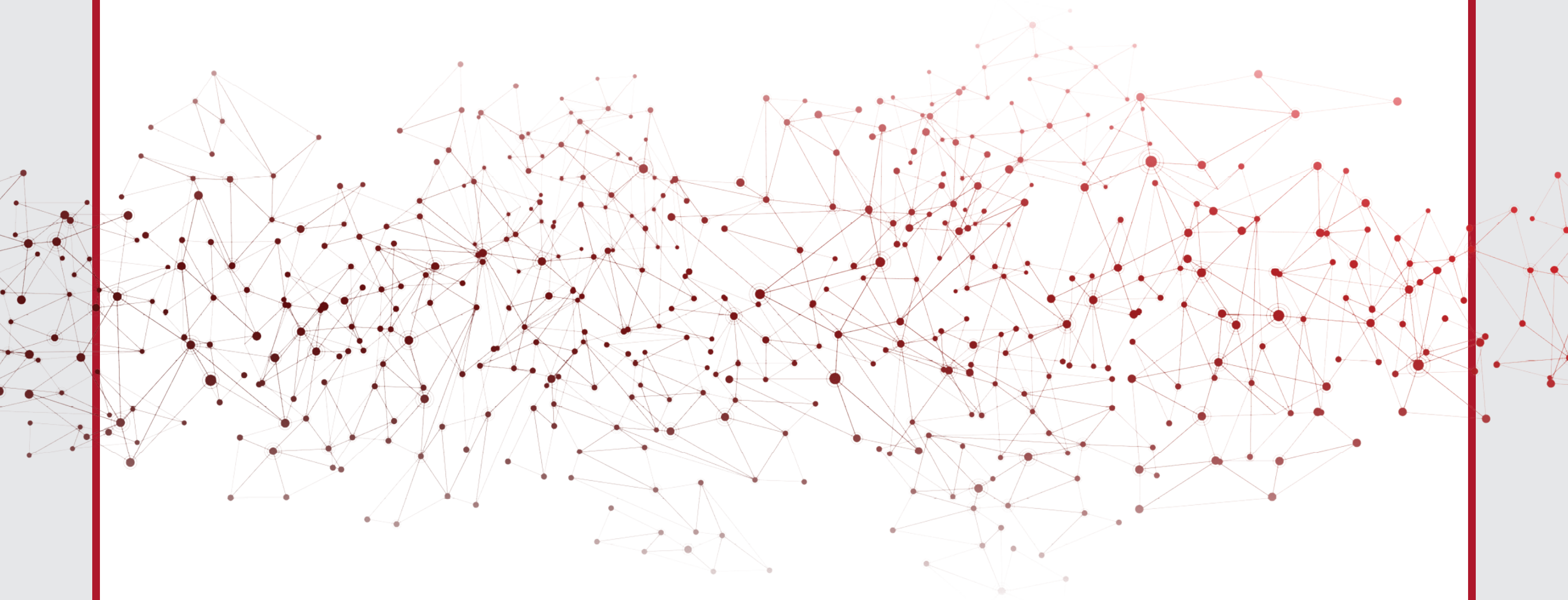
TECHNOLOGY PRIMERS FOR POLICYMAKERS

Artificial Intelligence & Machine Learning



HARVARD Kennedy School
BELFER CENTER
for Science and International Affairs
TECHNOLOGY AND PUBLIC PURPOSE PROJECT

April 2023

AUTHORS

Issam Eddine Abail (Harvard Kennedy School)
Gopal Nadadur (The Asia Group)
Enrico Santus (Bloomberg)
Ariel Higuchi, *Tech Primers Project Lead* (Harvard Kennedy School)
Amritha Jayanti (Harvard Kennedy School)

REVIEWERS

Nicolas Christin (Carnegie Mellon)
Susan Ritchie (US-India Strategic Partnership Forum)
Daniel Zhang (Stanford Institute for Human-Centered Artificial Intelligence)

FACULTY PRIMARY INVESTIGATOR

Ash Carter, TAPP Faculty Director, *in memoriam*

Statements and views expressed in this publication are solely those of the authors and do not imply endorsement by the Reviewers and their respective organizations, Harvard University, Harvard Kennedy School, or the Belfer Center for Science and International Affairs.

The Artificial Intelligence (AI) Technology Fact Sheet was originally published in January 2020. The Machine Learning (ML) Technology Fact Sheet was originally published in June 2019. The AI and ML Technology Fact Sheets were updated and combined into this document, “Artificial Intelligence & Machine Learning.”

CONTRIBUTORS OF THE ORIGINAL AI TECHNOLOGY FACTSHEET

Enrico Santus

Nicolas Christin

Harshini Jayaram

CONTRIBUTORS OF THE ORIGINAL ML TECHNOLOGY FACTSHEET

Amy Robinson

Ariel Herbert-Voss

ORIGINAL EDITORS

Amritha Jayanti

Bogdan Belei

The Technology Primers for Policymakers Series was designed to provide a brief overview of each technology and related policy considerations. These papers are not meant to be exhaustive.

Technology and Public Purpose Project
Belfer Center for Science and International Affairs
Harvard Kennedy School
79 John F. Kennedy Street, Cambridge, MA 02138

www.belfercenter.org/TAPP

Printed in the United States of America

Contents

Executive Summary

Artificial Intelligence (AI), can be defined as the theory and application of machines—especially computer programs—to perform tasks that typically require human intelligence, such as image captioning and generation, speech recognition and synthesis, natural language understanding and production, tool assembly and utilization, as well as various other perception-action based engagements. AI, in its current technological state, is being applied in various industries and domains, such as online advertising, financial trading, healthcare, pharmaceutical, and robotics. The lucrative market opportunities offered by AI applications have attracted investments from tech giants like Alphabet, Apple, Meta, Amazon, and Microsoft, as well as research universities and startups.

Machine Learning (ML), commonly categorized as a subfield of AI, is a field of study concerning the automatic discovery of historical patterns in data using statistical algorithms. ML's driving principle is that historical patterns are likely to reappear in the future. The discovered historical patterns can therefore be leveraged to make accurate predictions on data that has not been seen before. Once an algorithm is trained, it can be applied to new, larger streams of data. ML is already an integral component of many deployed commercial applications, such as content generation (e.g., text, image, audio, video generation), virtual assistants, social media feed ranking, content recommendation systems, financial market prediction, and healthcare screening and diagnostic tools, as well as administrative applications. In addition, ML is foundational in various other emerging technologies, such as autonomous vehicles and next-generation cybersecurity.

Currently, United States policy with regards to AI often derives from interpretations of various pre-existing legislations and legal precedents. However, with the increased awareness of AI-related risks (e.g., bias, accountability, misutilization, etc.), and the potential size of their impact, over the last decade, the number of proposed bills containing AI provisions significantly increased at both the state and federal levels (i.e., from two bills in 2012 to 131 in 2021), with 2% of them becoming law at the federal level and 20% of them becoming law at the state level.[1] Similarly, policies and regulatory frameworks are being crafted to guide the development and application of AI in other continents too, with Europe and Asia leading the process.[2] Acknowledging the potential impact of this technology on human life and societal dynamics, there is a pressing need for U.S. legislators and policymakers to remain engaged in the ethical and practical development of artificial intelligence.

[1] Morgan Lewis, "Increases in Global Artificial Intelligence Legislation Noted in AI Report," JDSupra, April 7, 2022, https://www.jdsupra.com/legalnews/increases-in-global-artificial-8328913/.

[2] Jonathan Keane, "China and Europe are Leading the Push to Regulate AI," CNBC, May 26, 2022, https://www.cnbc.com/2022/05/26/china-and-europe-are-leading-the-push-to-regulate-ai.htm.

PART 1: Technology

What Is Artificial Intelligence?

From Siri to Tesla vehicles, artificial intelligence is becoming increasingly prominent in the day-to-day lives of most people. Though there is currently no single universally accepted definition of AI, individual institutions and organizations have often provided their own definitions of the term to scope discussions and research initiatives. As previously stated, AI most often refers to the theory and application of machines to perform tasks that normally require human intelligence, either from a reasoning or a motion perspective.

The term *artificial intelligence* was coined at a conference at Dartmouth[3] in 1956 and has continued to gain public attention ever since due to the increased integration of AI systems into consumer-based technologies, government operations, and more.

