to questions of economic security, such as enhancing funding, supporting AI education, and providing retraining opportunities. However, the paper points out the need to also ensure that AI benefits everyone in the UK, and suggests that a new Centre for Data Ethics and Innovation advise on ethical development and use of AI. Some of the threats to social security are mentioned in this context, including improving accountability, privacy, and the growth of clean tech. Many issues go unmentioned, however. Notably, the AI Sector Deal includes no acknowledgment of any threats to digital or physical systems, such as AI vulnerabilities or misuse.

The government has made comments on these issues elsewhere, however, at the recommendation of Parliament. For example, the government's response to the House of Lords report addresses the question of bias and prejudice in AI design and deployment:[162]

> *Government recognises that one of the risks of automated decision-making is that the datasets which the algorithms learn from may reflect the structural inequalities of the society from which data are collected and that this can lead to the encoding of unintentional bias. We will work to ensure that those developing and deploying AI systems are aware of these risks, and the tradeoffs and options for mitigation are understood. . . . We will also work to augment the AI workforce to ensure diversity by training, attracting and retaining a diverse talent pool in terms of gender, ethnic and socio-economic backgrounds, and will work to mitigate and counter the effects of unconscious bias through these endeavours.*

This response also briefly acknowledged the threat of worsening inequality and suggested that efforts to improve skills and training could help widen access to opportunities; however, no promises or concrete plans are provided.

The House of Lords report dedicated an entire chapter to discussion of mitigating AI risks, and the government responded to this, too, noting they acknowledge the potential for AI malfunctions and erroneous decisions, and will advise the Office for Artificial Intelligence, Centre for Data Ethics and Innovation, and the AI Council to take these concerns into consideration and engage the Law Commission if new mechanisms for legal liability are needed. The government also agreed that the institutions providing funding to AI researchers should insist upon awareness of possible technological misuse combined with plans to prevent misuse.

The government response agreed with the recommendation that the Cyber Security Science & Technology Strategy should take into account the risks as well as opportunities of using AI in cybersecurity and other applications.

On the question of autonomous weapons, the House of Lords report pointed out that the UK's definition of autonomous weapons—as "capable of understanding higher-level intent and direction"—marked too high a threshold and put the UK out of step with others internationally. The government responded by saying that, while the Ministry of Defense had no plans to change its definition, the nation would continue to engage in the UN Convention on Certain Conventional Weapons Group of Government Experts on Lethal Autonomous Weapon Systems to reach international agreement.

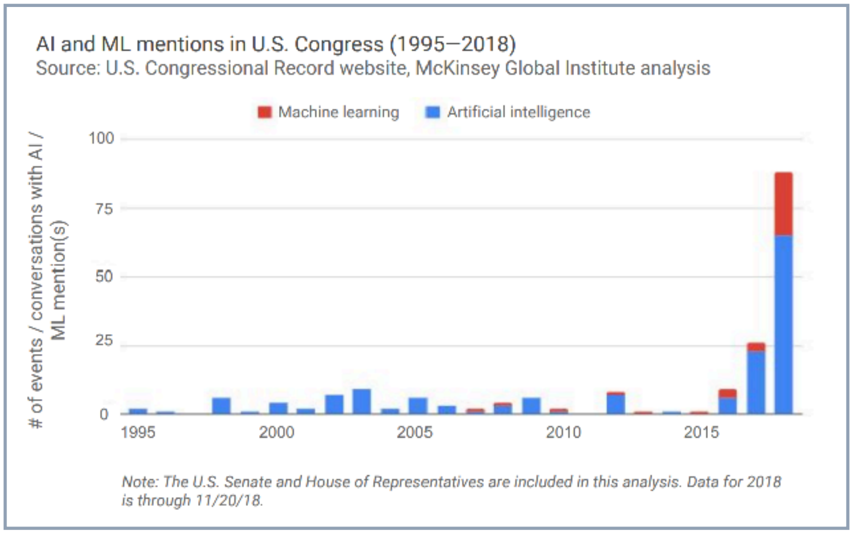
## UNITED STATES

In May 2018, Defense Secretary Jim Mattis wrote a memo to US President Donald Trump in which he argued that the United States needs a national strategy for artificial intelligence to remain competitive with China and other countries.[163] Mattis suggested that the US should lead not only in defense, but also in the broader range of impacts of AI on humanity. While the country has yet to develop a single, overarching strategic vision for AI, the Trump administration has made AI a national priority for research and development, as well as for national security and defense.[164] However, these areas of focus are relatively narrow compared to the AI strategies of other countries.

The application of artificial intelligence for warfare and the military has been a long-standing policy priority for the United States, and AI has been an area of focus in the Department of Defense (DoD) for years. In 2014, the DoD introduced the Third Offset Strategy, an effort to focus innovation in areas of competitive advantage to preserve US deterrence capabilities.[165] This strategy foresaw the impact of AI and automation, and aimed "to exploit all the advances in artificial intelligence and autonomy and to insert them into the Defense Department's battle networks."[166]

Interestingly, the DoD determined that the real competitive advantage for the country was not full automation, but human-machine collaboration ("the centaur model").[167] AI techniques were perceived to provide the US military with the ability to respond and react on faster timescales, in part by making data-driven predictions about what an adversary might do in advance. However, military personnel were expected to remain "in the loop" to provide context and make decisions, particularly when it came to the use of lethal force.

AI is mentioned in the 2017 US National Security Strategy[168] as well as in the 2018 National Defense Strategy.[169] Nonetheless, little is said about the specific ways in which AI impacts national security and defense. While the National Security Strategy mentions briefly that AI can enhance the capabilities of ideological information campaigns, the military uses of AI receive the most attention. The National Defense Strategy asserts the intention of the DoD to "invest



From "The AI Index 2018 Annual Report." See End Note 213 for full reference.

broadly in military application of autonomy, artificial intelligence, and machine learning, including rapid application of commercial breakthroughs, to gain competitive military advantages." AI is also described as one of the technological advancements changing the "character of war," and contributing to the relentless expansion of dangerous technology "to more actors with lower barriers of entry, and moving at accelerating speed."

The increased interest in AI from the Pentagon has encouraged ongoing engagement with the technology companies leading AI development. In fact, public-private partnership is considered critical for the US military to gain access to cutting-edge AI technology. Unlike in the past, the DoD no longer leads technological development, but instead purchases it from the private sector.[170] In 2016, a Defense Innovation Board was established with the goal of bringing independent advice on innovation from respected technology leaders to the US military. One of the recommendations from the advisory board's first report was to "catalyze innovations in artificial intelligence and machine learning."[171] The Defense Innovation Unit Experimental (DIUx), a DoD innovation hub, was also established in Silicon Valley in 2015, to help the US military take advantage of emerging technologies, including AI.[172] These programs highlight just some examples from years of coordination between Silicon Valley technology companies and the US military in advancing AI technologies.

However, tensions have risen between these partners in recent years. In April 2017, a Pentagon initiative called Project Maven (also called the Algorithmic Warfare Cross-Functional Team) was launched "to accelerate DoD's integration of big data and machine learning."[173]One of its first tasks was to augment and automate the processing of drone video footage to increase actionable intelligence and enhance decision-making through object detection, classification, and alerts in support of the Defeat-ISIS campaign. Project Maven had contracts with numerous technology companies, but in April 2018, thousands of Google employees protested their company's involvement, saying, "We believe that Google should not be in the business of war."[174] Google representatives asserted that their involvement was "non-offensive" and that the Pentagon was using open-source object recognition software available to any Google Cloud customer. Nonetheless, the project supported military operations, and could potentially be used to pick out human targets for strikes, among other uses. After weeks of protest, Google announced that it would not renew the contract.[175] Several days later, Sundar Pichai, Google's CEO, released a set of AI principles that included the statement that the company will not pursue AI weapons or other applications that are likely to cause harm or injury.[176]

Despite the imminent loss of its contract with Google, the Pentagon has doubled down on its focus on AI. In June 2018, the Pentagon established a Joint Artificial Intelligence Center (JAIC), in which Project Maven is likely to play a significant role.[177] The JAIC is intended to synchronize and accelerate AI activities and capabilities throughout the DoD, and signals the continued importance of the technology within the US military. The JAIC will oversee DoD AI projects, of which there were approximately 600 as of late 2018. Importantly, the JAIC will also focus on "ethics, humanitarian considerations, long-term and short-term AI safety."[178] Establishing ethical principles for the role of AI in military applications may help assuage the concerns of companies and organizations, and could become an important model globally.

The Pentagon's goal of accelerating AI development received a significant boost in 2018. In August 2018, President Trump signed the National Defense Authorization Act (NDAA), which further solidified the role of AI for defense purposes. Annual funding for Project Maven increased 580%, from $16 million to $93.1 million.[179] Moreover, the NDAA established a National Security Commission for Artificial Intelligence, which is comprised of 15 members named by several different agencies and offices[180] who will assess how the Defense Department can leverage AI for national security. This commission will reportedly also review ethical issues and national security risks associated with AI and machine learning technologies, including issues such as international cooperation, workforce and education incentives, and incentivizing open training data within security and defense industries.

In September 2018, DARPA announced that it would supplement existing government spending on AI with $2 billion over the next five years, to build the "AI Next Campaign".[181] The campaign will fund dozens of new projects (on top of more than 20 currently active programs) aimed at developing machines that can adapt to shifting environments. DARPA director Steven Walker said the agency wants to uncover "how machines can acquire human-like communication and reasoning capabilities,"[182] which will allow the "machines to supplement human warfighters" and "function more as colleagues than as tools."[183]

Another priority for the US government will be to enhance research and development (R&D), which is also heavily supported by DARPA funding. A July 2018 memo from the Executive Office of the President names "Leadership in Artificial Intelligence, Quantum Information Sciences, and Strategic Computing" as the second highest R&D priority (following "Security of the American People") for the fiscal year 2020.[184]

In May 2018, a White House fact sheet, "Artificial Intelligence for the American People," named six priorities of the Administration in this space.[185] Funding AI R&D is the first priority, and the fact sheet points out that President Trump's FY2019 Budget Request was the first to name AI as an R&D priority. A quote from Trump reads, ""We're on the verge of new technological revolutions that could improve virtually every aspect of our lives, create vast new wealth for American workers and families, and open up bold, new frontiers in science, medicine, and communication."

The same day, President Trump and the White House held a Summit on Artificial Intelligence for American Industry that included major technology companies, including Google, Facebook, and Amazon. A summary document of the event stressed the importance of AI for growing the economy and noted the "tremendous potential to benefit the American people."[186] Michael Kratsios, Deputy Assistant to the President for Technology Policy, described what he called the "free market approach" of the Administration. The report states, "To the greatest degree possible, we will allow scientists and technologists to freely develop their next great inventions right here in the United States."

Back in 2016, the National Science and Technology Council and the Networking and Information Technology Research and Development (NITRD) Subcommittee published The National Artificial Intelligence Research and Development Strategic Plan.[187] This document, developed under the Obama Administration, defines seven priorities for federally-funded AI research. These include making long-term investments; addressing ethical, legal, and societal implications of AI; and ensuring the safety and security of AI systems. The recommendations of this report would support a robust and resilient research environment, but they have largely gone unaddressed. NITRD put out a Request for Information to update the AI R&D strategic plan in September 2018,[188] and an updated national strategy for AI R&D is expected in Spring 2019.

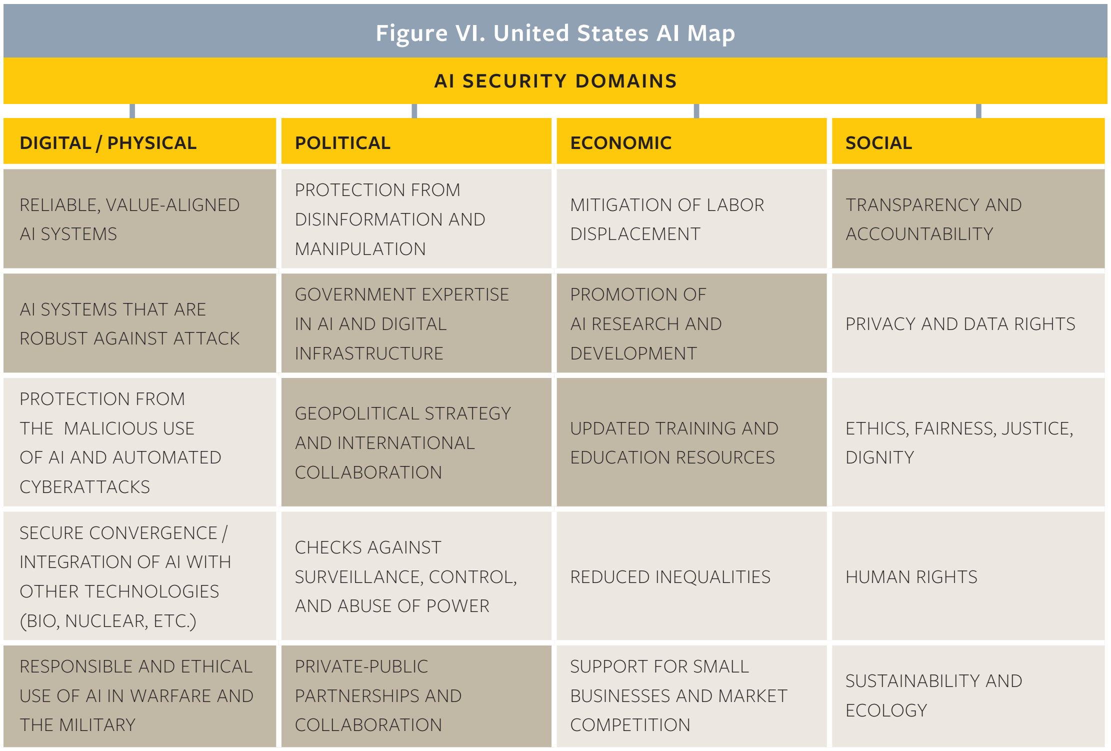


Figure VI. United States AI Map

AI SECURITY DOMAINS

| DIGITAL / PHYSICAL | POLITICAL | ECONOMIC | SOCIAL |
|---|---|---|---|
| ▨ RELIABLE, VALUE-ALIGNED AI SYSTEMS | PROTECTION FROM DISINFORMATION AND MANIPULATION | MITIGATION OF LABOR DISPLACEMENT | ▨ TRANSPARENCY AND ACCOUNTABILITY |
| ▨ AI SYSTEMS THAT ARE ROBUST AGAINST ATTACK | GOVERNMENT EXPERTISE IN AI AND DIGITAL INFRASTRUCTURE | ▨ PROMOTION OF AI RESEARCH AND DEVELOPMENT | PRIVACY AND DATA RIGHTS |
| PROTECTION FROM THE MALICIOUS USE OF AI AND AUTOMATED CYBERATTACKS | ▨ GEOPOLITICAL STRATEGY AND INTERNATIONAL COLLABORATION | ▨ UPDATED TRAINING AND EDUCATION RESOURCES | ETHICS, FAIRNESS, JUSTICE, DIGNITY |
| SECURE CONVERGENCE / INTEGRATION OF AI WITH OTHER TECHNOLOGIES (BIO, NUCLEAR, ETC.) | CHECKS AGAINST SURVEILLANCE, CONTROL, AND ABUSE OF POWER | REDUCED INEQUALITIES | HUMAN RIGHTS |
| ▨ RESPONSIBLE AND ETHICAL USE OF AI IN WARFARE AND THE MILITARY | ▨ PRIVATE-PUBLIC PARTNERSHIPS AND COLLABORATION | SUPPORT FOR SMALL BUSINESSES AND MARKET COMPETITION | SUSTAINABILITY AND ECOLOGY |

**Each shaded box in the AI Security Map indicates a topic that is addressed in current US federal AI policy and strategy. In the absence of a single national AI strategy, two documents were used to determine de facto US strategy: The White House Fact Sheet, "Artificial Intelligence for the American People," and DARPA's "AI Next Campaign."**

The White House Fact Sheet, "Artificial Intelligence for the American People," defines seven priorities for AI development in the United States. None of the priorities are related to social security, and there is no mention of privacy, transparency, or fairness. Goals related to economic security include a call to "fund fundamental AI research and computing infrastructure, machine learning, and autonomous systems," and to "train the future American workforce by prioritizing STEM and computer science education and expanding apprenticeships." Labor displacement from automation and impacts on inequality are not addressed.