AI is far from the all-capable human-intelligent C-3P0s of *Star Wars* or killer robots of *The Terminator*. Today's AI applications are ***narrow***, meaning they are designed to carry out one task, whether that be stock trading, playing chess, or responding to consumer complaints. ***Artificial general intelligence (AGI)*** is the scope of AI intelligence presented in movies listed above; AGI characterizes AI applications that can move between domains (and possibly modalities, such as language and vision) and ultimately apply their intelligence more broadly. While the distance between narrow AI and general AI is shrinking thanks to stronger ML models (see Deep Learning below), real AGI is still not foreseeable in the near future.

What Has Contributed to AI's Increasing Prominence?

Part of the reason why AI is becoming more prominent in our society is due to the heightened sophistication of algorithms and expanding computational capacity. Recent advancements in technical capabilities (e.g., computing power, graphics processing unit) have allowed for more use cases throughout industry and society. This increase in capability can be attributed to:

1. **Faster and more capable computer hardware**, which enables the processing of large datasets, to perform complex operations (e.g., convolutional neural networks). For example, AI systems can now examine thousands of medical records in just seconds to determine which symptoms indicate the presence of pathologies such as cancer.

[3] Rockwell Anyoha, "The History of Artificial Intelligence," Science in the News (blog), Harvard University, August 2018, 2017, http://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/.

2. **The Internet with both its enormous amount of data** (which is becoming a precious resource for training, testing, and applying AI) **and the recent possibility of sharing computational resources** (i.e., cloud computing). This is optimizing the resource allocation, allowing for faster scaling, and drastically reducing computational costs.

3. **Sensors**, which enable the monitoring of various types of parameters in real time. For example, wearables, such as smart watches, can collect vital signals from any location, helping monitor patient health and predict potential risks. Cars are filled with sensors that can proactively alert drivers about maintenance requirements.

4. **The Internet of Things (IoT)**, which thanks to recent networking advancements (e.g., 5G), allows data to be collected from sensors, processed in the cloud and deployed for multiple applications in real time. For instance, cars may send position signals to each other in low visibility contexts to avoid collisions.

5. **Advanced learning algorithms and architectures**, which enable the development of accurate models that better understand historical data and identify hidden patterns that are likely to reappear.

The Mechanics of AI

This section provides a high-level taxonomy of the mechanics behind AI. While policymakers can benefit from some knowledge of how the following tools work, to paraphrase Sendhil Mullainathan, the mechanics of AI is only a small part of the equation, just as knowing how an engine works is a small part of understanding how to drive.[4] Consequently, while this section provides a brief explanation of how AI systems work and the common terminology in the field, readers that are primarily interested in AI product applications may skip this section.

AI systems can be classified according to Approaches and Models:

Approaches: There are a few different technical approaches to the implementation of AI systems. Below, we describe Symbolic Reasoning, Classic Machine Learning, and Deep Learning. These different approaches are useful for policymakers to understand because they each provide different levels of visibility and transparency into an algorithm's decision.

[4] Sendhil Mullainathan, "ARTIFICIAL INTELLIGENCE (32200)," Sendhil Mullainathan, accessed February 9, 2023, https://sendhil.org/courses/artificial-intelligence-32200/.

Models: ML can either be *discriminative or generative*. Discriminative models (also called *classifiers*) classify new data (e.g., whether a picture represents a horse or a dog, or whether a patient has cancer or not). Generative models create new data points belonging to these classes (e.g., they can generate images of horses and dogs, or they can generate captions from medical images).

A DEEPER DIVE ON GENERATIVE MODELS

Generative AI models are designed to create new content such as text, images, audio, and video. Consumer applications powered by generative AI models such as OpenAI's chatbot, ChatGPT, and its image generator, DALL-E, have recently captured the public's attention.

Generative models use machine learning algorithms to learn patterns and structures in existing data, and then use that knowledge to create new data that is similar in style and content. These models are trained and fine-tuned based on human feedback, so the more they are used, the better they get. There are several types of generative models including variational autoencoders, generative adversarial networks, autoregressive models, and more.

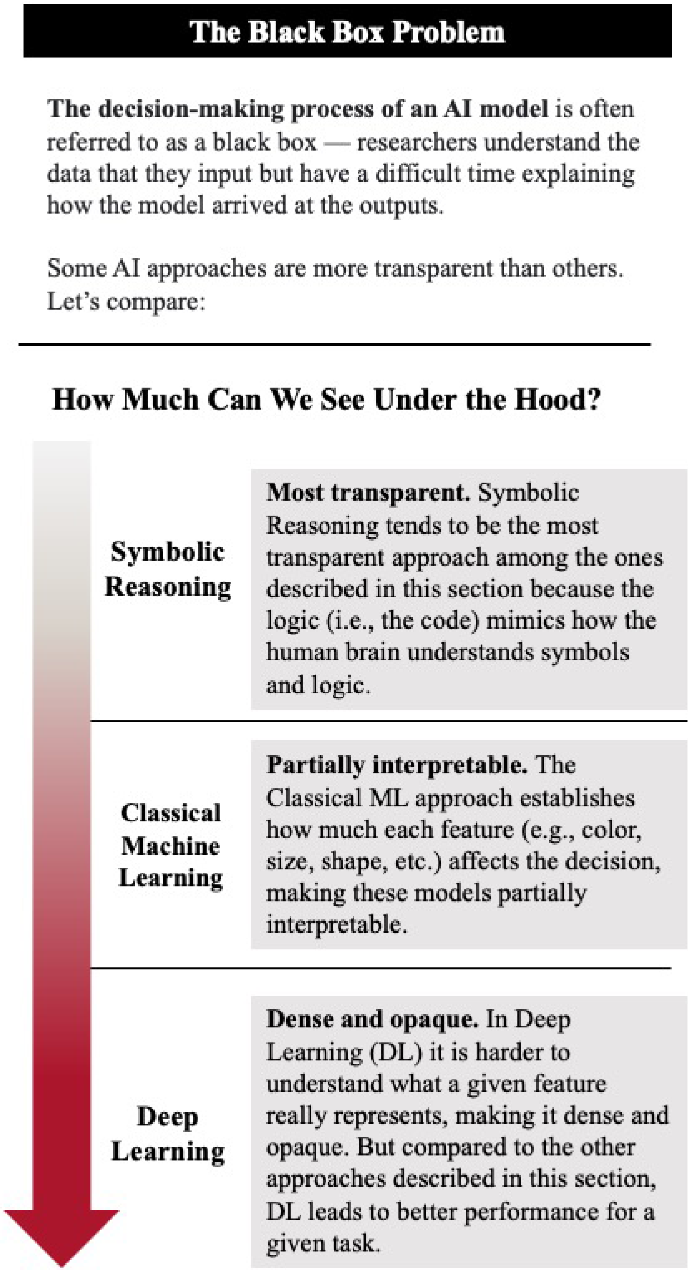
Public purpose considerations associated with AI-tools based on generative models such as the impact on labor and intellectual property considerations are discussed in Part 4: Public Purpose Considerations. The impact of generative AI on society will depend on how it is developed, implemented, and regulated.

Approaches to Implementing AI

Symbolic Reasoning (or Symbolic AI/Classical AI) is a branch of AI research that focuses on explicitly representing human knowledge in a declarative form through axioms and rules. Similarity is generally calculated as distance in knowledge graphs (e.g., "dog" is closer to "animal" than to "plant"). Symbolic AI became less popular after the late 1980s, when ML techniques began to become more prominent.

Classic Machine Learning (Classic ML) is a subfield of AI that leverages statistical methods or numerical optimization techniques to identify patterns in the input data and associate them with an expected output. In Classic ML the input data needs to be represented through numerical discrete features (e.g., color, size, shape, etc.). The ML learns the association between those features and expected outputs.

Deep Learning (DL) is a subfield of ML that further abstracts the representations by leveraging artificial neural networks, a type of computational architecture inspired by biological neural networks in the human brain. While Classic ML mostly relies on discrete features to describe the input data (color, size, shape, etc.), Deep Learning (DL) can automatically generate neural representations from the raw input data. The adjective "deep" in DL refers to the fact that DL algorithms consist of multiple communicating layers of neural networks, each representing the input data at different levels of abstraction (e.g., in a DL algorithm trained to recognize images, the first layer could represent colors, the second layer could represent edges, and the last layer could represent shapes).



The Black Box Problem

The decision-making process of an AI model is often referred to as a black box — researchers understand the data that they input but have a difficult time explaining how the model arrived at the outputs.

Some AI approaches are more transparent than others. Let's compare:

How Much Can We See Under the Hood?

Symbolic Reasoning	**Most transparent.** Symbolic Reasoning tends to be the most transparent approach among the ones described in this section because the logic (i.e., the code) mimics how the human brain understands symbols and logic.
Classical Machine Learning	**Partially interpretable.** The Classical ML approach establishes how much each feature (e.g., color, size, shape, etc.) affects the decision, making these models partially interpretable.
Deep Learning	**Dense and opaque.** In Deep Learning (DL) it is harder to understand what a given feature really represents, making it dense and opaque. But compared to the other approaches described in this section, DL leads to better performance for a given task.

How Do Machines Learn?

Learning paradigms: Learning is classified as *unsupervised* (i.e., automatically observing similarities between data points), *supervised* (i.e., through human annotated data), *weakly-supervised* (i.e., through semi-automatically made annotations), or *reinforcement* (i.e., using a "reward function" to provide feedback to systems after attempts of completing a given objective). Other derived paradigms include *adversarial learning* (i.e., two models learn from challenging each other).[5]

NOTE: While it is helpful to understand the various learning paradigms, these paradigms are not necessarily actionable for policymakers. We seek to reinforce to policymakers that since AI is currently built on algorithms that learn from data, biased datasets or human bias built into the data may create biased AI systems.

- ***Supervised learning***: The algorithm learns the association between patterns in the input data and the expected "correct" answers, which are provided in a training set.[6] For instance, to teach a computer system to detect spam, a practitioner would provide the algorithm with examples of emails that are manually labeled as spam or not spam—developers quite literally "supervise" the way an algorithm learns.

- ***Unsupervised learning***: The algorithm looks for common patterns in the data, with the goal of understanding how similar entities are to each other. An example would be feeding an algorithm with users' social media posts to discover which users discuss similar topics. Approaches based on unsupervised learning are preferred because the algorithm does not need to be trained on expensive annotations. However, as of today, supervised approaches generally outperform unsupervised ones, particularly when the user is looking for a specific outcome instead of general insights.[7]

- ***Weakly-supervised learning***: Although supervised methods generally outperform unsupervised ones, they are also the most expensive because labeling needs to be done by humans. Weakly-supervised methods are used to reduce the cost of labeling data by mixing supervised and unsupervised techniques. Weak supervision is also often used to increase the broader application and robustness of already trained algorithms and to mitigate biases.

[5] For visual learners, the authors recommend the diagrams on page 6 of J. D. Dulny et. al, The Artificial Intelligence Primer (McLean, VA: Booz Allen Hamilton Inc., 2018), https://www.boozallen.com/s/insight/thought-leadership/the-artificial-intelligence-primer.html. This report proϵ vides a simple and accessible diagram of supervised and unsupervised learning.

[6] Ibid, p. 6.

[7] "Supervised vs. Unsupervised Learning," accessed on February 4, 2023, https://www.alteryx.com/glossary/supervised-vs-unsupervised-learning.

- ***Reinforcement learning***: This kind of ML algorithm learns by attempting to solve a task in a trial-and-error fashion and receiving rewards when the proposed solution succeeds. This approach is promising in *constrained environments*, such as chess. For example, reinforcement learning gained fame when DeepMind's AI-powered software, AlphaGo Lee, beat the world champion at the game of Go. However, reinforcement learning is still far from satisfying performance expectations on unconstrained, open space problems.[8]

- ***Adversarial learning***: The above-mentioned *discriminative* and *generative* models can be combined in an adversarial training, where the generative model learns how to generate convincing counterfeit data points and the discriminative model learns how to recognize them from real ones. The models will optimize their respective abilities thanks to one another. An example of adversarial learning is DALL-E by OpenAI, which generates images from text.

What Can AI Do?

AI has been applied to sensorial (e.g., audio, video, and touch), cognitive (e.g., language and emotions), and physical (e.g., movement, action, events) domains, including overlaps across those domains:

1. **Image and video processings**: The field of AI applied to any task that is relevant for image and video processing and understanding is called **Computer Vision**. **Examples**: Classification (i.e., identify what entity or entities are represented), captioning (i.e., describing the content in natural language), processing (i.e., improving resolution, adding filters, etc.), generation (i.e., generating images from captions or other triggers, including other images).

2. **Human speech or writing**: The field of AI focused on assisting computer systems in understanding and utilizing human speech or writing is called **Natural Language Processing (NLP)**. Due to the pervasiveness of human language in any knowledge domain (including image processing), NLP has a crucial role in multiple fields such as medicine, finance, journalism, marketing, etc. **Examples**: NLP algorithms can classify documents, find relevant information, extract information from text and structure it in databases and knowledge graphs, answer questions, detect fake news, generate texts from tables, generate image and video captioning, turn texts to speech or speech to text, etc.

[8] Robert D. Hof, "Deep Learning," MIT Technology Review, April 23, 2013, https://www.technologyreview.com/technology/deep-learning/; Bernard Golstein, "A Brief Taxonomy of AI," SharperAI, November 1, 2018, https://www.sharper.ai/taxonomy-ai/.

3. **Processing and generation of sounds:** The field of AI focused on the understanding, processing, and generation of sounds, whether they are speech-related, noises, or music, is called **Audio Signal Processing (ASP)**. **Examples**: In the last few years, the quality of speech processors and generators has improved to the point that it is now common to interact with conversational agents like Siri and Alexa. ASP has also been applied to the screening of diseases, such as COVID-19, from cough recordings, as well to the automatic generation of music.

4. **Sensors and signal detection**: This field focuses on the processing of data from sensors in machines (e.g., car sensors about temperature or wheel pressure) or humans (e.g., wearables or medical devices) and make useful predictions about the status of the entity, including health-related risks. This field is currently growing thanks to the continuous diffusion of sensors and IoT devices, and thanks to improved networking infrastructures which allow multiple devices to interact with each other in real time.

5. **Robotics**: This field supports the development and life of robots, with all their functions (i.e., perception, motion, interaction with the environment, interaction with humans, etc.). Robots can be imagined as agents in an environment that need to be perceived and analyzed to take some possible actions (e.g., an automatic mop must identify the edges and obstacles and move around them, without missing parts of the floor). Decisions made by robots can be based on models trained by supervised, weakly-supervised, or reinforcement learning methods, for example. **Examples**: AI can support the decisions of self-driving cars (e.g., when to accelerate, brake, turn) in all their levels of automation (e.g., assistance, partial/conditional/full automation). While AI algorithms may be applied to robots, robotics and AI are not necessarily dependent technologies.

KEY INSIGHTS

- Artificial intelligence can be categorized according to various criteria, including the scope of intelligence (narrow vs. general), the approach (Symbolic Reasoning, Classic ML, and DL), and the learning paradigm.

- Artificial intelligence can be applied to any *sensorial* (e.g., audio, video, and touch), *cognitive* (e.g., language and emotions), and *physical* (e.g., movement, action, and events) domain, as well as to several domains at the same time.

- Because the domain space and specific applications of artificial intelligence differ in possible impacts to society, specific regulations should be considered.

PART 2: Product Applications and Market Development

Artificial intelligence has product applications in both physical and digital spaces. Examples of product applications with a physical presence are autonomous vehicles, automated manufacturing, precision farming, and robotics. Product applications that are primarily in the digital realm are virtual assistants, such as Siri and Alexa, recommendation algorithms, such as those used by Amazon or TikTok, and customer service chatbots.

Most technology companies and many defense agencies are investing in artificial intelligence. Over the years, much of the funding for AI research flowed from the Defense Advanced Research Projects Agency (DARPA). For example, Siri is a consumer application that began as a Defense Department research effort.[9]

The healthcare industry is also investing in AI to assist in cancer detection, precision medicine, and health insurance and the pharmaceutical industry is investing in AI to accelerate drug discovery and protein folding.

In recent years, some promising applications of AI include:

- **Protein folding**: Proteins have unique 3D dimensions that are difficult to predict. Using AI, researchers can predict more accurately the shape of proteins. Accurately identifying the shape of proteins from their amino-acid sequence would vastly accelerate efforts to identify the building blocks of cells and enable quicker and more advanced drug discoveries.[10]

- **Nuclear fusion**: Nuclear fusion,[11] or the process of creating roiling plasma that is hotter than the surface of the sun, has been mooted as the potential clean energy source of the future. AI can be used to learn what configuration and shapes of the plasma yields more power. This process requires a large amount of design, calculation, and engineering work that can be achieved by DL algorithms.