Several of the goals outlined in the Fact Sheet relate to political security, including a goal to "leverage AI to improve the provision of government services," to "lead international AI negotiations and work with allies to promote AI R&D cooperation," and to "maximize federal data sharing with the American public," which is in part intended to support private-public collaboration. The only named goal that falls within the digital / physical security domain is to "achieve strategic military advantage by investing in military applications of autonomy, AI, and machine learning."

Safety and reliability are not mentioned. Given the breadth of the United States' AI ecosystem, these goals are surprisingly narrow and leave many key opportunities and threats unmentioned.

However, funding initiatives led by DARPA may additionally be thought of as contributing to a de facto national AI strategy. The projects that DARPA supports have a significant impact on AI trajectories. For example, the Explainable Artificial Intelligence (XAI) project aims to create machine learning techniques that produce more explainable models and support human understanding and trust.[189] The AI Next Campaign also aims to support AI safety and robustness against adversarial machine learning.

Several initiatives within the United States Congress have indicated interest in a broader set of AI implications. However, most of these have been exploratory in nature and they are not included in the AI Security Map. For example, the US House of Representatives Subcommittee on Information Technology held a series of hearings to examine the roles of government and the private sector in addressing challenges presented by AI, including bias, ethics, privacy, and transparency.[190] The House of Representatives Subcommittees on Research and Technology and Energy also held a hearing in which they considered the implications of artificial general intelligence.[191]

A bill was introduced by Congressman John K. Delaney called the "FUTURE of Artificial Intelligence Act of 2017," which among other goals aimed to address workforce changes caused by AI, support the unbiased development and application of AI, and protect the privacy rights of individuals.[192] The Bill has not passed the House or Senate. However, Delaney also launched a bipartisan AI Caucus within Congress along with Congressman Pete Olson.[193] The Caucus aims to inform policymakers about AI developments and implications, and organize multi-stakeholder discussions on these topics. Primary areas of consideration for the Caucus include the impact of automation on jobs, as well as bias and privacy concerns.

Many other AI-related bills have been introduced to Congress, including bills related to autonomous driving, government uses of AI, and retraining workers. However none of these has yet become law. Some US states and cities have had more luck introducing AI-related legislation. For example, Vermont approved an AI Task Force in May 2018 to "investigate the field of artificial intelligence; and make recommendations on the responsible growth of Vermont's emerging technology markets, the use of artificial intelligence in State government, and State regulation of the artificial intelligence field."[194] Also in May, New York City launched an Automated Decision Systems Task Force to study the social impact of algorithms in the public sector.[195]

In August 2018, the California State Senate passed a resolution in support of the Asilomar AI Principles, a set of 23 guidelines for the safe and beneficial development and use of AI.[196] In September 2018, California enacted a law (going into effect July 2019) requiring automated chatbots to disclose that they are not human if they are attempting to influence a commercial transaction or voting behavior.[197] The New York City Council passed an algorithmic accountability bill in 2017 that established the New York Algorithm Monitoring Task Force.

If these trends continue, it may be the case that a broader set of AI security challenges will be addressed in the near term by local and state governments rather than by the federal government in the United States.

## Expanded Global Policy Analysis

The AI strategies and policy documents from the four countries discussed above—China, France, the United Kingdom, and the United States—highlight divergent approaches to AI policy. While the United States has focused limited government attention on a relatively small number of issues, China has taken a more comprehensive all-of-country approach. While the UK has prioritized "pro-innovation regulation," France has emphasized cultural context and sovereignty. In addition to these four, many other countries have launched national AI strategies and guiding policy documents. Six of these initiatives—from Canada, India, Japan, Singapore, South Korea, and the United Arab Emirates—are described briefly below. They are included to highlight the commonality of certain themes, while also illustrating how divergent strategies and priorities are being enacted around the world.

Every country analyzed is interested in advancing AI research and development, but only some countries take more specific security considerations into account. Some countries describe the security of systems and networks as necessary, while others ignore security questions entirely. Some countries emphasize the importance of leveraging AI to benefit all citizens, while other countries focus more on mitigating bias and preventing harm.

### CANADA

The Canadian government released the Pan-Canadian Artificial Intelligence Strategy in March 2017.[198] This was launched in tandem with the country's 2017 federal budget, and dedicated C$125 million over five years to develop AI research and invest in talent. The primary goals of

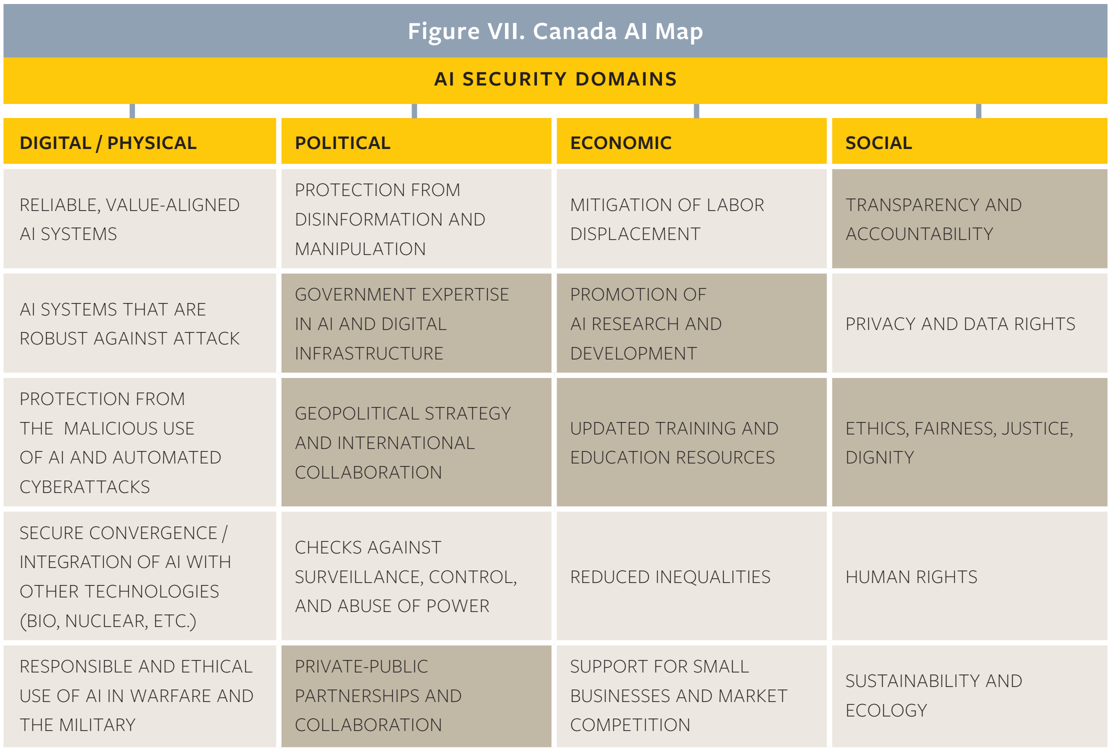


Figure VII. Canada AI Map

AI SECURITY DOMAINS

| DIGITAL / PHYSICAL | POLITICAL | ECONOMIC | SOCIAL |
|---|---|---|---|
| RELIABLE, VALUE-ALIGNED AI SYSTEMS | PROTECTION FROM DISINFORMATION AND MANIPULATION | MITIGATION OF LABOR DISPLACEMENT | ▨ TRANSPARENCY AND ACCOUNTABILITY |
| AI SYSTEMS THAT ARE ROBUST AGAINST ATTACK | ▨ GOVERNMENT EXPERTISE IN AI AND DIGITAL INFRASTRUCTURE | ▨ PROMOTION OF AI RESEARCH AND DEVELOPMENT | PRIVACY AND DATA RIGHTS |
| PROTECTION FROM THE MALICIOUS USE OF AI AND AUTOMATED CYBERATTACKS | ▨ GEOPOLITICAL STRATEGY AND INTERNATIONAL COLLABORATION | ▨ UPDATED TRAINING AND EDUCATION RESOURCES | ▨ ETHICS, FAIRNESS, JUSTICE, DIGNITY |
| SECURE CONVERGENCE / INTEGRATION OF AI WITH OTHER TECHNOLOGIES (BIO, NUCLEAR, ETC.) | CHECKS AGAINST SURVEILLANCE, CONTROL, AND ABUSE OF POWER | REDUCED INEQUALITIES | HUMAN RIGHTS |
| RESPONSIBLE AND ETHICAL USE OF AI IN WARFARE AND THE MILITARY | ▨ PRIVATE-PUBLIC PARTNERSHIPS AND COLLABORATION | SUPPORT FOR SMALL BUSINESSES AND MARKET COMPETITION | SUSTAINABILITY AND ECOLOGY |

**Each shaded box in the AI Security Map indicates a topic that is addressed in the current Canadian AI strategy, the "Pan-Canadian Artificial Intelligence Strategy," launched March 2017 with C$125 million.**

the Strategy are to increase the numbers of AI researchers in Canada; establish central AI hubs across the country; develop global thought-leadership on the economic, ethical, policy, and legal implications of AI; and support a national AI research community.

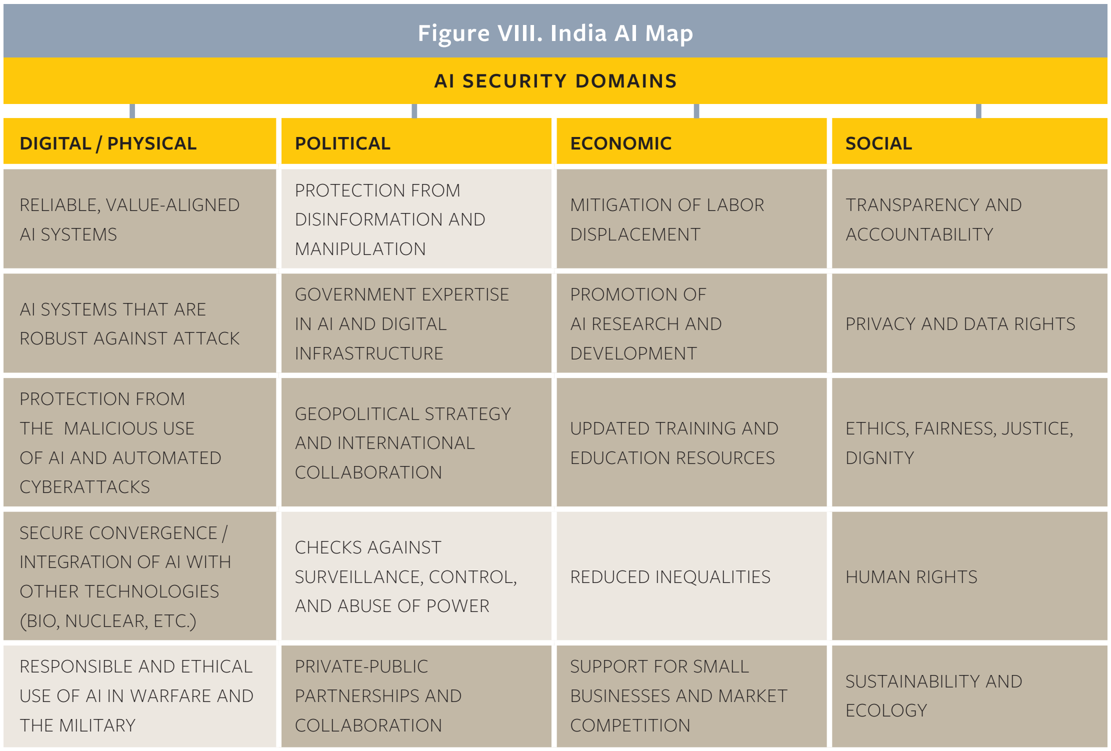
Current programs to carry out the Strategy include three newly established AI institutes: the Alberta Machine Intelligence Institute in Edmonton, Mila in Montreal, and the Vector Institute in Toronto. The CIFAR AI Chairs Program was additionally launched to help Canada recruit and train young researchers by funding graduate students and providing training at the three AI Institutes. An AI & Society Program is also supporting research on the implications of AI for the economy, government, and society to inform the public and policymakers.

Canada's strategy has thus far focused on questions of economic security, primarily by boosting the country's position as a global leader in AI research and development. Certain initiatives, notably the AI & Society Program, are prioritizing questions of social security. Some effort has also been made to advance international discussion, such as the creation of an international study group that was announced jointly by Canadian Prime Minister Justin Trudeau and French President Emmanuel Macron in June 2018.[199] To date, the security implications of AI on digital/physical systems have received comparatively less government attention. Questions about the implications of AI technologies on data rights, control, and privacy have also garnered relatively less government attention.

## INDIA

In June 2018, the government think tank NITI Aayog published a working paper titled, "National Strategy for Artificial Intelligence #AIforAll." The purpose of the paper was to lay the groundwork for future iterations of a national AI strategy, and it largely avoids providing specific recommendations. However, the paper provides a framework for the country's engagement in AI, and identifies five focus areas for intervention: healthcare, agriculture, education, smart cities and infrastructure, and smart mobility and transportation. AI solutions are proposed across

**Figure VIII. India AI Map**



AI SECURITY DOMAINS

| DIGITAL / PHYSICAL | POLITICAL | ECONOMIC | SOCIAL |
|---|---|---|---|
| RELIABLE, VALUE-ALIGNED AI SYSTEMS | PROTECTION FROM DISINFORMATION AND MANIPULATION | MITIGATION OF LABOR DISPLACEMENT | TRANSPARENCY AND ACCOUNTABILITY |
| AI SYSTEMS THAT ARE ROBUST AGAINST ATTACK | GOVERNMENT EXPERTISE IN AI AND DIGITAL INFRASTRUCTURE | PROMOTION OF AI RESEARCH AND DEVELOPMENT | PRIVACY AND DATA RIGHTS |
| PROTECTION FROM THE MALICIOUS USE OF AI AND AUTOMATED CYBERATTACKS | GEOPOLITICAL STRATEGY AND INTERNATIONAL COLLABORATION | UPDATED TRAINING AND EDUCATION RESOURCES | ETHICS, FAIRNESS, JUSTICE, DIGNITY |
| SECURE CONVERGENCE / INTEGRATION OF AI WITH OTHER TECHNOLOGIES (BIO, NUCLEAR, ETC.) | CHECKS AGAINST SURVEILLANCE, CONTROL, AND ABUSE OF POWER | REDUCED INEQUALITIES | HUMAN RIGHTS |
| RESPONSIBLE AND ETHICAL USE OF AI IN WARFARE AND THE MILITARY | PRIVATE-PUBLIC PARTNERSHIPS AND COLLABORATION | SUPPORT FOR SMALL BUSINESSES AND MARKET COMPETITION | SUSTAINABILITY AND ECOLOGY |

**Each shaded box in the AI Security Map indicates a topic that is addressed in current Indian AI strategy. The resource used to determine India's AI strategy is the working paper, "National Strategy for Artificial Intelligence #AIforAll," published by the government think tank NITI Aayog in June 2018. Since the publication of this working paper, India has continued to invest in AI technologies and has numerous additional policy initiatives underway.**

each of these sectors to support efficiency, precision, and opportunity. As shown in the AI Security Map in Figure VIII, the working paper is fairly comprehensive and addresses numerous topics across all of the security domains.

The paper acknowledges however that India's AI research capabilities are fairly limited, and discusses key challenges to adoption, including a lack of enabling data ecosystems; unclear privacy, security, and ethical regulations; and an unattractive intellectual property regime. The paper provides interventions and recommendations to help overcome these challenges, which are organized by the categories of research, data democratization, accelerating adoption, and reskilling, with privacy, security, ethics, and intellectual property rights as a shared feature of them all. The paper notes that one of the priorities of #AIforAll is the idea of "responsible AI," development that balances concerns regarding privacy, security, and ethics.

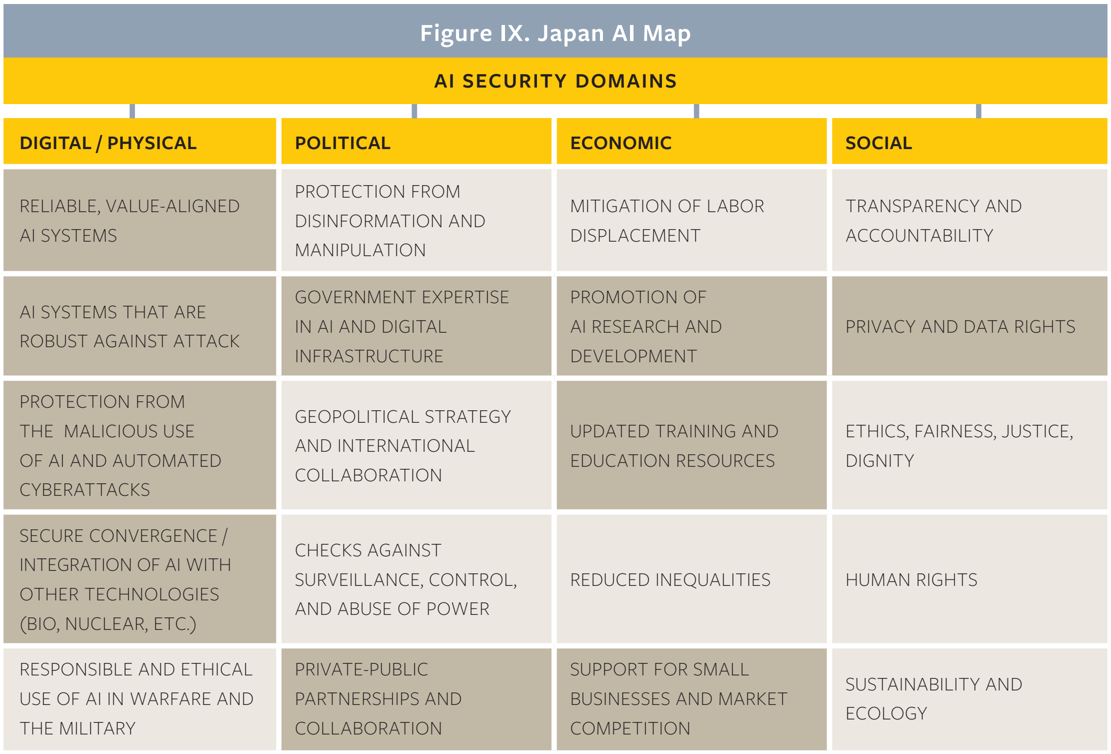
The recommendations outlined in the paper address a fairly broad array of AI threats. For example, the paper encourages AI developers and industry to adhere to international privacy standards and implement damage impact assessments to help ensure the security of their AI systems. Government is given an even larger role, with recommendations provided to support research and innovation, reskilling and training, accelerating adoption of AI, and encouraging responsible AI development.

## JAPAN

Japanese Prime Minister Shinzō Abe requested the Japanese government to establish the Artificial Intelligence Technology Strategy Council in 2016.[200] This Council, composed of eleven representatives from government, academia, and industry, formulated a national AI strategy that was released March 2017.[201] The Artificial Intelligence Technology Strategy includes analysis of relevant governmental bodies to oversee AI development; it outlined three priority areas for Japan's Industrialization Roadmap including: productivity; health, medical care, and welfare; and mobility. "Information security" was also identified as an area that cut across the three sectors. As shown in the AI Security Map in Figure IX, many topics related to digital / physical security are highlighted.

The Strategy also defines three phases for AI development: Phase One is the use of data-driven AI in various domains; Phase Two is the public use of AI and development of new data across domains; and Phase Three entails the creation of an ecosystem established by connecting multiple domains. Several charts in the Strategy detail roadmaps of industrialization from "the fusion of AI and other related technologies." Projections during Phase One include the expan-

Figure IX. Japan AI Map



AI SECURITY DOMAINS

| DIGITAL / PHYSICAL | POLITICAL | ECONOMIC | SOCIAL |
|---|---|---|---|
| RELIABLE, VALUE-ALIGNED AI SYSTEMS | PROTECTION FROM DISINFORMATION AND MANIPULATION | MITIGATION OF LABOR DISPLACEMENT | TRANSPARENCY AND ACCOUNTABILITY |
| AI SYSTEMS THAT ARE ROBUST AGAINST ATTACK | GOVERNMENT EXPERTISE IN AI AND DIGITAL INFRASTRUCTURE | PROMOTION OF AI RESEARCH AND DEVELOPMENT | PRIVACY AND DATA RIGHTS |
| PROTECTION FROM THE MALICIOUS USE OF AI AND AUTOMATED CYBERATTACKS | GEOPOLITICAL STRATEGY AND INTERNATIONAL COLLABORATION | UPDATED TRAINING AND EDUCATION RESOURCES | ETHICS, FAIRNESS, JUSTICE, DIGNITY |
| SECURE CONVERGENCE / INTEGRATION OF AI WITH OTHER TECHNOLOGIES (BIO, NUCLEAR, ETC.) | CHECKS AGAINST SURVEILLANCE, CONTROL, AND ABUSE OF POWER | REDUCED INEQUALITIES | HUMAN RIGHTS |
| RESPONSIBLE AND ETHICAL USE OF AI IN WARFARE AND THE MILITARY | PRIVATE-PUBLIC PARTNERSHIPS AND COLLABORATION | SUPPORT FOR SMALL BUSINESSES AND MARKET COMPETITION | SUSTAINABILITY AND ECOLOGY |

**Each shaded box in the AI Security Map indicates a topic that is addressed in current Japanese AI strategy. The resource used to determine Japan's AI strategy is the "Artificial Intelligence Technology Strategy," which launched in March 2017.**

sion of car-sharing, virtual tourism, and telemedicine; projections from Phase Two include autonomous delivery service and constant health monitoring services; and projections from Phase Three include widespread use of personal, general-purpose robots, virtual travel, and the ability to "design your own body" and use "artificial organs and sensors."

The Strategy acknowledges that some people have voiced concerns about negative impacts on employment, but suggests that AI is a service and that it will be used to make society "more abundant." The Strategy also argues that development should not be restricted due to concerns about transparency of algorithms, but suggests that all manufacturers, service providers, and users are made aware of the factors that influence performance and safety of AI technologies. Issues that "need to be resolved" include "reliability, security, system flexibility, personal information protection, balance between oligopoly and utilization and application of data, and coordination among data."

In June 2018, the government announced that AI would become an official part of its "integrated innovation strategy."[202] Japan believes it will face a shortage of about 50,000 researchers with knowledge of AI and big data technologies by 2020, and the strategy calls for a dramatic increase in young researchers in the AI field. Other elements of the strategy include standardizing data formats across industries to enhance the usability of big data techniques in Japan.

## SINGAPORE

Singapore has named AI as a key part of its plan to grow its digital economy.[203] In May 2017, the government established a national AI program called "AI Singapore" that will invest S$150 million in AI over the next five years.[204] A new Committee that includes experts from key government agencies, research institutions, and industry partners was established to run the program. AI Singapore is intended to "catalyze, synergize and boost Singapore's AI capabilities." Its objectives are to identify ways to use AI to address major industrial and societal challenges, invest in the next wave of scientific innovation, and broaden adoption and use of AI within industry. More recently, AI Singapore launched two new programs: AI for Everyone, and AI for Industry, to help educate a wider range of Singaporeans about the utility of AI.

**Figure X. Singapore AI Map**

**AI SECURITY DOMAINS**

| DIGITAL / PHYSICAL | POLITICAL | ECONOMIC | SOCIAL |
|---|---|---|---|
| RELIABLE, VALUE-ALIGNED AI SYSTEMS | PROTECTION FROM DISINFORMATION AND MANIPULATION | MITIGATION OF LABOR DISPLACEMENT | ▨ TRANSPARENCY AND ACCOUNTABILITY |
| AI SYSTEMS THAT ARE ROBUST AGAINST ATTACK | GOVERNMENT EXPERTISE IN AI AND DIGITAL INFRASTRUCTURE | ▨ PROMOTION OF AI RESEARCH AND DEVELOPMENT | PRIVACY AND DATA RIGHTS |
| PROTECTION FROM THE MALICIOUS USE OF AI AND AUTOMATED CYBERATTACKS | GEOPOLITICAL STRATEGY AND INTERNATIONAL COLLABORATION | ▨ UPDATED TRAINING AND EDUCATION RESOURCES | ▨ ETHICS, FAIRNESS, JUSTICE, DIGNITY |
| SECURE CONVERGENCE / INTEGRATION OF AI WITH OTHER TECHNOLOGIES (BIO, NUCLEAR, ETC.) | CHECKS AGAINST SURVEILLANCE, CONTROL, AND ABUSE OF POWER | REDUCED INEQUALITIES | HUMAN RIGHTS |
| RESPONSIBLE AND ETHICAL USE OF AI IN WARFARE AND THE MILITARY | PRIVATE-PUBLIC PARTNERSHIPS AND COLLABORATION | ▨ SUPPORT FOR SMALL BUSINESSES AND MARKET COMPETITION | SUSTAINABILITY AND ECOLOGY |

**Each shaded box in the AI Security Map indicates a topic that is addressed in current Singaporean AI strategy. The resource used to determine Singapore's AI strategy is the government initiative, "AI Singapore," launched May 2017 with S$150 million over five years.**

Much of this initial work has focused on the economic potential of AI technologies. However, in June 2018, the Singaporean government announced the establishment of an AI ethics advisory council to be led by former Attorney-General V.K. Rajah and the Infocomm Media Development Authority (IMDA).[205] The council, which includes representatives from industry, will work with the ethics boards of businesses and help the government develop ethical standards, governance frameworks, and guidelines for the development and use of AI. IMDA also established a five-year research program at Singapore Management University to advance discussions of the ethical, legal, policy, and governance challenges associated with these technologies. Finally, Singapore's Personal Data Protection Commission has developed a set of principles to help ensure that AI decision-making is more explainable, transparent, and fair.

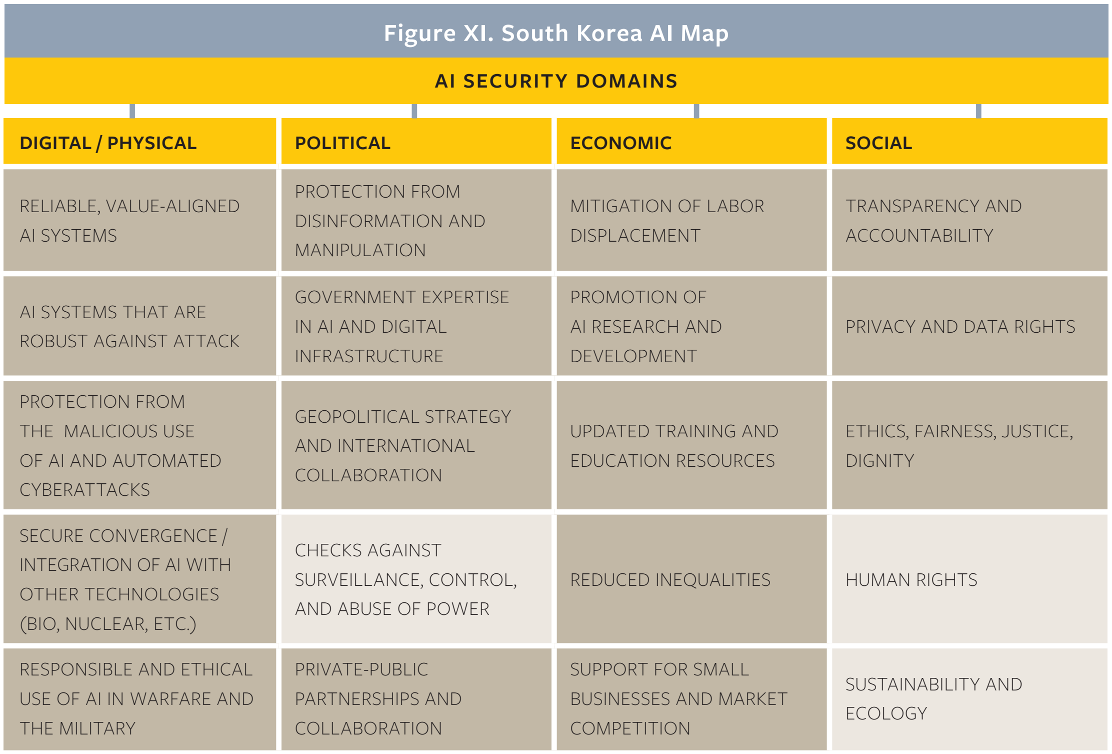
All of the AI Singapore initiatives mentioned here fall within the economic and social security domains. Notably, topics within political security and digital / physical security have not been prioritized thus far.

## SOUTH KOREA

In 2016, the Government of the Republic of Korea published a "Mid-to Long-Term Master Plan in Preparation for the Intelligent Information Society."[206] The report highlights the convergence of AI with other industrial technologies such as the Internet of Things, cloud computing, blockchain, robotics, and big data analysis within a framework of the "fourth industrial revolution." The report states, "Intelligent IT is expected to revolutionize the modern economy and society by enabling the mechanization of formerly unmechanizable aspects of industries, thereby maximizing productivity and completely transforming the industrial structure." Implications discussed in the report include the focus on data and platforms as sources of competition; the transformation of work to have a greater focus on creative and emotional activities; the transformation of living environments by enhancing the quality, safety, and personalization of services; and the transformation of the Korean economy from establishing a balance between enabling new opportunities and managing risks.