[9] Geoffrey Ingersoll, "That's Right, Apple's Famous Siri Began in the Military," *Business Insider*, October 4, 2012, https://www.businessinsider.com/ thats-right-folks-apples-siri-is-totally-a-military-brat-2012-10.

[10] Ewen Callaway, "'It Will Change Everything': DeepMind's AI Makes Gigantic Leap in Solving Protein Structures," *Nature Magazine*, November 30, 2020, https://www.nature.com/articles/d41586-020-03348-4. .

[11] Amit Katwala, "DeepMind Has Trained an AI to Control Nuclear Fusion," *Wired*, February 16, 2022, https://www.wired.com/story/deepmind-ai-nuclear-fusion/.

- **Energy efficiency**: AI can help increase energy efficiency and accelerate the energy transition towards renewable energy sources. Specifically, AI is used to coordinate turbines in entire wind farms to maximize their output and make them act as one turbine instead of isolated units. AI is also used to accurately predict peaks of energy demand and automatically increase or decrease power generation when needed. Furthermore, AI is being used to predict weather to manage the energy mix efficiently (e.g., wind vs. solar).

For productized AI solutions, some of the top players are:[12] Google (DeepMind), Amazon, Microsoft and OpenAI, Salesforce, Baidu, Apple, Intel, Meta (formerly Facebook), Nvidia, Tencent, and IBM. Many of these companies have developed AI Principles that guide their work and engagement with the space.

Many companies investing in AI are growing to be among the dominant players. Some of the front-runners noted most frequently are: Google DeepMind, CrowdStrike, AIBrain, CloudMinds, SenseTime, and Twitter. It is important to note that many leaders in the space, including fourteen startups valued at $1 billion and above, are Chinese companies.[13] Over time, AI will likely become embedded in all computer applications, at least to some extent.

Additionally, several universities, nonprofits, and research organizations are pioneering technical progress for AI as well as catalyzing important conversations around what the implications of AI are for humanity. Some of these include:

- OpenAI, a San Francisco-based company, focuses on building safe AGI.[14] OpenAI has recently captured the public's attention with the development of its AI-tools based on generative models.

- The Partnership on AI pulls together over eighty partners ranging from the private sector to civil society representatives to shape the dialogue and practices around artificial intelligence.[15]

- The Future of Life Institute focuses on ways to create positive AI applications while working heavily on risk mitigation.[16]

[12] John Divine,"Artificial Intelligence Stocks: 10 of the Best AI Stocks to Buy," *U.S. News & World Report*, January 10, 2020. https://money.usnews.com/investing/stock-market-news/slideshows/artificial-intelligence-stocks-the-10-best-ai-companies.

[13] WEF Digital Media Team, "Meet China's Five Biggest AI Companies," *CommonWealth Magazine*, September 21, 2018, https://english.cw.com.tw/article/article. action?id=2122.

[14] "OpenAI," accessed on February 4, 2023, https://openai.com.

[15] "Partnership on AI," accessed on February 4, 2023, https://www.partnershiponai.org/.

[16] "Future of Life Institute," accessed on February 4, 2023, https://futureoflife.org/?cn-reloaded=1.

- The Future of Humanity Institute at the University of Oxford[17] and Leverhulme Centre for the Future of Intelligence[18] look at big-picture questions regarding AI and technology-specific risks.

- Almost all universities in the U.S. are heavily engaged with artificial intelligence research. Carnegie Mellon University (CMU), Stanford University, the Massachusetts Institute of Technology (MIT), and the University of California at Berkeley are generally considered among the most prolific.[19]

KEY INSIGHTS

- Artificial Intelligence has product applications in both the physical and digital spaces with applications ranging from autonomous vehicles to chatbots. The wide range of possible applications makes it a particularly transformative and powerful set of technologies.

- Investment and interest in AI have grown significantly over the last few years with many large technology companies such as Google and Microsoft investing in AI, along with more specialized AI companies such as DeepMind and AI Brain. Research organizations and universities have also been particularly interested in AI with Berkeley, CMU, MIT, and Stanford considered among the leading AI research institutions.

[17] "Future of Humanity Institute," accessed on February 4, 2023, https://www.fhi.ox.ac.uk/

[18] "Leverhulme Centre for the Future of Intelligence," accessed on February 4, 2023, http://lcfi.ac.uk/.

[19] Monica Nickelsburg,"Top Schools for AI: New Study Ranks the Leading U.S. Artificial Intelligence Grad Programs," *GeekWire*, March 20, 2018, https://www.geekwire. com/2018/top-schools-ai-new-study-ranks-leading-u-s-artificial-intelligence-grad-programs/.

PART 3: Current State of Strategies and Regulations

As of 2021, globally there were over 250 national strategies, agendas, and plans involving AI.[20] Much of the news and attention today focuses on AI research and development strategies rather than governance. Many people see AI development as a "race"contributing to the steady rise in national AI strategies, though many scholars are encouraging developers and policymakers to reject this narrative.[21]

The following are examples of national strategies and funding related to AI and ML:

- In the U.S., both defense and non-defense agencies are playing pivotal roles. The Department of Defense's (DoD) Joint Artificial Intelligence Center—now called the Chief Digital and Artificial Intelligence Office—is an effort to combine various AI efforts throughout the defense ecosystem. Detailed DoD and DARPA budgets are not publicly available, but the DoD requested over $870 million for AI projects in the fiscal year 2022. In non-defense areas, total budgets for AI have grown from $560 million in 2018 to over $1.6 billion in 2022 (requested budget). Of the 2022 budget request for AI initiatives, more than 41 percent (over $690 million) was from the National Science Foundation and more than 21 percent (over $350 million) was from the National Institutes of Health.[22] Examples of major initiatives that have been funded by these defense and non-defense requests include the DoD's Artificial Intelligence Strategy, Executive Order13859 on Maintaining American Leadership in Artificial Intelligence, the National AI Initiative Act, the National AI R&D Strategic Plan, the National Security Commission on AI, the National AI Research Resource, and the National Strategy for Critical and Emerging Technologies.[23] More recently developed national frameworks include the AI Risk Management Framework released by NIST and the Blueprint for an AI Bill of Rights developed by OSTP.[24]

[20] National AI Policies & Strategies, OECD.AI Policy Observatory database, European Commission and Better Policies for Better Lives, accessed on June 6, 2022, , https://oecd.ai/en/dashboards.

[21] Julian E. Barnes and Josh Chin, "The New Arms Race in AI," *Wall Street Journal*, March 2, 2018, https://www.wsj.com/articles/the-new-arms-race-in-ai-1520009261.

[22] "Artificial Intelligence R&D Investments: Fiscal Year 2018 - Fiscal Year 2022," Networking and Information Technology Research and Development, https://www.nitrd.gov/apps/itdashboard/ai-rd-investments/.

[23] Policies for United States, OECD.AI Policy Observatory database, European Commission and Better Policies for Better Lives, accessed on June 6, 2022, https://oecd.ai/en/dashboards/policy-initiatives?conceptUris=http:%2F%2Fkim.oecd.org%2FTaxonomy%2FGeographicalAreas%23UnitedStates.

[24] "Blueprint for an AI Bill of Rights," White House Office of Science and Technology Policy, October 2022, https://www.whitehouse.gov/ostp/ai-bill-of-rights/.

- The European Union built on its April 2018 Strategy for AI by developing a Coordinated Plan to run until 2027. The plan lays out around seventy coordinated actions between Member States and the European Commission in areas including research, investment, market uptake, skills and talent, data, and international cooperation. One of the plan's ambitious targets is to attract over 20 billion euros per year of public and private investment into AI.[25]

- China is aiming to be the world's primary AI innovation center by 2030 and has launched numerous state-funded R&D initiatives to this end. Estimated 2018 government expenditures on AI range from $300 million to $2.7 billion for military R&D, and a lower bound[26] estimate of $1.7 to $5.7 billion for civilian R&D.[27] In addition to the Chinese central government's initiatives, local governments (provinces and municipalities) are investing substantially and setting up other incentives to attract AI talent and businesses.[28]

- In Russia, planned state funding for AI projects from 2020 to 2024 is over $4.5 billion.[29] The state-funded Sberbank is leading key elements of the country's AI strategy and associated funding. In 2019, Sberbank led the development of a national technological development roadmap for AI, which also specified investment targets. In 2019, Russia also finalized its Strategy for the Development of AI through 2030, although funding cuts due to COVID-19 have impacted the roll out of the strategy.