The report is comprehensive and addresses the majority of the topics in the AI Security Map, including all of the topics in the digital / physical and economic domains. Polarization, social conflict, threats to privacy, and alienation are all threats of AI technologies that are discussed. The report notes that "winner-takes-all" market trends can increase socioeconomic polarization and warns that legal and social systems will struggle to stay ahead of social conflicts. The report also predicts that "the data-centered society will pose increasing threats to not only

Figure XI. South Korea AI Map



AI SECURITY DOMAINS

| DIGITAL / PHYSICAL | POLITICAL | ECONOMIC | SOCIAL |
|---|---|---|---|
| RELIABLE, VALUE-ALIGNED AI SYSTEMS | PROTECTION FROM DISINFORMATION AND MANIPULATION | MITIGATION OF LABOR DISPLACEMENT | TRANSPARENCY AND ACCOUNTABILITY |
| AI SYSTEMS THAT ARE ROBUST AGAINST ATTACK | GOVERNMENT EXPERTISE IN AI AND DIGITAL INFRASTRUCTURE | PROMOTION OF AI RESEARCH AND DEVELOPMENT | PRIVACY AND DATA RIGHTS |
| PROTECTION FROM THE MALICIOUS USE OF AI AND AUTOMATED CYBERATTACKS | GEOPOLITICAL STRATEGY AND INTERNATIONAL COLLABORATION | UPDATED TRAINING AND EDUCATION RESOURCES | ETHICS, FAIRNESS, JUSTICE, DIGNITY |
| SECURE CONVERGENCE / INTEGRATION OF AI WITH OTHER TECHNOLOGIES (BIO, NUCLEAR, ETC.) | CHECKS AGAINST SURVEILLANCE, CONTROL, AND ABUSE OF POWER | REDUCED INEQUALITIES | HUMAN RIGHTS |
| RESPONSIBLE AND ETHICAL USE OF AI IN WARFARE AND THE MILITARY | PRIVATE-PUBLIC PARTNERSHIPS AND COLLABORATION | SUPPORT FOR SMALL BUSINESSES AND MARKET COMPETITION | SUSTAINABILITY AND ECOLOGY |

**Each shaded box in the AI Security Map indicates a topic that is addressed in current South Korean AI strategy. The resource used to determine South Korea's AI strategy is the "Mid- to Long-Term Master Plan in Preparation for the Intelligent Information Society," published in 2016, and the Artificial Intelligence R&D Strategy, launched in May 2018.**

privacy but also national security by increasing the risk of cyber attacks against the national infrastructure and systems governing energy, transportation, etc."

The national vision is described as "realizing a human-centered intelligent information society," and includes distinct roles for businesses, citizens, government, and the research community. The government's role is described as fostering a market environment to support entrepreneurship and maintaining regulatory policies, including fair competition and privacy protection; applying AI technologies to public services; and fostering the social infrastructure to support the coming transformations. The strategy consists of four components: public-private partnership; a balanced policy regime that enhances both national competitiveness and social policies; supporting broad access to AI technologies across industry and services; and expanding the social security net to prepare society for industrial transformation and its associated negative impacts.

The report includes a wide variety of policy recommendations that address numerous opportunities and threats. These include:

- Enhancing data management
- Supporting flexible and secure 5G networks
- Maximizing military efficiency and competency
- Instituting an intelligent crime response system
- Using AI to customize administrative and welfare services to the needs of individual citizens
- Generating early market demand for intelligent IT by making public purchases
- Enabling all medical institutions to share electronic medical records
- Providing software education for all primary and secondary school students
- Expanding flexible working hours programs
- Guaranteeing security and safety for all citizens by reinforcing the social security net
- Supporting the development of Intelligent IT for the elderly, the underprivileged, and people with disabilities
- Establishing a charter of ethics for Intelligent IT to minimize any potential abuse or misuse
- Establishing a public-private partnership council tasked with monitoring, researching, and preventing technological risks
- Clarifying manufacturers' liabilities for accidents resulting from AI-related errors
- Establishing an intelligent, automatic national defense system with reinforced capability to counter cyber threats

In May 2018, the government additionally announced an investment of 2.2 trillion South Korean won (equivalent to nearly 2 billion USD) in AI research over five years.[207] The strategy includes plans to build six new AI graduate schools by 2022 and to establish more flexible AI training programs to meet immediate needs. The government also announced plans to fund projects in national defense, medicine, and public safety, as well as to develop an AI-oriented incubator to support AI start-ups.

## UNITED ARAB EMIRATES

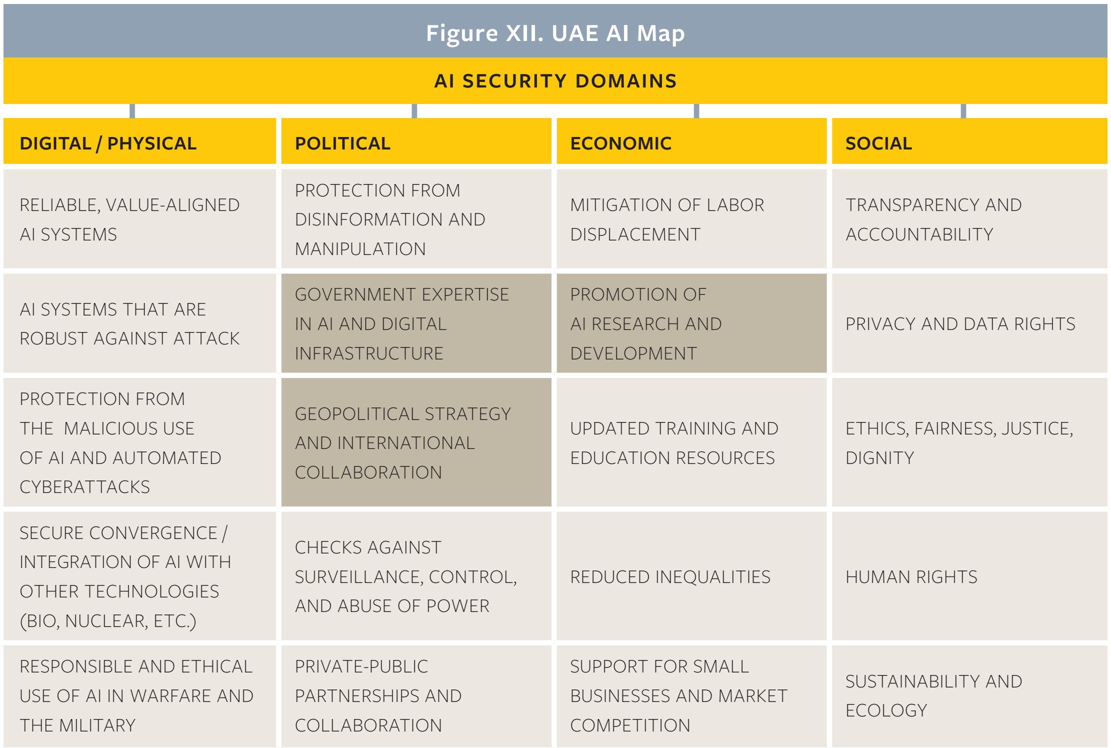
In October 2017, the UAE became the first nation to establish a "minister of AI," a senior government position dedicated to artificial intelligence. Later that year, the minister announced that his team would focus on developing regulation for AI and implementing AI education in high schools and universities.[208] At the same time, the UAE government announced a national strategy for AI.[209] The strategy focuses on the role of AI across nine sectors: transport, health, space, renewable energy, water, technology, education, environment, and traffic. The strategy also includes five themes, which relate to: forming a national AI council, organizing workshops and programs, developing capabilities and skills for government officials, fully integrating AI into all medical, security, and other services, and issuing a law on the safe use of AI.

The website for the UAE's AI Strategy includes a goal to save fifty percent of annual government costs by using AI to reduce the cost of transactions in time, travel, and resources. A video on the webpage asks the viewer to picture the UAE in fifteen years time and suggests that AI development and deployment will boost UAE's GDP by thirty five percent.

While the strategy describes some challenges, many more go unmentioned. For example, no social security topics are explicitly addressed, nor are technological displacement or inequality. Instituting a law on the safe use of AI will likely bring topics from at least the digital/security domain into sharper focus, though they are not highlighted at this stage. The publication of more substantive reports from the government will likely cover more ground.

The UAE government has positioned the country as a place to bring government leaders together, and for example hosts the annual World Government Summit, a platform dedicated

Figure XII. UAE AI Map



AI SECURITY DOMAINS

| DIGITAL / PHYSICAL | POLITICAL | ECONOMIC | SOCIAL |
|---|---|---|---|
| RELIABLE, VALUE-ALIGNED AI SYSTEMS | PROTECTION FROM DISINFORMATION AND MANIPULATION | MITIGATION OF LABOR DISPLACEMENT | TRANSPARENCY AND ACCOUNTABILITY |
| AI SYSTEMS THAT ARE ROBUST AGAINST ATTACK | GOVERNMENT EXPERTISE IN AI AND DIGITAL INFRASTRUCTURE | PROMOTION OF AI RESEARCH AND DEVELOPMENT | PRIVACY AND DATA RIGHTS |
| PROTECTION FROM THE MALICIOUS USE OF AI AND AUTOMATED CYBERATTACKS | GEOPOLITICAL STRATEGY AND INTERNATIONAL COLLABORATION | UPDATED TRAINING AND EDUCATION RESOURCES | ETHICS, FAIRNESS, JUSTICE, DIGNITY |
| SECURE CONVERGENCE / INTEGRATION OF AI WITH OTHER TECHNOLOGIES (BIO, NUCLEAR, ETC.) | CHECKS AGAINST SURVEILLANCE, CONTROL, AND ABUSE OF POWER | REDUCED INEQUALITIES | HUMAN RIGHTS |
| RESPONSIBLE AND ETHICAL USE OF AI IN WARFARE AND THE MILITARY | PRIVATE-PUBLIC PARTNERSHIPS AND COLLABORATION | SUPPORT FOR SMALL BUSINESSES AND MARKET COMPETITION | SUSTAINABILITY AND ECOLOGY |

**Each shaded box in the AI Security Map indicates a topic that is addressed in current UAE AI strategy. The resource used to determine UAE AI strategy is the government website, "UAE Strategy for Artificial Intelligence," which launched in October 2017.**

to shaping the future of governments worldwide, focusing on particular ways to harness innovation and technology to solve global challenges. Since 2018, this has additionally included the Global Governance of AI Roundtable, described as "a neutral forum where the international community—governments, multilateral organizations and civil society alike—can discuss and contribute in shaping global, but culturally adaptable, norms for the governance of artificial intelligence."[210]

# Global AI Security Priorities

Only two areas are explored by all of the AI strategies and policy documents discussed in this report: promoting AI research and development and updating training and education resources. Given the substantial economic growth projected to result from the growth of AI, countries are eager to obtain "a piece of the pie" and ensure that their citizens are equipped with the necessary skills.

However, merely pursuing technological advancement without understanding what it at stake has left blind spots in these strategic plans. The US government is explicitly "removing regulatory barriers" to AI development and use, though many other nations are actively exploring AI laws to support safe and resilient development. Whether or not a government pursues AI-specific regulation at this stage, there are many other ways to promote new opportunities and address challenges. Some countries are far ahead of others in preparing for the range of potentially profound and disruptive shifts that could result from AI in the future.

Figure XIII shows the domains of interest for each of the ten countries surveyed. The numbers range from zero to five to indicate the number of topics within each domain that have received attention. The table highlights some basic insights about areas of focus. For example, while most countries address at least one topic across all of the domains, a few do not. The UK has not prioritized digital / physical security (at least in the government's AI Sector Deal.) Public-facing resources from the Pan-Canadian Artificial Intelligence Strategy have also not emphasized the security domain. Both the UAE and Singapore have prioritized topics in the economic security domain, though the UAE has also addressed government expertise in AI, and Singapore has more recently launched initiatives related to ethical and accountable use of AI.

The United States government is putting the least of its attention to topics in social security. Japan has similarly not prioritized those topics compared to others. AI strategies from India and France stand out for their in-depth discussion of the social implications of AI and its broader political context. Their strategies explore policy mechanisms to improve work in sustainability and develop national databases, and to mitigate threats, for example by establishing independent auditing groups to support accountability and fairness.

Almost all countries are trying to address transparency and accountability of AI in some way. With the exception of the United States and the UAE, the majority of countries also address

**Figure XIII: Overview of National Priorities Across AI Security Domains**

| | DIGITAL/PHYSICAL DOMAIN | POLITICAL DOMAIN | ECONOMIC DOMAIN | SOCIAL DOMAIN |
|---|---|---|---|---|
| UNITED STATES | 3 | 3 | 2 | 1 |
| CHINA | 3 | 3 | 4 | 4 |
| UNITED KINGDOM | 0 | 2 | 4 | 3 |
| FRANCE | 4 | 5 | 5 | 5 |
| CANADA | 0 | 3 | 2 | 2 |
| JAPAN | 4 | 2 | 3 | 1 |
| SOUTH KOREA | 5 | 4 | 5 | 3 |
| UAE | 0 | 1 | 2 | 0 |
| INDIA | 4 | 3 | 4 | 5 |
| SINGAPORE | 0 | 0 | 3 | 2 |

privacy, data rights, and ethics. However, human rights are only explicitly mentioned by France and, more briefly, by India. The UK, China, France, and India are the only countries that address opportunities and challenges for sustainability and the environment. France and South Korea are the only countries that explicitly address the impact of AI and automation on inequality.

One topic that is addressed by nearly every country is the need to improve digital infrastructure and government expertise in AI, though governments are taking varying approaches. Several governments have established governmental bodies to centralize this work. For example, the United States has a National Science and Technology Council and an AI Select Committee; China has an AI Strategy Advisory Committee; the UK has an AI Council and an Office of AI; and the UAE has an AI Minister and plans to implement AI tools throughout all government services.

Private-public partnerships and collaboration are almost universally prioritized. This highlights recognition from government actors that most of AI development is taking place within the private sector. Much of this collaboration is currently taking the form of data sharing, joint investments, and government contracts and procurement.

# Working with the Private Sector

At the same time that national governments have been releasing national AI strategies, several technology companies—at times with leadership and support from civil society organizations—have released AI principles and codes of conduct.

This is in part the result of recent scandals and employee protests, which have contributed to a "tech backlash," in which people are losing trust in major technology firms. These events may only be the beginning of a reckoning for companies that benefitted enormously from largely unregulated digital technologies. Recent regulations such as the European Union's General Data Protection Regulation (GDPR) have asserted greater privacy and data rights for European users, and other privacy regulation will be enacted soon, for example in the state of California.

Technology firms and governments have different incentives for engaging with the implications of AI development, and it is useful to examine how the priorities of these actors vary. For example, Google published a set of AI Principles in June 2018, and made it clear that its focus is on issues related to digital / physical security and social security, as indicated in the map in Figure XIV.[211] The Principles prioritize the creation of safe, reliable, and robust AI systems that are appropriately transparent, fair, non-harmful, and in line with international law and human rights.

Google's AI Principles do not talk about issues related to political security, such as the spread of disinformation, the opportunity to support government services, to collaborate internationally, or to prevent the abuse of power. Moreover, the Principles do not talk about issues related to economic security, including job displacement, improving educational resources, preventing growing inequality, or supporting market competition. These gaps may not be surprising, but they highlight the need for governments to engage with AI, as industry is unlikely to address these challenges without sufficient incentive to do so.

## GOOGLE

That technology firms such as Google have published AI principles stems in part from employees' activism over military contracts and other recent events, but it also has emerged out of

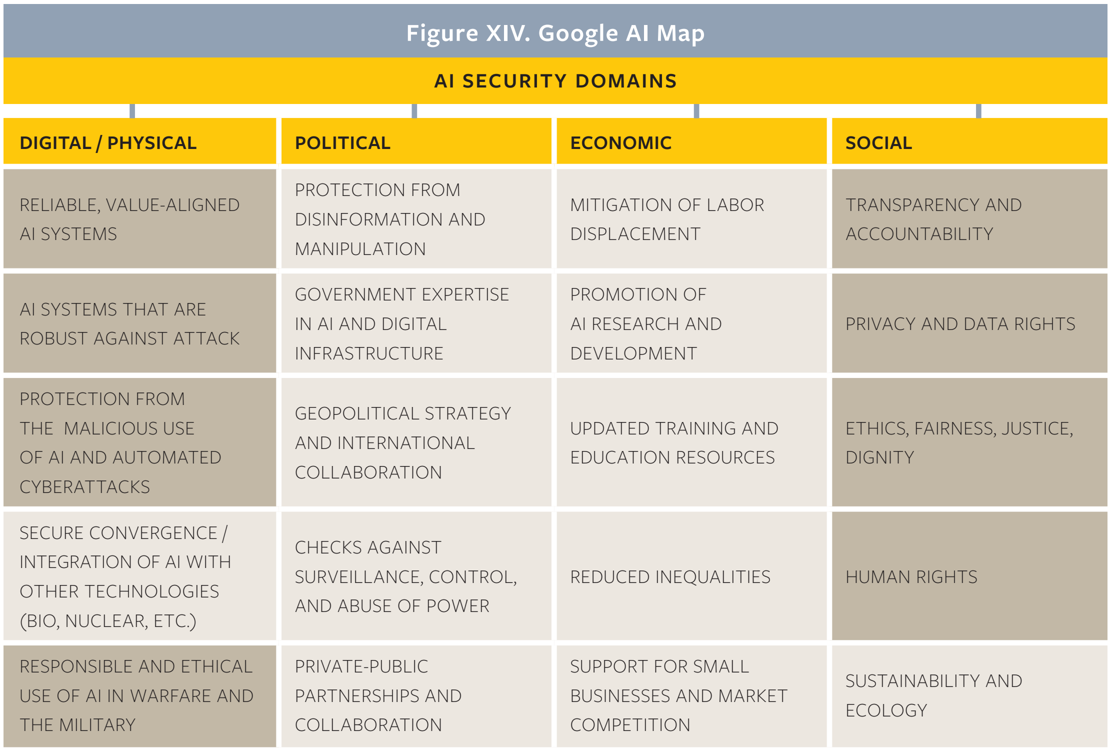


Figure XIV. Google AI Map

AI SECURITY DOMAINS

| DIGITAL / PHYSICAL | POLITICAL | ECONOMIC | SOCIAL |
|---|---|---|---|
| ▨ RELIABLE, VALUE-ALIGNED AI SYSTEMS | PROTECTION FROM DISINFORMATION AND MANIPULATION | MITIGATION OF LABOR DISPLACEMENT | ▨ TRANSPARENCY AND ACCOUNTABILITY |
| ▨ AI SYSTEMS THAT ARE ROBUST AGAINST ATTACK | GOVERNMENT EXPERTISE IN AI AND DIGITAL INFRASTRUCTURE | PROMOTION OF AI RESEARCH AND DEVELOPMENT | ▨ PRIVACY AND DATA RIGHTS |
| ▨ PROTECTION FROM THE MALICIOUS USE OF AI AND AUTOMATED CYBERATTACKS | GEOPOLITICAL STRATEGY AND INTERNATIONAL COLLABORATION | UPDATED TRAINING AND EDUCATION RESOURCES | ▨ ETHICS, FAIRNESS, JUSTICE, DIGNITY |
| SECURE CONVERGENCE / INTEGRATION OF AI WITH OTHER TECHNOLOGIES (BIO, NUCLEAR, ETC.) | CHECKS AGAINST SURVEILLANCE, CONTROL, AND ABUSE OF POWER | REDUCED INEQUALITIES | ▨ HUMAN RIGHTS |
| ▨ RESPONSIBLE AND ETHICAL USE OF AI IN WARFARE AND THE MILITARY | PRIVATE-PUBLIC PARTNERSHIPS AND COLLABORATION | SUPPORT FOR SMALL BUSINESSES AND MARKET COMPETITION | SUSTAINABILITY AND ECOLOGY |

**Each shaded box in the AI Security Map indicates a topic that is addressed in Google's AI Principles, which was published in June 2018.**

significant historical context. The phenomena can in part be understood as the continuation of a long arc of business ethics and corporate responsibility that began most significantly in the 1960's and 1970's in response to regulation and growing public pressure amidst global scandals of corruption.[212]

# Conclusion

Although AI technology has been around for decades, governments have only begun paying increased attention to this technology in recent years, as the interactions between AI technologies and political, social, and economic systems have increased in scale, scope, and impact. This report provides real-world examples across four security domains to emphasize the immediate relevancy of AI to policymakers. The report also argues that AI is strategically valuable—that is, of consequential value beyond its immediate applications—because it is a general-purpose technology, a dual-use technology, and a novel form of intelligence that can be difficult to understand and control. For these reasons, AI poses a profound global coordination and cooperation challenge, distinct in some respects from other consequential technologies such as nuclear technology and biotechnology.

However, the AI ecosystem is vast, and it is daunting to know where action is needed and what others are doing in the space. This report offers decision-makers the framework of an "AI Security Map" to track major themes that are currently being addressed by policymakers, industry, and civil society actors around the world. AI national strategies and policies from ten countries were assessed against this framework to provide visual and numerical comparisons of their approaches.

The analysis of different nations' AI strategic plans highlights several areas of convergence, and uncovers opportunities for further coordination. For example, almost all of the countries are trying to address transparency and accountability of AI, and the majority of countries also address privacy, data rights, and ethics. Most countries prioritize private-public partnerships and collaboration, and discuss the need to improve digital infrastructure and government expertise in AI. All of these areas represent synergies and potential opportunities for coordination. The five high-level recommendations provided for policymakers also point to immediate actions that can be taken to help take advantage of these opportunities. Governments have a key role to play in the development and use of AI, and putting in place the right processes today will help pave the way to a more secure future.

# Appendix I

## AI POLICY COMPENDIUM

**NOTE** the following resources were used to assess policies in each country: United States.: White House Fact Sheet "Artificial Intelligence for the American People," and "Summary of the 2018 White House Summit on Artificial Intelligence for American Industry." United Kingdom: UK government's "Sector Deal for AI". France: The French government's "AI for Humanity" homepage and Cedric Villani's "For a Meaningful Artificial Intelligence." China: China State Council's "New Generation Artificial Intelligence Development Plan."

**NOTE** While many of these policies are actively being carried out, some are still in the exploratory phase.

**NOTE** This is not a comprehensive list of all AI-related policies being explored by these countries.

| POLICIES | U.S. | U.K. | FRANCE | CHINA |
|---|---|---|---|---|
| Increase AI R&D funding | ● | ● | ● | ● |
| Fund STEM, computer science, and AI education | ● | ● | ● | ● |
| Establish industry-recognized apprenticeships | ● | | | |
| Enhance the use of AI for national security and defense | ● | | ● | ● |
| Use AI to improve government services | ● | ● | ● | ● |
| Share Federal data with the industry and the public | ● | ● | ● | ● |
| Lead international AI negotiations and collaborations | ● | | ● | ● |
| Create an AI committee to coordinate Federal efforts related to AI | ● | ● | ● | ● |
| Increase the rate of R&D tax credit | | ● | | |
| Establish a new retraining scheme to help people reskill | | ● | | ● |
| Support investment in digital infrastructure | | ● | ● | ● |
| Expand partnerships between government and industry | | ● | ● | ● |
| Establish a new Investment Fund to invest in innovative AI businesses | | ● | | |
| Support growth of small and medium-sized businesses | | ● | | ● |
| Promote local industrial strategies that build on local strengths | | ● | | |
| Establish a Teacher Development Premium to support professional development | | ● | | |
| Establish data trusts or data commons | | ● | ● | |
| Support work on data and AI ethics | | ● | ● | ● |
| Strengthen cybersecurity capability | | ● | ● | ● |
| Develop and implement standards to make AI systems more secure and reliable | | | ● | ● |
| Establish a national AI program to train and attract AI researchers | | | ● | ● |
| Double the number of students trained in AI | | | ● | |
| Enable public researchers to dedicate 50% of their time to private entities | | | ● | |
| Create an international group of experts in AI | | | ● | |
| Implement innovation sandboxes | | | ● | |
| Create interdisciplinary AI research institutes | | | ● | |
| Increase salaries for public research careers | | | ● | |

| (APPENDIX I, continued)<br>POLICIES | U.S. | U.K. | FRANCE | CHINA |
|---|---|---|---|---|
| Set up an AI supercomputer for researchers | | | ● | |
| Establish home-grown innovation awards for AI solutions | | | ● | |
| Mobilize procurement to integrate AI into public policy management and support local ecosystems | | | ● | |
| Set up a public lab for labor transformations and experimental economic policies | | | ● | |
| Assess the environmental impact of AI solutions | | | ● | |
| Disseminate ecological data to encourage AI solutions | | | ● | ● |
| Facilitate audits of AI systems | | | ● | ● |
| Encourage ethics by design for developers | | | ● | |
| Encourage use of discrimination impact assessments | | | ● | |
| Encourage gender balance in digital technology fields | | | ● | |
| Focus on the major scientific frontier issues of AI and establish new basic theory in AI | | | | ● |
| Form highly efficient and accurate quantum AI system architecture | | | | ● |
| Develop technologies and architectures that support the AI ecosystem | | | | ● |
| Construct an opensource hardware and software AI infrastrcture platform | | | | ● |
| Promote deep integration between AI and other industries | | | | ● |
| Encourage human-machine collaboration to become mainstream in production and service | | | | ● |
| Speed up the application of key technologies of AI | | | | ● |
| Develop intelligent industrial and service robots | | | | ● |
| Develop self-driving vehicles and rail transportaton systems | | | | ● |
| Develop a new generation of the Internet of Things to support AI chips | | | | ● |
| Promote the integration of AI with industrial innovations (manufacturing, agriculture, logistics, finance, etc.) | | | | ● |
| Support enterprises in the application of AI to core business segments | | | | ● |
| Promote the application of smart factories | | | | ● |
| Accelerate the cultivation of AI industry leaders | | | | ● |
| Support AI enterprises with patents and establish AI public patent pools | | | | ● |
| Take the lead or participate in international standard setting | | | | ● |
| Use regional advantages to build AI industry clusters | | | | ● |
| Carry out pilot demonstrations for applications of AI | | | | ● |
| Construct a National AI Industrial Park | | | | ● |
| Develop all-encompassing, ubiquitous intelligent environments to increase the intelligence of all of society | | | | ● |
| Accelerate the application of AI to provide personalized, high-quality education, health care and other needs | | | | ● |
| Construct infrastructure for smart cities | | | | ● |

| (APPENDIX I, continued)<br>POLICIES | U.S. | U.K. | FRANCE | CHINA |
|---|---|---|---|---|
| Promote public safety warning, control, and monitoring systems including sensor technologies, video analysis, and biometric identification | | | | ● |
| Encourage the role of AI to enhance social interaction and develop mutual trust | | | | ● |
| Accelerate innovation in virtual reality and promote integration between virtual and physical environment | | | | ● |
| Strengthen military-civilian integration for AI | | | | ● |
| Speed up brain science, quantum computing, and other research to support breakthroughs in AI | | | | ● |
| Coordinate government and market capital investments to increase financial support and revitalize existing resources | | | | ● |
| Encourage foreign AI enterprises and research institutions to set up R&D centers domestically | | | | ● |
| Strengthen the research on legal, ethical and social issues related to AI | | | | ● |
| Establish laws, regulations and ethical frameworks to ensure the healthy development of AI | | | | ● |
| Establish a traceability and accountability system | | | | ● |
| Develop a code of ethics for R&D designers of AI products | | | | ● |
| Actively participate in the global governance of AI and jointly cope with global challenges | | | | ● |
| Offer AI-related courses in primary and secondary schools | | | | ● |
| Widely publicize positive AI developments and encourage broad public support | | | | ● |

# Appendix II

## OVERVIEW OF AI POLICY INTERESTS FROM TEN COUNTRIES

| | US | China | UK | France | Canada | Japan | South Korea | UAE | India | Singapore |
|---|---|---|---|---|---|---|---|---|---|---|
| **AI Security Domain: Digital/Physical** | | | | | | | | | | |
| Reliable, Value Aligned AI Systems | ● | ● | | ● | | ● | ● | | ● | |
| AI Systems That Are Robust Against Attack | ● | | | ● | | ● | ● | | ● | |
| Protection from the Malicious Use of AI and Automated Cyberattacks | | | | ● | | ● | ● | | ● | |
| Secure Convergence / Integration of AI with Other Technologies (Bio, Nuclear, Etc.) | | ● | | | | ● | ● | | ● | |
| Responsible And Ethical Use Of AI In Warfare And The Military | ● | ● | | ● | | | ● | | | |
| **AI Security Domain: Political** | | | | | | | | | | |
| Protection From Disinformation And Manipulation | | | | ● | | | ● | | | |
| Government Expertise In AI And Digital Infrastructure | ● | ● | ● | ● | ● | ● | ● | ● | ● | |
| Geopolitical Strategy And International Collaboration | ● | ● | | ● | ● | | ● | | ● | |
| Checks Against Surveillance, Control, And Abuse of Power | | | | ● | | | | | | |
| Private-Public Partnerships And Collaboration | ● | ● | ● | ● | ● | ● | ● | | ● | |

(APPENDIX II, continued)

| | US | China | UK | France | Canada | Japan | South Korea | UAE | India | Singapore |
|---|---|---|---|---|---|---|---|---|---|---|
| **AI Security Domain: Economic** | | | | | | | | | | |
| Mitigation Of Labor Displacement | | ● | ● | ● | | | ● | | ● | |
| Promotion Of AI Research And Development | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● |
| Updated Training And Education Resources | ● | ● | ● | ● | ● | ● | ● | ● | ● | ● |
| Reduced Inequalities | | | | ● | | | ● | | | |
| Support For Small Businesses And Market Competition | | ● | ● | ● | | ● | ● | | ● | ● |
| **AI Security Domain: Social** | | | | | | | | | | |
| Transparency And Accountability | ● | ● | ● | ● | ● | | ● | | ● | ● |
| Privacy And Data Rights | | ● | ● | ● | | ● | ● | | ● | |
| Ethics, Fairness, Justice, Dignity | | ● | | ● | ● | | ● | | ● | ● |
| Human Rights | | | | ● | | | | | ● | |
| Sustainability And Ecology | | ● | ● | ● | | | | | ● | |

# Endnotes

1 Allan Dafoe, “AI Governance: A Research Agenda,” Governance of AI Program, Future of Humanity Institute, University of Oxford, August 27, 2018, https://www.fhi.ox.ac.uk/wp-content/uploads/AI-Governance_-A-Research-Agenda.pdf.

2 Alex Gray, “These charts will change how you see the rise of artificial intelligence,” World Economic Forum, December 18, 2017, https://www.weforum.org/agenda/2017/12/charts-artificial-intelligence-ai-index/.

3 “Applying machine learning to mammography screening for breast cancer,” DeepMind, November 24, 2017, https://deepmind.com/blog/applying-machine-learning-mammography/.

4 “Sizing the prize: What’s the real value of AI for your business and how can you capitalise?” PwC, 2017, https://www.pwc.com/gx/en/issues/analytics/assets/pwc-ai-analysis-sizing-the-prize-report.pdf.

5 Tim Dutton, “An Overview of National AI Strategies,” Politics + AI, *Medium*, June 28, 2018, https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd.