- Many other countries (such as India, Japan, UK, France, Germany, Canada) have also accelerated the development, funding, and implementation of national strategies for AI:

    - Canada was a leader in developing guidelines for the responsible use of AI,[30] and in ensuring that multilateral groups like the G7 and Organisation for Economic Co-operation and Development (OECD) incorporated AI into their priority areas.

[25] European Commission, "White Paper: On Artificial Intelligence: A European Approach to Excellence and Trust," February 19, 2020, https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.

[26] Why this is a lower bound estimate. Based on China's National Natural Science Foundation and National Key R&D Programs, the estimate excludes spending through state-funded venture capital funds, "megaprojects" to accelerate innovation in a few key industries, and the "bases and talents" program for centrally supported labs and research teams.

[27] Ashwin Acharya and Zachary Arnold, *Chinese Public AI R&D Spending: Provisional Findings* (Washington, DC: Center for Security and Emerging Technology, December 2019), https://cset.georgetown.edu/publication/chinese-public-ai-rd-spending-provisional-findings/.

[28] Jaqueline Ives and Anna Holzmman, "Local Governments Power Up to Advance China's National AI Agenda," Mercator Institute for China Studies, April 26, 2018, https://www.merics.org/en/analysis/local-governments-power-advance-chinas-national-ai-agenda.

[29] Nikolai Markotkin and Elena Chernenko, "Developing Artificial Intelligence in Russia: Objectives and Reality," Carnegie Endowment for International Peace, May 8, 2020, https://carnegiemoscow.org/commentary/82422.

[30] Responsible Use of Artificial Intelligence, accessed on February 3, 2023, Government of Canada, https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai.html.

- India is developing its governance and operational principles for AI.[31] These are informed by other frameworks (e.g., in the U.S., EU, and Australia) and are being customized to India's context.

A challenge around AI is that the field itself is very broad and extremely intersectional, so trying to regulate it through a single governing body is challenging. Instead, relevant frameworks may depend on its specific application. Some examples of how jurisdiction is already dispersed in the United States is that autonomous vehicles are subject to the Department of Transportation, the Transportation Safety Administration, and similar agencies, while health diagnoses applications are subject to the Health Insurance Portability and Accountability Act under the Department of Health and Human Services.

Below are some examples of countries that have introduced AI-related regulations and governance initiatives. *NOTE: The following list is not exhaustive.*

U.S. Existing Laws Applied to AI

- **Product liability and tort laws**: Several judicial cases already have applied product liability and tort laws to cases that involve injury resulting from artificial intelligence applications such as GPS and smart robotics. For example, *Cruz v. Talmadge, Calvary Coach; Nilsson v. General Motors LLC; and Holbrook v. Prodomax Automation Ltd.*[32]

U.S. Regulation/Strategies Specifically Created for AI and AI Applications

- **Algorithmic Accountability Act and American Data Privacy and Protection Act**: These two proposed bills are currently the closest to constituting federal AI regulation. The Algorithmic Accountability Act would require the Federal Trade Commission to house a public repository of impact assessments of AI in terms of bias, effectiveness, etc.[33] The Federal Trade Commission would also be given the resources to enforce the law. The American Data Privacy and Protection Act, which incorporates many key clauses from the Algorithmic Accountability Act, is a draft bipartisan bill to introduce a national data privacy and data security framework.[34]

[31] *Responsible AI #AIForAll: Approach Document for India: Part 2—Operationalizing Principles for Responsible AI* (New Delhi: NITI Aayog 2021), https://www.niti.gov.in/sites/default/files/2021-08/Part2-Responsible-AI-12082021.pdf.

[32] John Villasenor, "Products Liability Law as a Way to Address AI Harms," Brookings Institution, October 31, 2019, https://www.brookings.edu/research/ products-liability-law-as-a-way-to-address-ai-harms/.

[33] Ron Wyden, "Wyden, Booker and Clarke Introduce Algorithmic Accountability Act of 2022 to Require New Transparency and Accountability for Automated Decision Systems," press release, February 3, 2022, https://www.wyden.senate.gov/news/press-releases/wyden-booker-and-clarke-introduce-algorithmic-accountability-act-of-2022-to-require-new-transparency-and-accountability-for-automated-decision-systems.

[34] Frank Pallone, Jr., "House and Senate Leaders Release Bipartisan Discussion Draft of Comprehensive Data Privacy Bill," House Committee on Energy and Commerce, press release, June 3, 2022, https://energycommerce.house.gov/newsroom/press-releases/house-and-senate-leaders-release-bipartisan-discussion-draft-of.

- **AI Risk Management Framework**: The National Institute of Standards and Technology (NIST) is developing this risk management framework, with extensive public comments and inputs. NIST is aiming to release a Version 1.0 of the framework in early 2023.[35]

- **William M. (Mac) Thornberry National Defense Authorization Act for Fiscal Year 2021**. This act included numerous consequential provisions related to AI. Examples include provisions on AI ethics for DoD, risk management produced by the National Institute of Standards and Technology (NIST), and a national initiative led by the White House.[36]

- **Defense Innovation Board AI Principles**: Starting in 2019, the DoD tasked the Defense Innovation Board with proposing AI Ethics Principles for the design, development, and deployment of AI for combat and non-combat purposes.[37]

- **American AI Initiative**: Launched in 2019, the executive order supports funneling federal funding and resources toward AI-specific research while also implementing U.S.-led international AI standards. Additionally, the initiative calls for new research into increasing AI literacy in U.S. workers.[38]

- **National Artificial Intelligence Research and Development Strategic Plan**: This plan supports the American AI Initiative by identifying priority areas for federal investments into R&D related to AI. The plan was first developed in 2016 and was updated in 2019; the update lists eight strategic priorities ranging from investments to standards, workforce requirements to public-private partnerships.[39]

- **Local and state regulations on facial recognition technologies**: Since 2019, regulations at the local and state levels have been enacted to provide government oversight of or to narrowly ban facial recognition technology.[40] However, there are no regulations at the federal level, despite increased use of facial recognition technology by numerous federal agencies.[41]

[35] "AI Risk Management Framework Concept Paper," National Institute of Standards and Technology, December 13, 2021, https://www.nist.gov/system/files/documents/2021/12/14/AI%20RMF%20Concept%20Paper_13Dec2021_posted.pdf.

[36] "Summary of AI Provisions from the National Defense Authorization Act 2021," accessed on February 4, 2023, Stanford University Human-Centered Artificial Intelligence, https://hai.stanford.edu/policy/policy-resources/summary-ai-provisions-national-defense-authorization-act-2021.

[37] Policies for United States, OECD.AI Policy Observatory database.

[38] Mark Minevich, "The American AI Initiative: A Good First Step, of Many," *TechCrunch*, August 20, 2019, https://techcrunch.com/2019/08/20/the-american-ai-initiative-a-good-first-step-of-many/.

[39] "The National Artificial Intelligence Research and Development Strategic Plan: 2019 Update," Select Committee on Artificial Intelligence of the National Science & Technology Council, June 2019, https://www.nitrd.gov/pubs/National-AI-RD-Strategy-2019.pdf.

[40] Ibid.; Taylor K. Lively, "Facial Recognition in the United States: Privacy Concerns and Legal Developments," December 1, 2021, https://www.asisonline.org/security-management-magazine/monthly-issues/security-technology/archive/2021/december/facial-recognition-in-the-us-privacy-concerns-and-legal-developments/.

[41] Tate Ryan-Mosley, "US Government Agencies Plan to Increase Their Use of Facial Recognition Technology," *MIT Technology Revie*w, August 24, 2021, https://www.technologyreview.com/2021/08/24/1032967/us-government-agencies-plan-to-increase-their-use-of-facial-recognition-technology/.

- **AI Next Campaign**: Started in 2018, this effort builds on DARPA's five decades of AI technology creation. Its objectives are to invest in high-risk, high-payoff research to tackle the most difficult national security challenges.[42]

- **Automated Vehicles 3.0**: Started in 2017, this initiative introduces guiding principles and describes the Department of Transportation's strategy to address existing barriers to safety innovation and progress.[43]

- **Preparing for the Future of Transportation**: Published by the National Highway and Transportation Safety Administration in 2016, this document provides an outline for the evolution of driving from no automation to full automation, and the roles of humans and vehicles at each stage.[44]

- **Preparing for the Future of Artificial Intelligence**: Created in 2016 by the Executive Office of the President National Science and Technology Council Committee on Technology, this report evaluates the state of AI, the role of agencies, and more.[45]

International Regulations/Strategies

In the EU:

- **Digital Services Act and Digital Markets Act**: This package of legislation was passed by the European Parliament in July 2022 and must now be approved by the Council of the EU. This legislation includes powers to require large online platforms to explain their algorithms to guard against anti-competitive practices, biases, illegal transactions, misinformation, etc.[46]

- **Data Governance Act**: This regulation, which entered into force in June 2022, includes measures to regulate commercial data sharing by businesses (both business-to-business and business-to-consumer-to-business) and institute governance mechanisms for data standardization.[47]

[42] Policies for United States, OECD.AI Policy Observatory database.

[43] Policies for United States, OECD.AI Policy Observatory database.

[44] "Preparing for the Future of Transportation," U.S. Department of Transportation, October 2018, https://www.transportation.gov/sites/dot.gov/files/docs/policy-initiatives/ automated-vehicles/320711/preparing-future-transportation-automated-vehicle-30.pdf.

[45] "Preparing for the Future of Artificial Intelligence," Executive Office of the President National Science and Technology Council Committee on Technology, October 2016, https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf.

[46] "The Digital Services Act Package," accessed on February 4, 2023, European Commission, https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package.

[47] Policies for European Union, OECD.AIPolicy Observatory database, European Commission and Better Policies for Better Lives, accessed on June 6, 2022, https://oecd.ai/en/dashboards/policy-initiatives?conceptUris=http:%2F%2Fkim.oecd.org%2FTaxonomy%2FOrganisations%23EuropeanUnion.

- **The Artificial Intelligence Act**: Published by the European Commission in 2021, this legislative package includes: i) a proposed European regulatory approach for AI; ii) an updated and Coordinated Plan on the development and use of AI across EU Member States; and iii) proposed regulations on machinery. This draft legislation categorizes uses of AI into minimal risk, low risk, high risk, and unacceptable risk. The unacceptable risk category is proposed to be strictly banned, while the minimal risk category is proposed to be used freely. Appropriate regulation is proposed for each of the other two categories. The proposal also encourages European countries to facilitate AI R&D under strict regulatory oversight.[48]

- **Framework of Ethical Aspects of AI, Robotics, and Related Technologies**: This 2020 resolution by the European Parliament includes provisions to supervise R&D development and implementation technology that involves ML.[49]

- **General Data Protection Regulation**: While this 2018 regulation covers many aspects of privacy, it includes a specific clause providing specific rights to individuals impacted by decision-making driven by AI. Further, the implications of data storage, ownership, and rights have an impact on how companies can leverage data for their AI initiatives.[50]

In China:

- **Internet Information Service Algorithmic Recommendation Management Provisions**: Introduced in March 2022 by the powerful regulator Cyberspace Administration of China,[51] this is among the most concrete regulations worldwide to govern internet recommendation algorithms. Regulators elsewhere in the world might obtain useful insights from the implementation of these laws and from companies' approaches for compliance.[52]

- **Guidelines for Building New Generation AI Standard System**: This initiative aims to: i) establish an AI Ethical Standard Code by 2021; and ii) publish and test the implementation of standards in key areas (e.g., algorithms, systems, services) by 2023.[53]

[48] Policies for European Union, OECD.AI Policy Observatory database.

[49] Policies for European Union, OECD.AI Policy Observatory database.

[50] Mike Kaput, "How the European Union's GDPR Rules Impact Artificial Intelligence and Machine Learning," Marketing Artificial Intelligence Institute, May 24, 2018, https:// www.marketingaiinstitute.com/blog/how-the-european-unions-gdpr-rules-impact-artificial-intelligence-and-machine-learning.

[51] Matt Sheehan, "China's New AI Governance Initiatives Shouldn't Be Ignored," Carnegie Endowment for International Peace, January 4, 2022, https://carnegieendowment.org/2022/01/04/china-s-new-ai-governance-initiatives-shouldn-t-be-ignored-pub-86127.

[52] Rogier Creemer, Graham Webster, and Helen Toner, "Translation: Internet Information Service Algorithmic Recommendation Management Provisions," DigiChina, January 10, 2022, https://digichina.stanford.edu/work/translation-internet-information-service-algorithmic-recommendation-management-provisions-effective-march-1-2022/.

[53] Policies for China, OECD.AI Policy Observatory database.

- **Governance Principles for New Generation AI**: This 2019 initiative lays out eight principles to guide the development of responsible AI: harmony, friendliness, fairness, inclusiveness, respect for privacy, security and controllability, shared responsibility, open collaboration, and agile governance.[54]

- **AI Standardization White Paper**: Issued by the Ministry for Industry and Information Technology in 2018, it outlines a Chinese development plan and broader strategy for AI.[55]

- **A Next Generation AI Development Plan**: This is China's main national AI strategy document, issued in 2017 by the State Council, which is equivalent to the cabinet in the U.S. Federal Government. This strategy includes basic principles, objectives, and operational details (e.g., focus tasks, resource allocation, organizational structure).[56]

In some other countries, examples of regulation include:

- **Russia**: The Sberbank-led AI-Russia Alliance, which is experimenting with innovation-friendly technology regulation and attempting to improve public-private cooperation.[57]

- **Australia**: i) Issue paper on Automated Decision-Making and AI, which seeks public comment that will lead to a discussion paper and then an overarching Digital Age Policy Framework;[58] ii) AI Standards Roadmap, which identifies priority areas for AI standards development and outlines a strategy for Australian leadership on international standards development; iii) AI Ethics Framework; iv) National Enforcement Guidelines for Automated Vehicles; and v) Australian Code for the Responsible Conduct of Research.[59]

[54] Policies for China, OECD.AI Policy Observatory database, European Commission and Better Policies for Better Lives, accessed on June 6, 2022, https://oecd.ai/en/dashboards/policy-initiatives?conceptUris=http:%2F%2Fkim.oecd.org%2FTaxonomy%2FGeographicalAreas%23China.

[55] Yan Luo, Ashwin Kaja, and Theodore J. Karch, "China's Framework of AI Standards Moves Ahead," *National Law Review*, July 16, 2018, https://www.natlawreview.com/article/china-s-framework-ai-standards-moves-ahead.

[56] Graham Webster, Rogier Creemers, Paul Triolo, and Elsa Kania, "Full Translation: China's 'New Generation Artificial Intelligence Development Plan'," *New America*, August 1, 2017, https://www.newamerica.org/cybersecurity-initiative/digichina/blog/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017/.

[57] Nikolai Markotkin and Elena Chernenko, "Developing Artificial Intelligence in Russia: Objectives and Reality," Carnegie Endowment for International Peace, May 8, 2020, https://carnegiemoscow.org/commentary/82422.

[58] *Positioning Australia as a Leader in Digital Economy Regulation* (Canberra: Australian Government, 2022), https://www.pmc.gov.au/sites/default/files/automated-decision-making-ai-regulation-issues-paper.pdf.

[59] Policies for Australia, OECD.AI Policy Observatory database, European Commission and Better Policies for Better Lives, accessed on June 6, 2022, https://oecd.ai/en/dashboards/policy-initiatives?conceptUris=http:%2F%2Fkim.oecd.org%2FTaxonomy%2FGeographicalAreas%23Australia.

- **India**: i) AI Standardization Committee, under the Department of Telecommunications, that is tasked with developing AI standards; and ii) Operationalizing Principles for Responsible AI.[60]

- **Japan**: i) Governance Guidelines for the Implementation of AI Principles, compiled by the Ministry of Economy, Trade and Industry based on public comment;[61] ii) AI R&D Guidelines, which were prepared as a basis for discussions with the G7 and OECD; and iii) Legal Regulation of Autonomous Driving Technology.[62]

Examples of multilateral efforts include:

- **G7**: In 2018, the G7 issued the Charlevoix Common Vision for the Future of Artificial Intelligence. This statement listed twelve commitments by the leaders of the G7 countries.[63]

- **G20**: The ministers of Trade and Digital Economy of G20 countries issued a statement in 2019 that included an emphasis on human-centered artificial intelligence.[64] The statement described principles for trustworthy AI and outlined an approach for international cooperation.[65]

- **OECD**: The OECD AI Principles were adopted in 2019. They were the first major joint declaration on AI by a large number of countries, including non-OECD member countries like India. OECD's first-mover advantage in the field of AI regulations has resulted in a rich repository of information and has catapulted this organization into a leadership position on the issue.[66]

While many of these existing pieces of legislation and strategies have focused on individual sovereignties, world leaders like French President Emmanuel Macron[67] and Chinese President Xi Jinping[68] have talked about the need for international cooperation on AI regulation.