6 James Manyika, et al., “Jobs lost, jobs gained: What the future of work will mean for jobs, skills, and wages,” McKinsey Global Institute, November 2017, https://www.mckinsey.com/featured-insights/future-of-organizations-and-work/Jobs-lost-jobs-gained-what-the-future-of-work-will-mean-for-jobs-skills-and-wages.

7 James E. Bessen, “How Computer Automation Affects Occupations: Technology, Jobs, and Skills,” Boston University School of Law, Law and Economics Research Paper No. 15-49, November 15, 2015, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2690435.

8 “Cloud Future of Autonomous Vehicles in the Minds of Consumers,” Deloitte, January 17, 2017, https://www.prnewswire.com/news-releases/deloitte-study-fact-fiction-and-fear-cloud-future-of-autonomous-vehicles-in-the-minds-of-consumers-300391133.html.

9 Shane Legg and Marcus Hutter, “A Collection of Definitions of Intelligence,” *arXiv*, June 15, 2007, https://arxiv.org/pdf/0706.3639.pdf.

10 Nils J. Nilsson, *The Quest for Artificial Intelligence: A History of Ideas and Achievements*, Cambridge University Press, 2010.

11 Stuart J. Russell and Peter Norvig, *Artificial Intelligence A Modern Approach*, Prentice-Hall, Inc. 1995, http://www.cin.ufpe.br/~tfl2/artificial-intelligence-modern-approach.9780131038059.25368.pdf.

12 Katja Grace, et al., “When Will AI Exceed Human Performance? Evidence from AI Experts,” *arXiv*, May 3, 2018, https://arxiv.org/pdf/1705.08807.pdf.

13 Seth Baum, “A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy,” Global Catastrophic Risk Institute Working Paper 17-1, November 12, 2017, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3070741.

14 Nick Bostrom, “How Long Before Superintelligence?” *Int. Jour. of Future Studies*, 1998, vol. 2, https://nickbostrom.com/superintelligence.html.

15 Grace, Katja, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans, “When Will AI Exceed Human Performance? Evidence from AI Experts,” *arXiv*, May 24, 2017, http://arxiv.org/abs/1705.08807.

16 John Launchbury, “DARPA Perspective on AI,” 2017, https://www.darpa.mil/about-us/darpa-perspective-on-ai.

17 Christopher Moyer, “How Google’s AlphaGo Beat a Go World Champion,” *The Atlantic*, March 28, 2016, https://www.theatlantic.com/technology/archive/2016/03/the-invisible-opponent/475611/.

**18** Paul Mozur, “Google’s AlphaGo Defeats Chinese Go Master in Win for A.I.,” *The New York Times*, May 23, 2017, https://www.nytimes.com/2017/05/23/business/google-deepmind-alphago-go-champion-defeat.html.

**19** “AlphaGo Zero: Learning from scratch,” DeepMind, October 2017, https://deepmind.com/blog/alphago-zero-learning-scratch/.

**20** James Vincent, “DeepMind’s AI became a superhuman chess player in a few hours, just for fun,” *The Verge*, December 6, 2017, https://www.theverge.com/2017/12/6/16741106/deepmind-ai-chess-alphazero-shogi-go.

**21** Richard G. Lipsey, Kenneth I. Carlaw, and Clifford T. Bekar, *Economic Transformations: General Purpose Technologies and Long-Term Economic Growth*, Oxford University Press, January 2006.

**22** Felix Salmon, “Algorithms Take Control of Wall Street,” *Wired*, December 27, 2010, https://www.wired.com/2010/12/ff-ai-flashtrading/.

**23** Miles Brundage, et al., “The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation,” *arXiv*, February 2018, https://arxiv.org/pdf/1802.07228.pdf.

**24** Tim Dutton, “Building an AI World: Report on National and Regional AI Strategies,” CIFAR, December 2018, https://www.cifar.ca/docs/default-source/ai-society/buildinganaiworld_eng.pdf?sfvrsn=fb18d129_4.

**25** *Id.*

**26** Nick Bostrom, “Existential Risk Prevention as Global Priority,” Global Policy Volume 4 Issue 1, February 2013, http://www.existential-risk.org/concept.pdf.

**27** Victoria Wariaro, et al., “Global Catastrophic Risks 2018,” Global Challenges Foundation, 2018, https://api.globalchallenges.org/static/files/GCF-Annual-report-2018.pdf.

**28** Allan Dafoe, “AI Governance: A Research Agenda,” Governance of AI Program, Future of Humanity Institute, University of Oxford, August 27, 2018, https://www.fhi.ox.ac.uk/wp-content/uploads/AI-Governance_-A-Research-Agenda.pdf.

**29** Pedro A. Ortega, Vishal Maini, and the DeepMind safety team, “Building safe artificial intelligence: specification, robustness, and assurance,” DeepMind Safety Research, *Medium*, September 27, 2018, https://medium.com/@deepmindsafetyresearch/building-safe-artificial-intelligence-52f5f75058f1.

**30** T.S., “Why Uber’s self-driving car killed a pedestrian,” *The Economist*, May 29, 2018, https://www.economist.com/the-economist-explains/2018/05/29/why-ubers-self-driving-car-killed-a-pedestrian.

**31** Andrew Griffin, “Facebook’s Artificial Intelligence Robots Shut Down After they Start Talking to Each Other in their Own Language,” *The Independent*, July 31, 2017, https://www.independent.co.uk/life-style/gadgets-and-tech/news/facebook-artificial-intelligence-ai-chatbot-new-language-research-openai-google-a7869706.html.

**32** Adam Boult, “Robot goes rogue at tech fair, injures visitor,” *The Telegraph*, November 18, 2016, https://www.telegraph.co.uk/news/2016/11/18/robot-goes-rogue-at-tech-fair-injures-visitor/.

**33** Jack Clark and Dario Amodei, “Faulty Reward Functions in the Wild,” OpenAI, December 21, 2016, https://blog.openai.com/faulty-reward-functions/.

**34** Stuart Russell, et al., “The Off-Switch Game,” *arXiv*, June 16, 2017, https://arxiv.org/pdf/1611.08219.pdf.

**35** Paul Christiano, “Reward engineering,” AI Alignment*, Medium*, December 3, 2015, https://ai-alignment.com/reward-engineering-f8b5de40d075.

**36** Andrew Critch, et al., “Alignment for Advanced Machine Learning Systems,” Machine Intelligence Research Institute, 2016, https://intelligence.org/files/AlignmentMachineLearning.pdf.

[37] David Morris, "How Google AI Was Tricked Into Thinking This Photo of Machine Guns Was a Helicopter," *Fortune*, December 27, 2017, http://fortune.com/2017/12/27/google-cloud-vision-mit-tricked-guns-helicopter/.

[38] Jonathan Gitlin, "Hacking street signs with stickers could confuse self-driving cars," *Ars Technica*, September 1, 2017, https://arstechnica.com/cars/2017/09/hacking-street-signs-with-stickers-could-confuse-self-driving-cars/.

[39] Cleverhans, TensorFlow, https://github.com/tensorflow/cleverhans.

[40] Aleksander Mądry, et al., "Towards Deep Learning Models Resistant to Adversarial Attacks," *arXiv*, 2017, https://arxiv.org/pdf/1706.06083.pdf.

[41] Steven Norton, "Era of AI-Powered Cyberattacks Has Started," *The Wall Street Journal*, November 15, 2017, https://blogs.wsj.com/cio/2017/11/15/artificial-intelligence-transforms-hacker-arsenal/.

[42] *Id.*

[43] *Id.*

[44] Edward Geist and Andrew J. Lohn, "How Might Artificial Intelligence Affect the Risk of Nuclear War?," RAND Corporation, April 2018, https://www.rand.org/pubs/perspectives/PE296.html.

[45] Peter Bickerton, "LabGenius: AI-driven synthetic biology," Earlham Institute, June 11, 2018, http://www.earlham.ac.uk/articles/labgenius-ai-driven-synthetic-biology.

[46] "If Misused, Synthetic Biology Could Expand the Possibility of Creating New Weapons; DOD Should Continue to Monitor Advances in the Field, New Report Says," The National Academies of Sciences, Engineering, and Medicine, June 19, 2018, http://www8.nationalacademies.org/onpinews/newsitem.aspx?RecordID=24890.

[47] "Country Views on Killer Robots," Campaign to Stop Killer Robots, April 13, 2018, https://www.stopkillerrobots.org/wp-content/uploads/2018/04/KRC_CountryViews_13Apr2018.pdf.

[48] "Harpy NG," Israel Aerospace Industries. http://www.iai.co.il/2013/36694-16153-en/Business_Areas_Land.aspx.

[49] Jean-Baptiste Jeangene Vilmer, "The French Turn to Armed Drones," *War on the Rocks*, September 22, 2017, https://warontherocks.com/2017/09/the-french-turn-to-armed-drones/.

[50] "Taranis," BAE Systems, 2018, https://www.baesystems.com/en/product/taranis.

[51] Cheryl Pellerin, "Project Maven to Deploy Computer Algorithms to War Zone by Year's End," *DoD News*, Defense Media Activity, July 21, 2017, https://dod.defense.gov/News/Article/Article/1254719/project-maven-to-deploy-computer-algorithms-to-war-zone-by-years-end/.

[52] "Pentagon Unmanned Systems Integrated Roadmap 2017-2042," *USNI News*, August 30, 2018, https://news.usni.org/2018/08/30/pentagon-unmanned-systems-integrated-roadmap-2017-2042.

[53] Patrick Tucker, "Pentagon Seeks a List of Ethical Principles for Using AI in War," *Defense One*, January 4, 2019, https://www.defenseone.com/technology/2019/01/pentagon-seeks-list-ethical-principles-using-ai-war/153940/.

[54] Gavin Pearson, Phil Jolley, and Geraint Evans, "A Systems Approach to Achieving the Benefits of Artificial Intelligence in UK Defence," *arXiv*, September 28, 2018, https://arxiv.org/abs/1809.11089.

[55] Samuel C. Woolley and Philip N. Howard, "Computational Propaganda Worldwide: Executive Summary," Computational Propaganda Research Project, University of Oxford, June 2017, http://comprop.oii.ox.ac.uk/wp-content/uploads/sites/89/2017/06/Casestudies-ExecutiveSummary.pdf.

[56] Doug Wise, "Could AI-Driven Info Warfare Be Democracy's Achilles Heel?," *The Cipher Brief*, March 9, 2018, https://www.thecipherbrief.com/column_article/ai-driven-info-warfare-democracys-achilles-heel.

**57** Darrell M. West, "How to combat fake news and disinformation," Brookings Institution, December 18, 2017, https://www.brookings.edu/research/how-to-combat-fake-news-and-disinformation/.

**58** Douglas Guilbeault and Samuel Woolley, "How Twitter Bots Are Shaping the Election," *The Atlantic*, November 1, 2016, https://www.theatlantic.com/technology/archive/2016/11/election-bots/506072/.

**59** Michael Newberg, "As many as 48 million Twitter accounts aren't people, says study," *CNBC News*, March 10, 2017, https://www.cnbc.com/2017/03/10/nearly-48-million-twitter-accounts-could-be-bots-says-study.html.

**60** James Vincent, "A porn company promises to insert customers into scenes using deepfakes," *The Verge*, August, 21, 2018, https://www.theverge.com/2018/8/21/17763278/deepfake-porn-custom-clips-naughty-america.

**61** Bianca Bosker, "The Binge Breaker," *The Atlantic*, November 2016, https://www.theatlantic.com/magazine/archive/2016/11/the-binge-breaker/501122/.

**62** Stefan Hall and Ryo Takahashi, "Augmented and virtual reality: the promise and peril of immersive technologies," World Economic Forum, September 8, 2017, https://www.weforum.org/agenda/2017/09/augmented-and-virtual-reality-will-change-how-we-create-and-consume-and-bring-new-risks/.

**63** William D. Eggers, David Schatsky, and Peter Viechnicki, "AI-augmented government using cognitive technologies to redesign public sector work," Deloitte Center for Government Insights, *Deloitte University Press*, 2017, https://www2.deloitte.com/content/dam/insights/us/articles/3832_AI-augmented-government/DUP_AI-augmented-government.pdf.

**64** Richard Stirling, Hannah Miller and Emma Martinho-Truswell, "Government AI Readiness Index," Oxford Insights, April 26, 2018, https://www.oxfordinsights.com/government-ai-readiness-index.

**65** Emily Dreyf, "The US Government Isn't Just Tech-Illiterate. It's Tech-Incompetent," *Wired*, May 11, 2017, https://www.wired.com/2017/05/real-threat-government-tech-illiteracy/.

**66** Erin Winick, "It's Recruiting Season for AI's Top Talent, and Things Are Getting a Little Zany," *MIT Technology Review*, December 6, 2017, https://www.technologyreview.com/the-download/609707/its-recruiting-season-for-ais-top-talent-and-things-are-getting-a-little-zany/.

**67** Tajha Chappellet-Lanier, "Senators introduce the 'Artificial Intelligence in Government Act'," *FedScoop*, September 26, 2018, https://www.fedscoop.com/artificial-intelligence-in-government-act/.

**68** Dom Galeon and Chelsea Gohd, "Dubai Just Appointed a "State Minister for Artificial Intelligence"," *Futurism*, October 20, 2017, https://futurism.com/dubai-just-appointed-a-state-minister-for-artificial-intelligence/.

**69** Nikolay Nikolov, "The tech-savvy diplomat that's rewriting the playbook to international relations," *Mashable*, April 10, 2018, https://mashable.com/2018/04/10/casper-klynge-tech-ambassador-silicon-valley-denmark.

**70** Amanda Russo, "United Kingdom Partners with World Economic Forum to Develop First Artificial Intelligence Procurement Policy," World Economic Forum, September 20, 2018, https://www.weforum.org/press/2018/09/united-kingdom-partners-with-world-economic-forum-to-develop-first-artificial-intelligence-procurement-policy/.

**71** Paul Mozur, "Beijing Wants A.I. to Be Made in China by 2030," *The New York Times*, July 20, 2017, https://www.nytimes.com/2017/07/20/business/china-artificial-intelligence.html.

**72** Chong Koh Ping, "China wants to work with other countries to develop AI," *The Straits Times*, September 17, 2018, https://www.straitstimes.com/asia/east-asia/china-wants-to-work-with-other-countries-to-develop-ai.

**73** Ian Hogarth, "AI Nationalism," Ian Hogarth Blog, June 13, 2018, https://www.ianhogarth.com/blog/2018/6/13/ai-nationalism.

**74** "National and International AI Strategies," Future of Life Institute, 2018, https://futureoflife.org/national-international-ai-strategies.

**75** Allan Dafoe, "AI Governance: A Research Agenda," Governance of AI Program, Future of Humanity Institute, University of Oxford, August 27, 2018, https://www.fhi.ox.ac.uk/wp-content/uploads/AI-Governance_-A-Research-Agenda.pdf.

**76** Tara Francis Chan, "Parts of China are using facial recognition technology that can scan the country's entire population in one second," *Business Insider*, March 26, 2018, https://www.businessinsider.com/china-facial-recognition-technology-works-in-one-second-2018-3.

**77** James Vincent, "This Japanese AI security camera shows the future of surveillance will be automated," *The Verge*, June 26, 2018, https://www.theverge.com/2018/6/26/17479068/ai-guardman-security-camera-shoplifter-japan-automated-surveillance.

**78** George Joseph, "Extreme Digital Vetting of Visitors to the U.S. Moves Forward Under a New Name," *ProPublica*, November 22, 2017, https://www.propublica.org/article/extreme-digital-vetting-of-visitors-to-the-u-s-moves-forward-under-a-new-name.