[60] Policies for India, OECD.AI Policy Observatory database, European Commission and Better Policies for Better Lives, accessed on June 6, 2022, https://oecd.ai/en/dashboards/policy-initiatives?conceptUris=http:%2F%2Fkim.oecd.org%2FTaxonomy%2FGeographicalAreas%23India.

[61] "Governance Guidelines for Implementation of AI Principles Ver. 1.1," Ministry of Economy, Trade, and Industry, January 28, 2022, https://www.meti.go.jp/english/press/2022/0128_003.html.

[62] Policies for Japan, OECD.AI Policy Observatory database, European Commission and Better Policies for Better Lives, accessed on June 6, 2022, https://oecd.ai/en/dashboards/policy-initiatives?conceptUris=http:%2F%2Fkim.oecd.org%2FTaxonomy%2FGeographicalAreas%23Japan.

[63] "Charlevoix Common Vision for the Future of Artificial Intelligence," G7 2018, Government of Canada, June 6, 2018, https://www.international.gc.ca/world-monde/international_relations-relations_internationales/g7/documents/2018-06-09-artificial-intelligence-artificielle.aspx?lang=eng.

[64] "G20 Ministerial Statement on Trade and Digital Economy," Ministry of Economy, Trade and Industry of Japan, June 2019, https://www.meti.go.jp/press/2019/06/20190610010/20190610010-1.pdf

[65] "G20 AI Principles," Annex, G20 Insights, July 2019, https://www.g20-insights.org/wp-content/uploads/2019/07/G20-Japan-AI-Principles.pdf

[66] "OECD AI Principles Overview," OECD.AI Policy Observatory, 2019,https://oecd.ai/en/ai-principles.

[67] 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/.

[68] Will Knight, "China's Leaders Are Softening Their Stance on AI," *MIT Technology Review*, September 18, 2018, https://www.technologyreview.com/s/612141/ chinas-leaders-are-calling-for-international-collaboration-on-ai/.

KEY INSIGHTS

- Over the last decade, governments in the United States and elsewhere have developed hundreds of strategies, initiatives, and regulations on AI.
- China is at the leading edge of introducing governance structures for algorithms driving AI. The EU is making quick progress too.
- **To lead the writing of international rules and norms on AI, the United States needs to prioritize the development of a comprehensive AI strategy.**

Part 4: Public Purpose Considerations

In general, AI applications are designed to improve human life. However, as with any other tool, AI can be intentionally or unintentionally misused and consequently harm individuals and society. As a result, the field of "AI ethics" has become increasingly popular.

Some noteworthy areas of concern include:

- **Degree of autonomy**: With ML, once a system learns a task, it can carry it out without human intervention. It is important to consider which tasks are appropriate to automate and which tasks should require some degree of human oversight and/or control (referred to as "human-in-the-loop").

- **Transparency and interpretability**: Concerns about transparency and interpretability involve two sets of issues:

    - **Opaque decision-making processes**, which lead to concerns about meaningful interpretability, responsibility, accountability, and feedback. For example, in military applications of AI (e.g., in the intelligence community), it is important for human end-point operators to consider why and how an AI-based analytics tool came to the recommendation(s) that it did. This is crucial for system-wide responsibility and accountability.[69] Particularly with more complex AI systems, it is hard to know why they make certain decisions. The systems are trained from extremely large data sets and often are "black boxed" in their decision-making process. For instance, in DL systems based on neural networks there are "hidden layers"of neurons between input data and output data, which make it difficult to understand how decisions were made within the algorithm. Even in AI systems not using neural networks, companies often do not release the code for proprietary reasons. This makes it difficult to understand how the system navigates highly interpretable situations.

    - **People not being transparent about when they are using an AI system**, for example, police officers not telling suspects that they were identified using facial recognition systems. This prevents the suspect from contesting the identification with full information.

- **Discrimination and bias**: Both development and usage of AI systems can result in discrimination.

[69] Sarah Marquart, "Transparent and Interpretable AI: an Interview with Percy Liang," Future of Life Institute, June 5, 2018, https://futureoflife.org/2018/02/13/ transparent-interpretable-ai/.

    - **In development**: Since AI is currently built on algorithms that learn from data, biased datasets or human bias built into the data may create biased AI systems.[70] For example, one résumé-analyzing system may consistently choose men over women.[71] Forty percent of false matches in Amazon's facial Rekognition software involved people of color, which could place them at greater risk of being considered potential criminals.[72]

    - **In usage**: AI systems like predictive policing tools may be used more often in low-income neighborhoods than in high-income neighborhoods.[73]

- **Privacy**: Because more data yields better AI, there is a strong incentive for companies to collect as much data about individuals as possible.[74] Furthermore, AI can sometimes help companies determine extremely personal information, such as someone's sexual identity, bipolar diagnosis, or likelihood to commit suicide.[75] Governments may also be able to gain unprecedented knowledge about their citizens, even without their consent, through a combination of surveillance and AI tools, such as speech and face recognition.[76]

- **Security**: Many investments in and applications of AI are associated with military and defense organizations.[77] Moreover, safety-critical AI systems like autonomous vehicles need assurances about the system's robustness to unforeseen circumstances, hacking, and malicious data or other inputs that may corrupt the AI's functionality.[78] For example, there are concerns that targeted interference could get an autonomous vehicle to read a stop sign as an increased speed limit sign or cause the vehicle to veer into incoming traffic.[79]

[70] AI and bias, IBM Research-US, accessed January 23, 2020, https://www.research.ibm.com/5-in-5/ai-and-bias/.

[71] Jeffrey Dastin, "Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women," Reuters, October 10, 2018, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN-1MK08G.

[72] Queenie Wong, "Why Facial Recognition's Racial Bias Problem is So Hard to Crack," *CNET*, March 27, 2019, https://www.cnet.com/science/why-facial-recognitions-racial-bias-problem-is-so-hard-to-crack/.

[73] Dhruv Mehrotra, Surya Mattu, Annie Gilbertson, and Aaron Sankin, "How We Determined Predictive Policing Software Disproportionately Targeted Low-Income, Black, and Latino Neighborhoods," *Gizmodo*, December 2, 2021, https://gizmodo.com/how-we-determined-predictive-policing-software-dispropo-1848139456.

[74] Darren Shou, "The Next Big Privacy Hurdle? Teaching AI to Forget," *Wired*, June 11, 2019, https://www.wired.com/story/the-next-big-privacy-hurdle-teaching-ai-to-forget/.

[75] Emerging Technology, "Your Tweets Could Show If You Need Help for Bipolar Disorder," *MIT Technology Review*, January 5, 2018, https://www.technologyreview.com/2018/01/05/146407/your-tweets-could-show-if-you-need-help-for-bipolar-disorder/; Mason Marks, "Suicide Prediction Technology Is Revolutionary. It Badly Needs Oversight," Washington Post, December 28, 2018, https://www.washingtonpost.com/ outlook/suiscide-prediction-technology-is-revolutionary-it-badly-needs-oversight/2018/12/20/214d2532-fd6b-11e8-ad40-cdfd0e0dd65a_story.html.

[76] Lily Kuo, "Chinese Surveillance Company Tracking 2.5m Xinjiang Residents," *Guardian*, February 18, 2019, https://www.theguardian.com/world/2019/feb/18/ chinese-surveillance-company-tracking-25m-xinjiang-residents.

[77] M. L. Cummings, "Artificial Intelligence and the Future of Warfare," Chatham House research paper, January 2017, https://www.chathamhouse.org/sites/default/files/publications/research/2017-01-26-artificial-intelligence-future-warfare-cummings-final.pdf.

[78] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kowk, "Fooling Neural Networks in the Physical World," labsix, October 31, 2017, https://www.labsix.org/ physical-objects-that-fool-neural-nets/; Evan Ackerman, "Slight Street Sign Modifications Can Completely Fool Machine Learning Algorithms," *IEEE Spectrum*, August 4, 2017, https://spectrum.ieee.org/cars-that-think/transportation/sensors/slight-street-sign-modifications-can-fool-machine-learning-algorithms.