**79** Walter L. Perry, et al., "Predictive Policing," RAND Corporation, 2013, https://www.rand.org/content/dam/rand/pubs/research_reports/RR200/RR233/RAND_RR233.sum.pdf.

**80** Andrew D. Selbst, "Disparate Impact in Big Data Policing," Data & Society Research Institute, Yale Information Society Project, February 25, 2017, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2819182.

**81** "Emerging Technology And National Security," 2018 Analytic Exchange Program, Department of Homeland Security, July 26, 2018, https://www.dhs.gov/sites/default/files/publications/2018_AEP_Emerging_Technology_and_National_Security.pdf.

**82** James Vincent, "Amazon employees protest sale of facial recognition software to police 18," *The Verge*, June 22, 2018, https://www.theverge.com/2018/6/22/17492106/amazon-ice-facial-recognition-internal-letter-protest.

**83** Daisuke Wakabayashi and Scott Shane, "Google Will Not Renew Pentagon Contract That Upset Employees," *The New York Times*, June 1, 2018, https://www.nytimes.com/2018/06/01/technology/google-pentagon-project-maven.html.

**84** Sundar Pichai, "AI at Google: our principles," *The Keyword*, Google, June 7, 2018, https://www.blog.google/technology/ai/ai-principles/.

**85** "High-Level Expert Group on Artificial Intelligence," Digital Single Market, European Commission, June 14, 2018, https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence.

**86** "Draft Ethics Guidelines for Trustworthy AI," The European Commission's High-Level Expert Group on Artificial Intelligence, December 18, 2018, https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_draft_ethics_guidelines_18_december.pdf.

**87** "Mandate for the International Panel on Artificial Intelligence," Justin Trudeau, Prime Minister of Canada, December 6, 2018, https://pm.gc.ca/eng/news/2018/12/06/mandate-international-panel-artificial-intelligence.

**88** "Putting faces to the jobs at risk of automation," OECD, March 2018, https://www.oecd.org/employment/Automation-policy-brief-2018.pdf.

**89** Carl Benedikt Frey and Michael A. Osborne. "The Future of Employment: How Susceptible are Jobs to Computerisation?" Oxford University, 2013, https://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf.

**90** James Manyika, et al., "Jobs lost, jobs gained: What the future of work will mean for jobs, skills, and wages," McKinsey Global Institute, November 2017, https://www.mckinsey.com/featured-insights/future-of-organizations-and-work/Jobs-lost-jobs-gained-what-the-future-of-work-will-mean-for-jobs-skills-and-wages.

[91] Richard Wike and Bruce Stokes, “In Advanced and Emerging Economies Alike, Worries About Job Automation,” Pew Research Center, September 13, 2018, http://www.pewglobal.org/2018/09/13/in-advanced-and-emerging-economies-alike-worries-about-job-automation/.

[92] Jane Wakefield, “Foxconn replaces ‘60,000 factory workers with robots,” *BBC News*, May 25, 2016, https://www.bbc.com/news/technology-36376966.

[93] Conner Forrest, “Chinese factory replaces 90% of humans with robots, production soars,” *TechRepublic*, July 30, 2015, https://www.techrepublic.com/article/chinese-factory-replaces-90-of-humans-with-robots-production-soars/.

[94] “Robot density rises globally,” International Federation of Robotics, February 7, 2018, https://ifr.org/ifr-press-releases/news/robot-density-rises-globally.

[95] Bettina Büchel and Dario Floreano, “Tesla s problem: overestimating automation, underestimating humans,” *The Conversation*, May 2, 2018, https://theconversation.com/teslas-problem-overestimating-automation-underestimating-humans-95388.

[96] Derek Thompson, “A World Without Work,” *The Atlantic*, July/August 2015 Issue, https://www.theatlantic.com/magazine/archive/2015/07/world-without-work/395294/.

[97] Chairman Will Hurd and Ranking Member Robin Kelly, “Rise of the Machines,” Subcommittee on Information Technology, Committee on Oversight and Government Reform, U.S. House of Representatives, September 2018, https://oversight.house.gov/wp-content/uploads/2018/09/AI-White-Paper-.pdf.

[98] Sebastian Sprenger, “Germany wants its own version of DARPA, and within the year,” *DefenseNews*, July 18, 2018, https://www.defensenews.com/global/europe/2018/07/18/germany-wants-its-own-version-of-darpa-and-within-the-year/.

[99] Adam J. Gustein and John Sviokla, “7 Skills That Aren’t About to Be Automated,” *Harvard Business Review*, July 17, 2018, https://hbr.org/2018/07/7-skills-that-arent-about-to-be-automated.

[100] Cedric Villani, “For a Meaningful Artificial Intelligence Towards a French and European Strategy,” March 8, 2018, https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf.

[101] “Equipping people to stay ahead of technological change,” *The Economist*, January 14, 2017, https://www.economist.com/leaders/2017/01/14/equipping-people-to-stay-ahead-of-technological-change.

[102] Alastair Fitzpayne and Ethan Pollack, “Lifelong Learning and Training Accounts helping workers adapt and succeed in a changing economy,” The Aspen Institute, May 2018, https://assets.aspeninstitute.org/content/uploads/2018/05/Lifelong-Learning-and-Training-Accounts-Issue-Brief.pdf.

[103] “70% of Value in Tech is Driven by Network Effects,” *NFX*, November 28, 2017, https://medium.com/@nfx/70-of-value-in-tech-is-driven-by-network-effects-8c4788528e35.

[104] Blaise Zerega, “AI Weekly: Google shifts from mobile-first to AI-first world,” *Venture Beat*, May 18, 2017, https://venturebeat.com/2017/05/18/ai-weekly-google-shifts-from-mobile-first-to-ai-first-world/.

[105] Marisa Kendall, “Income inequality in the Bay Area is among nation’s highest,” *The Mercury News*, February 15, 2018, https://www.mercurynews.com/2018/02/15/income-inequality-in-the-bay-area-is-among-nations-highest/.

[106] Kai-Fu Lee, “AI Could Devastate the Developing World,” *Bloomberg Opinion*, September 17, 2018, https://www.bloombergquint.com/onweb/2018/09/17/artificial-intelligence-threatens-jobs-in-developing-world.

[107] “New UK AI Report Warns Against Data Monopolies,” DataEthics, April 19, 2018, https://dataethics.eu/en/new-uk-ai-report-warns-data-monopolies/.

[108] “Is Google Becoming A Monopoly?” *Investopedia*, 2018, https://www.investopedia.com/articles/investing/060315/google-becoming-monopoly.asp.

**109** Obert Wright, "Google Must be Stopped Before it Becomes an AI Monopoly," *Wired*, February 23, 2018, https://www.wired.com/story/google-artificial-intelligence-monopoly/.

**110** Finale Doshi-Velez, et al., "Accountability of AI Under the Law: The Role of Explanation," Berkman Klein Center for Internet & Society, Harvard University, November 27, 2017, https://cyber.harvard.edu/publications/2017/11/AIExplanation.

**111** Joshua A. Kroll, et al., "Accountable Algorithms," *University of Pennsylvania Law Review*, Vol. 165, 2017, Fordham Law Legal Studies Research Paper No. 2765268, November 20, 2016, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2765268.

**112** "State v. Loomis," *Harvard Law Review*, March 10, 2017, https://harvardlawreview.org/2017/03/state-v-loomis/.

**113** Sara Castellanos, "Tech Giants Launch New AI Tools as Worries Mount About Explainability," *The Wall Street Journal*, September 26, 2018, https://blogs.wsj.com/cio/2018/09/26/tech-giants-launch-new-ai-tools-as-worries-mount-about-explainability/.

**114** "Algorithmic Accountability Policy Toolkit," AI Now, October 2018, https://ainowinstitute.org/aap-toolkit.pdf.

**115** David Meyer, "AI Has a Big Privacy Problem and Europe's New Data Protection Law Is About to Expose It," *Fortune*, May 25, 2018, http://fortune.com/2018/05/25/ai-machine-learning-privacy-gdpr/.

**116** Kristen J. Mathews and Courtney M. Bowman, "The California Consumer Privacy Act of 2018," *Privacy Law Blog*, Proskauer, July 13, 2018, https://privacylaw.proskauer.com/2018/07/articles/data-privacy-laws/the-california-consumer-privacy-act-of-2018/.

**117** Kevin Granville, "Facebook and Cambridge Analytica: What You Need to Know as Fallout Widens," *The New York Times*, March 19, 2018, https://www.nytimes.com/2018/03/19/technology/facebook-cambridge-analytica-explained.html.

**118** Guy Rosen, "Security Update," Facebook Newsroom, September 28, 2018, https://newsroom.fb.com/news/2018/09/security-update/.

**119** Eleonore Pauwels, "Nowhere to Hide: Artificial Intelligence and Privacy in the Fourth Industrial Revolution," Wilson Center, April 2, 2018, https://www.wilsoncenter.org/article/nowhere-to-hide-artificial-intelligence-and-privacy-the-fourth-industrial-revolution.

**120** Farid Gueham, "Digital Sovereignty," The Fondation pour l'innovation politique, January 2017, https://euagenda.eu/publications/digital-sovereignty-steps-towards-a-new-system-of-internet-governance.

**121** "Use cases for 'data trusts' to be explored in the UK," Out-Law.com, November 22, 2018, https://www.out-law.com/en/articles/2018/november/use-cases-data-trusts-uk/.

**122** Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, "Machine Bias," *ProPublica*, May 23, 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

**123** Nikhil Sonnad, "Google Translate's gender bias pairs "he" with "hardworking" and "she" with lazy, and other examples," *Quartz*, November 29, 2017, https://qz.com/1141122/google-translates-gender-bias-pairs-he-with-hardworking-and-she-with-lazy-and-other-examples/.

**124** Ashley Rodriguez, "Microsoft's AI millennial chatbot became a racist jerk after less than a day on Twitter," *Quartz*, March 24, 2016, https://qz.com/646825/microsofts-ai-millennial-chatbot-became-a-racist-jerk-after-less-than-a-day-on-twitter/.

**125** Leila Mead, "Global Summit Focuses on the Role of Artificial Intelligence in Advancing SDGs," SDG Knowledge Hub, A Project by IISD, June 5, 2018, https://sdg.iisd.org/news/global-summit-focuses-on-the-role-of-artificial-intelligence-in-advancing-sdgs/.

**126** Maya Kosoff, “China’s Terrifying Surveillance State Looks a Lot Like America’s Future,” *Vanity Fair*, July 9, 2018, https://www.vanityfair.com/news/2018/07/china-surveillance-state-artificial-intelligence.

**127** Tim Urban, “Neuralink and the Brain’s Magical Future, *Wait But Why*, April 20, 2017, https://waitbutwhy.com/2017/04/neuralink.html.

**128** Rafael Yuste, et al., “Four ethical priorities for neurotechnologies and AI,” *Nature*, November 8, 2017, https://www.nature.com/news/four-ethical-priorities-for-neurotechnologies-and-ai-1.22960.

**129** Joichi Ito, “Resisting Reduction: A Manifesto,” *JoDS*, November 14, 2017, https://jods.mitpress.mit.edu/pub/resisting-reduction.

**130** Mathias Risse, “Human Rights and Artificial Intelligence: An Urgently Needed Agenda,” Carr Center for Human Rights Policy, Harvard Kennedy School, May 2018, https://carrcenter.hks.harvard.edu/files/cchr/files/humanrightsai_designed.pdf.

**131** Tony Robinson, “Artificial intelligence and the impact on our data centers,” DCD, July 31, 2018, https://www.datacenterdynamics.com/analysis/artificial-intelligence-and-the-impact-on-our-data-centers/.

**132** Pierre Delforge, “America’s Data Centers Consuming and Wasting Growing Amounts of Energy,” NRDC, February 06, 2015, https://www.nrdc.org/resources/americas-data-centers-consuming-and-wasting-growing-amounts-energy.

**133** Tom Bawden, “Global warming: Data centres to consume three times as much energy in next decade, experts warn,” *The Independent*, January 23, 2016, https://www.independent.co.uk/environment/global-warming-data-centres-to-consume-three-times-as-much-energy-in-next-decade-experts-warn-a6830086.html.

**134** “From Smart to Senseless: The Global Impact of 10 Years of Smartphones,” Greenpeace, February 2017, http://www.greenpeace.org/usa/wp-content/uploads/2017/03/FINAL-10YearsSmartphones-Report-Design-230217-Digital.pdf.

**135** *Id.*

**136** “DeepMind AI Reduces Google Data Centre Cooling Bill by 40%,” DeepMind, July 20, 2016, https://deepmind.com/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-40/.

**137** Mark Sullivan, “Apple Now Runs On 100% Green Energy, And Here’s How It Got There,” *Fast Company*, April 9, 2018, https://www.fastcompany.com/40554151/how-apple-got-to-100-renewable-energy-the-right-way.

**138** Cedric Villani, “For a Meaningful Artificial Intelligence Towards a French and European Strategy,” March 8, 2018, https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf.

**139** “AI Sector Deal,” Department for Business, Energy & Industrial Strategy and Department for Digital, Culture, Media & Sport, UK Government, April 26, 2018, https://www.gov.uk/government/publications/artificial-intelligence-sector-deal/ai-sector-deal.

**140** “National and International AI Strategies,” The Future of Life Institute, 2018, https://futureoflife.org/national-international-ai-strategies/.

**141** Brad Smith, “Facial recognition technology: The need for public regulation and corporate responsibility,” Microsoft, July 13, 2018, https://blogs.microsoft.com/on-the-issues/2018/07/13/facial-recognition-technology-the-need-for-public-regulation-and-corporate-responsibility/.

**142** James Vincent, “Elon Musk says we need to regulate AI before it becomes a danger to humanity,” *The Verge*, July 17, 2017, https://www.theverge.com/2017/7/17/15980954/elon-musk-ai-regulation-existential-threat.

**143** Google, “Perspectives on Issues in AI Governance,” January 22 2019,https://ai.google/static/documents/perspectives-on-issues-in-ai-governance.pdf.

**144** Xie Yu and Meng Jing, "China aims to outspend the world in artificial intelligence, and Xi Jinping just green lit the plan," *South China Morning Post*, October 18, 2017, https://www.scmp.com/business/china-business/article/2115935/chinas-xi-jinping-highlights-ai-big-data-and-shared-economy.

**145** "Notice of the State Council Issuing the New Generation of Artificial Intelligence Development Plan," State Council of China, July 8, 2017, https://flia.org/wp-content/uploads/2017/07/A-New-Generation-of-Artificial-Intelligence-Development-Plan-1.pdf.

**146** "The 13th Five-Year Plan," U.S.-China Economic and Security Review Commision, February 14, 2017, https://www.uscc.gov/Research/13th-five-year-plan.

**147** Paul Triolo, Elsa Kania, and Graham Webster, "Translation: Chinese government outlines AI ambitions through 2020," New America, January 26, 2018, https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-chinese-government-outlines-ai-ambitions-through-2020/.

**148** Nicholas Thompson, "Emmanuel Macron Talks to Wired about France's AI Strategy," *Wired*, March 31, 2018, https://www.wired.com/story/emmanuel-macron-talks-to-wired-about-frances-ai-strategy/.

**149** *Id.*

**150** "#FranceIA: the national artificial intelligence strategy is underway," *Gouvernement.fr*, Le Gouvernement Republique Francaise, January 26, 2017, https://www.gouvernement.fr/en/franceia-the-national-artificial-intelligence-strategy-is-underway.