[79] Kevin Eykholt, et al. "Robust Physical-World Attacks on Deep Learning Visual Classification," arXiv 1707.08945v5 (April 2018), https://arxiv.org/pdf/1707.08945.pdf; Karen Hao, "Hackers Trick a Tesla into Veering into the Wrong Lane," *MIT Technology Review*, April 1, 2019, https://www.technologyreview.com/2019/04/01/65915/hackers-trick-teslas-autopilot-into-veering-towards-oncoming-traffic/.

- **Impact on labor**: These impacts fit into two sets of issues:

    - **AI may increasingly lead to replacing people with computer-based systems**. These shifts can happen in disparate industries ranging from truck drivers (autonomous vehicles) to journalism (so-called robot reporters).[80] However, several experts predict that AI will displace but not replace jobs.[81] There is much uncertainty around the "future of work," and it warrants increased public policy attention given the magnitude of its potential impact.

    - **AI may increase exploitative working conditions and reduce the ability of workers to improve their working conditions**. Proliferation of AI is supported by jobs with labor-intensive and potentially exploitative working conditions (e.g., content moderation, data labeling, safety drivers). Some economists argue that AI gouges out the middle part of the career ladder, creating jobs at the top and bottom rungs of the ladder, while automating out the middle. This results in fewer opportunities for upward mobility out of the lower-rung positions. Additionally, AI algorithms can now manage workers remotely (such as with gig work). This makes it harder for workers to organize and improve their working conditions.

- **False information**: AI generative models can produce fake photos, videos, text, and sounds that appear convincing.[82] This may lead to increased generation of false information. AI algorithms can also help amplify false information to huge audiences. All this can also reduce trust in facts and reliable information.

- **Monopolization of data**: AI systems, in their current state, are extremely data hungry.[83] As previously noted, there is a large incentive for companies and institutions to aggregate large amounts of data that will be advantageous for their learning models and deployment goals. This could cause issues of monopoly around data, computational resources, and ultimately, AI services.

[80] Evan Ackerman, "Slight Street Sign Modifications Can Completely Fool Machine Learning Algorithms," *IEEE Spectrum*, August 4, 2017, https://spectrum.ieee.org/cars-that-think/transportation/sensors/slight-street-sign-modifications-can-fool-machine-learning-algorithms.

[81] Tom Relihan, "Machine Learning Will Redesign, Not Replace, Work," *MIT Sloan*, June 26, 2018, https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-will-redesign-not-replace-work.

[82] Tero Karras, "This Person Does Not Exist," accessed January 23, 2020. https://thispersondoesnotexist.com/; Alec Radford, "Better Language Models and Their Implications," *OpenAI* (blog), December 13, 2019, https://openai.com/blog/better-language-models/.

[83] Willem Sundblad, "Data Is the Foundation for Artificial Intelligence and Machine Learning," *Forbes*, October 18, 2018, https://www.forbes.com/sites/ willemsundbladeurope/2018/10/18/data-is-the-foundation-for-artificial-intelligence-and-machine-learning/#3d9bd53351b4.

  - **Monopoly of AI services and national security**. There is debate about whether Big Tech's growing monopoly of AI services advances[84] or hinders[85] the ability of the United States to compete with China's Big Tech companies. While some argue that U.S. Big Tech companies are a critical alternative/competitor to Chinese services, others argue that stricter antitrust regulation is needed to create more robust domestic competition to advance the U.S. defense industrial base.

- **Environmental footprint:** The continuous growth of artificial neural network parameters—necessary to obtain high performance, robustness, and generalizability—also has a heavy carbon footprint. Using internet data to train the Generative Pre-trained Transformer 3 (GPT-3) language model, for example, OpenAI has spent over $4.6 million in computational resources, which have released an estimated 552 tons of CO2, the equivalent of 460 round-trip flights between San Francisco and New York.[86]

- **Intellectual Property (IP):** Both the data used to train generative AI models and the content produced by those models raise IP concerns.[87] The legal landscape for IP infringement with AI powered tools is still evolving, and it is unclear how courts will handle these types of cases.

  - **Training Data:** Generative AI models are trained on large datasets that include publicly available text, audio, images, video, etc. to generate new content. This can lead to unintended infringement of patents, trademarks, or copyrights. Fair use is a legal doctrine that allows for the limited use of copyrighted material without obtaining permission from the copyright owner, under certain circumstances. Considering fair use doctrine, when are unauthorized uses of copyrighted works prohibited? What responsibility should developers of generative AI models have to conduct IP clearance searches before training their models?

  - **Generated Content:** It is unclear who *owns* the content generated by generative AI tools. Is it the developer of the AI system, the owner of the data used to train the system, or the user who inputs the parameters for generating the content? Additionally, given that AI-based tools are capable of creating content without direct human input, it may not be clear who is *responsible* for the generated content. If generative AI systems are used to create infringing content, should the developer, user, and/or owner of the system be legally liable?

[84] Nitasha Tiku, "Big Tech: Breaking Us Up Will Only Help China," *Wired*, May 23, 2019, https://www.wired.com/story/big-tech-breaking-will-only-help-china/.

[85] Ganesha Sitaraman, "The National Security Case for breaking up big tech," Knight First Amendment Institute, January 30, 2020, https://knightacolumbia.org/content/the-national-security-case-for-breaking-up-big-tech.

[86] David Patterson et al., "Carbon Emissions and Large Neural Network Training" (arXiv, April 23, 2021), https://doi.org/10.48550/arXiv.2104.10350.

[87] "IP Implications of Generative AI." Cooley. Cooley, March 20, 2023. https://www.cooley.com/services/industry/artificial-intelligence/generative-ai.

KEY INSIGHTS

- The performance of AI systems on previously unseen data is at least partially unpredictable. There are processes and methods that can be followed to reduce the risk of obtaining undesired results. Among them, ensuring the quality of training and test data, maximizing models' robustness and generalizability, ensuring that the system decisions are reproducible and interpretable, and, last but not least, monitoring bias, social impact, costs, research opportunities, and environmental footprint.

- AI is a powerful tool that can be used to improve human life, but like many other technologies, it can be used intentionally or unintentionally to create harm. Some of the key concerns around AI is how the technology handles data and whether it could lead to invasion of privacy, discrimination and bias, and a lack of transparency in how it reaches a decision.

Long-Term Concerns

While nearer-term issues require our immediate attention, it is important to consider the potential long-term threats of AI as well. Most common is the idea of the "technological singularity," which is defined as the point when artificial intelligence surpasses human intelligence. While this is conjecture and may never happen, it does raise two additional important questions around AI: (1) how can we align AI systems with the values and goals of humanity (known as the "alignment problem")?; and (2) if achieving some form of the singularity is technologically possible, should we be building toward it at all given the risks and uncertainty around what it could mean for humanity?

That said, it is important to guard against unrealistic expectations or hype over the concept of the singularity. There are significant profit motives behind over-promoting this concept and its possibility. The more it is portrayed as a race that must be won, the easier it is to justify investing vast amounts of public funding and other forms of support.

About the Technology and Public Purpose (TAPP) Project

The arc of innovative progress has reached an inflection point. It is our responsibility to ensure it bends toward public good.

Technological change has brought immeasurable benefits to billions through improved health, productivity, and convenience. Yet as recent events have shown, unless we actively manage their risks to society, new technologies may also bring unforeseen destructive consequences.

Making technological change positive for all is the critical challenge of our time. We ourselves - not only the logic of discovery and market forces - must manage it. To create a future where technology serves humanity as a whole and where public purpose drives innovation, we need a new approach.

Found by former U.S. Secretary of Defense Ash Carter, the TAPP Project works to ensure that emerging technologies are developed and managed in ways that serve the overall public good.

TAPP Project Principles:

1. Technology's advance is inevitable, and it often brings with it much progress for some. Yet, progress for all is not guaranteed. We have an obligation to foresee the dilemmas presented by emerging technology and to generate solutions to them.

2. There is no silver bullet; effective solutions to technology-induced public dilemmas require a mix of government regulation and tech-sector self-governance. The right mix can only result from strong and trusted linkages between the tech sector and government.

3. Ensuring a future where public purpose drives innovation requires the next generation of tech leaders to act; we must train and inspire them to implement sustainable solutions and carry the torch.

For more information, visit: www.belfercenter.org/TAPP