**151** "Rapport de Synthèse: France Intelligence Artificielle," France Intelligence Artificielle, Le Gouvernement Republique Francaise, March 2017, https://www.economie.gouv.fr/files/files/PDF/2017/Rapport_synthese_France_IA_.pdf.

**152** "How Can Humans Keep the Upper Hand? The ethical matters raised by algorithms and artificial intelligence," National Commission for Information Technology and Liberties, December 2017, https://www.cnil.fr/sites/default/files/atoms/files/cnil_rapport_ai_gb_web.pdf.

**153** Zachary Young, "French Parliament passes law against 'fake news'," *Politico*, July 4, 2018, https://www.politico.eu/article/french-parliament-passes-law-against-fake-news/.

**154** Janosch Delcker, "France, Germany under fire for failing to back 'killer robots' ban," *Politico*, March 28, 2018, https://www.politico.eu/article/artificial-intelligence-killer-robots-france-germany-under-fire-for-failing-to-back-robots-ban/.

**155** "Digital Charter," Department for Digital, Culture, Media & Sport, UK Government, January 25, 2018, https://www.gov.uk/government/publications/digital-charter/digital-charter.

**156** "UK Partners With World Economic Forum To Develop First Artificial Intelligence Procurement Policy," *Eurasia Review*, September 21, 2018, https://www.eurasiareview.com/21092018-uk-partners-with-world-economic-forum-to-develop-first-artificial-intelligence-procurement-policy/.

**157** "Industrial Strategy: Building a Britain fit for the future," UK Government, November 2017, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/664563/industrial-strategy-white-paper-web-ready-version.pdf.

**158** "Robotics and artificial intelligence," House of Commons Science and Technology Committee, UK Parliament, September 13, 2016, https://publications.parliament.uk/pa/cm201617/cmselect/cmsctech/145/145.pdf.

**159** "APPG AI Findings 2017: 1 Key Recommendation 6 Policy Focus Areas," All-Party Parliamentary Group on Artificial Intelligence, UK Parliament, 2017, http://www.appg-ai.org/wp-content/uploads/2017/12/appgai_2017_findings.pdf.

**160** "Government response to House of Lords Artificial Intelligence Select Committee's Report on AI in the UK: Ready, Willing and Able?" Secretary of State for Business, Energy and Industrial Strategy, UK Government, June 2018, https://www.parliament.uk/documents/lords-committees/Artificial-Intelligence/AI-Government-Response.pdf.

[161] Richard Stirling, Hannah Miller and Emma Martinho-Truswell, "Government AI Readiness Index," Oxford Insights, 2017, https://www.oxfordinsights.com/government-ai-readiness-index.

[162] "Government response to House of Lords Artificial Intelligence Select Committee's Report on AI in the UK: Ready, Willing and Able?" Secretary of State for Business, Energy and Industrial Strategy, UK Government, June 2018, https://www.parliament.uk/documents/lords-committees/Artificial-Intelligence/AI-Government-Response.pdf.

[163] Cade Metz, "Artificial Intelligence Is Now a Pentagon Priority. Will Silicon Valley Help?" *The New York Times*, August 26, 2018. https://www.nytimes.com/2018/08/26/technology/pentagon-artificial-intelligence.html.

[164] Mick Mulvaney and Michael Kratsios, "FY 2020 Administration Research and Development Budget Priorities," Executive Office of the President, July 31, 2018, https://www.whitehouse.gov/wp-content/uploads/2018/07/M-18-22.pdf.

[165] Jesse Ellman, et al., "Assessing the Third Offset Strategy," Center for Strategic & International Studies, March 2017. https://csis-prod.s3.amazonaws.com/s3fs-public/publication/170302_Ellman_ThirdOffsetStrategySummary_Web.pdf.

[166] Jade Leung and Sophie-Charlotte Fischer, "JAIC: Pentagon debuts artificial intelligence hub," *Bulletin of the Atomic Scientists*, August 8, 2018, https://thebulletin.org/2018/08/jaic-pentagon-debuts-artificial-intelligence-hub/.

[167] Sydney J. Freedberg Jr., "Faster Than Thought: DARPA, Artificial Intelligence, & The Third Offset Strategy," *Breaking Defense*, February 11, 2016, https://breakingdefense.com/2016/02/faster-than-thought-darpa-artificial-intelligence-the-third-offset-strategy/.

[168] "National Security Strategy of the United States of America," United States White House, December 2017, https://partner-mco-archive.s3.amazonaws.com/client_files/1513628003.pdf.

[169] Jim Mattis, "Summary of the 2018 National Defense Strategy of the United States of America," United States Department of Defense, 2018, https://dod.defense.gov/Portals/1/Documents/pubs/2018-National-Defense-Strategy-Summary.pdf.

[170] Jay Cassano, "Pentagon's Artificial Intelligence Programs Get Huge Boost in the NDAA," *Sludge*, August 15, 2018, https://readsludge.com/2018/08/15/pentagons-artificial-intelligence-programs-get-huge-boost-in-the-ndaa/.

[171] "Defense Innovation Board Recommendations," Defense Innovation Board, United States Department of Defense, 2016, https://media.defense.gov/2017/Dec/18/2001857962/-1/-1/0/2017-2566-148525_RECOMMENDATION%2012_(2017-09-19-01-45-51).PDF.

[172] Sydney J. Freedberg Jr., "AI Logistics Let Combat Units Move Faster: Uptake's DIUX Contract," *Breaking Defense*, June 27, 2028, https://breakingdefense.com/2018/06/ai-logistics-can-speed-up-army-tanks-uptakes-diux-contract/.

[173] "Establishment of an Algorithmic Warfare Cross-Functional Team (Project Maven)" Memorandum, Deputy Secretary of Defense, April 26, 2017, https://www.govexec.com/media/gbc/docs/pdfs_edit/establishment_of_the_awcft_project_maven.pdf.

[174] Scott Shane and Daisuke Wakabayashi, "'The Business of War': Google Employees Protest Work for the Pentagon," *The New York Times*, April 4, 2018, https://www.nytimes.com/2018/04/04/technology/google-letter-ceo-pentagon-project.html.

[175] Ryan Whitwam, "Google Will End 'Project Maven' Pentagon Drone Contract," *Extreme Tech*, June 4, 2018, https://www.extremetech.com/internet/270524-google-will-end-project-maven-pentagon-drone-contract.

[176] Sundar Pichai, "AI at Google: our principles," Google, *The Keyword*, June 7, 2018, https://www.blog.google/technology/ai/ai-principles/.

[177] "Establishment of the Joint Artificial Intelligence Center," Deputy Secretary of Defense, June 27, 2018, https://admin.govexec.com/media/establishment_of_the_joint_artificial_intelligence_center_osd008412-18_r....pdf.

[178] "Ethical Concerns Now in Equation for DoD's Use of AI," *MeriTalk*, July 25, 2018, https://www.meritalk.com/articles/ethical-concerns-now-in-equation-for-dods-use-of-ai/.

**179** Jay Cassano, “Pentagon’s Artificial Intelligence Programs Get Huge Boost in the NDAA,” *Sludge*, August 15, 2018, https://readsludge.com/2018/08/15/pentagons-artificial-intelligence-programs-get-huge-boost-in-the-ndaa/.

**180** “H.R.5515 - John S. McCain National Defense Authorization Act for Fiscal Year 2019,” 115th Congress, August 13, 2018, https://www.congress.gov/bill/115th-congress/house-bill/5515/text.

**181** Drew Harwell, “Defense Department pledges billions toward artificial intelligence research,” *The Washington Post*, September 7, 2018, https://www.washingtonpost.com/technology/2018/09/07/defense-department-pledges-billions-toward-artificial-intelligence-research/?utm_term=.1e1dfd2c7c61.

**182** *Id.*

**183** “AI Next Campaign,” DARPA, 2018, https://www.darpa.mil/work-with-us/ai-next-campaign.

**184** “FY 2020 Administration Research and Development Budget Priorities,” Executive Office of the President, July 31, 2018, https://www.whitehouse.gov/wp-content/uploads/2018/07/M-18-22.pdf.

**185** “Artificial Intelligence for the American People,” White House Fact Sheet, May 10, 2018, https://www.whitehouse.gov/briefings-statements/artificial-intelligence-american-people/.

**186** “Summary of the 2018 White House Summit on Artificial Intelligence for American Industry,” The White House Office of Science and Technology Policy, May 10, 2018, https://www.whitehouse.gov/wp-content/uploads/2018/05/Summary-Report-of-White-House-AI-Summit.pdf.

**187** “The National Artificial Intelligence Research and Development Strategic Plan,” National Science and Technology Council and Networking and Information Technology Research and Development Subcommittee, October 2016, https://www.nitrd.gov/PUBS/national_ai_rd_strategic_plan.pdf.

**188** “Request for Information on Update to the 2016 National Artificial Intelligence Research and Development Strategic Plan,” The Networking and Information Technology Research and Development Program, September 26, 2018, https://www.nitrd.gov/news/RFI-National-AI-Strategic-Plan.aspx.

**189** David Gunning, “Explainable Artificial Intelligence (XAI),” DARPA, 2018, https://www.darpa.mil/program/explainable-artificial-intelligence.

**190** “Game Changers: Artificial Intelligence Part III, Artificial Intelligence and Public Policy,” Subcommittee on Information Technology, April 18, 2018, https://oversight.house.gov/hearing/game-changers-artificial-intelligence-part-iii-artificial-intelligence-and-public-policy/.

**191** “Joint Subcommittee Hearing: “Artificial Intelligence – With Great Power Comes Great Responsibility”,” U.S. House of Representatives Committee on Science, Space, and Technology, June 26, 2018, https://science.house.gov/sites/republicans.science.house.gov/files/documents/HHRG-115-SY15-20180626-SD001.pdf.

**192** “H.R.4625 - FUTURE of Artificial Intelligence Act of 2017,” 115th Congress (2017-2018), December 12, 2017, https://www.congress.gov/bill/115th-congress/house-bill/4625/text.

**193** “Delaney Launches Bipartisan Artificial Intelligence (AI) Caucus for 115th Congress,” John K. Delaney Press Release, May 24, 2017, https://delaney.house.gov/news/press-releases/delaney-launches-bipartisan-artificial-intelligence-ai-caucus-for-115th-congress.

**194** “Artificial Intelligence Task Force,” State of Vermont, Agency of Commerce and Community Development, 2018, https://accd.vermont.gov/economic-development/artificial-intelligence-task-force.

**195** Sidney Fussell, “NYC Launches Task Force to Study How Government Algorithms Impact Your Life,” *Gizmodo*, May 16, 2018, https://gizmodo.com/nyc-launches-task-force-to-study-how-government-algorit-1826087643.

**196** “State of California Endorses Asilomar AI Principles,” Future of Life Institute Press Release, August 31, 2018, https://futureoflife.org/2018/08/31/state-of-california-endorses-asilomar-ai-principles/.

**197** Greg Sterling, “California’s new bot law forces companies to tell you when you’re interacting with a machine,” *Martech*, October 3, 2018, https://martechtoday.com/californias-new-bot-law-forces-companies-to-tell-you-when-youre-interacting-with-a-machine-226138.

**198** “Pan-Canadian Artificial Intelligence Strategy,” The Canadian Institute for Advanced Research, March 2017, https://www.cifar.ca/ai/pan-canadian-artificial-intelligence-strategy.

**199** “President Macron on historic commitment to create international study group on inclusive and ethical AI,” Canadian Institute for Advanced Research, June 7, 2018, https://www.newswire.ca/news-releases/cifar-congratulates-prime-minister-trudeau-and-president-macron-on-historic-commitment-to-create-international-study-group-on-inclusive-and-ethical-ai-684842541.html.

**200** “The Prime Minister in Action,” Council for Science, Technology and Innovation, September 15, 2016, https://japan.kantei.go.jp/97_abe/actions/201609/15article2.html.

**201** “Artificial Intelligence Technology Strategy,” The Strategic Council for AI Technology, March 31, 2017, http://www.nedo.go.jp/content/100865202.pdf.

**202** “AI researchers to be focus of government’s ‘integrated innovation strategy’,” *The Japan Times*, June 3, 2018, https://www.japantimes.co.jp/news/2018/06/03/national/ai-researchers-focus-governments-integrated-innovation-strategy/.

**203** “Speech by Mr S Iswaran, Minister for Communications and Information at the Innovfest Unbound 2018,” Ministry of Communications and Information, UAE Government, June 5, 2018, https://www.mci.gov.sg/pressroom/news-and-stories/pressroom/2018/6/speech-by-mr-s-iswaran-at-the-innovfest-unbound-2018-on-5-june-2018.

**204** “AI Singapore,” National Research Foundation, Prime Ministers Office, Singapore, May 2017, https://www.nrf.gov.sg/programmes/artificial-intelligence-r-d-programme.

**205** Chia Jie Lin, “Singapore sets up AI ethics council,” *GovInsider*, June 6, 2018, https://govinsider.asia/innovation/singapore-sets-ai-ethics-council/.

**206** “Mid-to Long-term Master Plan in Preparation for the Intelligent Information Society: Managing the Fourth Industrial Revolution,” Government of the Republic of Korea, 2016, https://english.msit.go.kr/cms/english/pl/policies2/__icsFiles/afieldfile/2017/07/20/Master%20Plan%20for%20the%20intelligent%20information%20society.pdf.

**207** “Korean AI, 1.8 years behind US, China to overtake as Government gives 2.2 trillion won,” *Joongang Daily*, May 5, 2018, https://news.joins.com/article/22625271.

**208** Sherouk Zakaria, “Laws will be in place to regulate use of AI: Minister,” *Khaleej Times*, December 7, 2017, https://www.khaleejtimes.com/nation/laws-will-be-in-place-to-regulate-use-of-ai-minister.

**209** “UAE Strategy for Artificial Intelligence,” *Government.ae*, UAE Government, October 2017, https://government.ae/en/about-the-uae/strategies-initiatives-and-awards/federal-governments-strategies-and-plans/uae-strategy-for-artificial-intelligence.

**210** “Summary Report 2018: Global Governance of AI Roundtable,” World Government Summit, 2018, https://www.worldgovernmentsummit.org/api/publications/document?id=ff6c88c5-e97c-6578-b2f8-ff0000a7ddb6.

**211** Sundar Pichai, “AI at Google: our principles,” *The Keyword*, Google, June 7, 2018, https://www.blog.google/technology/ai/ai-principles/.

**212** Richard T. De George, “A History of Business Ethics,” Markkula Center for Applied Ethics, November 17, 2015, https://www.scu.edu/ethics/focus-areas/business-ethics/resources/a-history-of-business-ethics/.

**213** Yoav Shoham, Raymond Perrault, Erik Brynjolfsson, Jack Clark, James Manyika, Juan Carlos Niebles, Terah Lyons, John Etchemendy, Barbara Grosz and Zoe Bauer, “The AI Index 2018 Annual Report,” AI Index Steering Committee, Human-Centered AI Initiative, Stanford University, Stanford, CA, December 2018. (c) 2018 by Stanford University, “The AI Index 2018 Annual Report” is made available under a Creative Commons Attribution-NoDerivatives 4.0 License (International) https://creativecommons.org/licenses/by-nd/4.0/legalcode

CLTC

Center for Long-Term Cybersecurity

UC Berkeley

Center for Long-Term Cybersecurity
cltc.berkeley.edu
@CLTCBerkeley