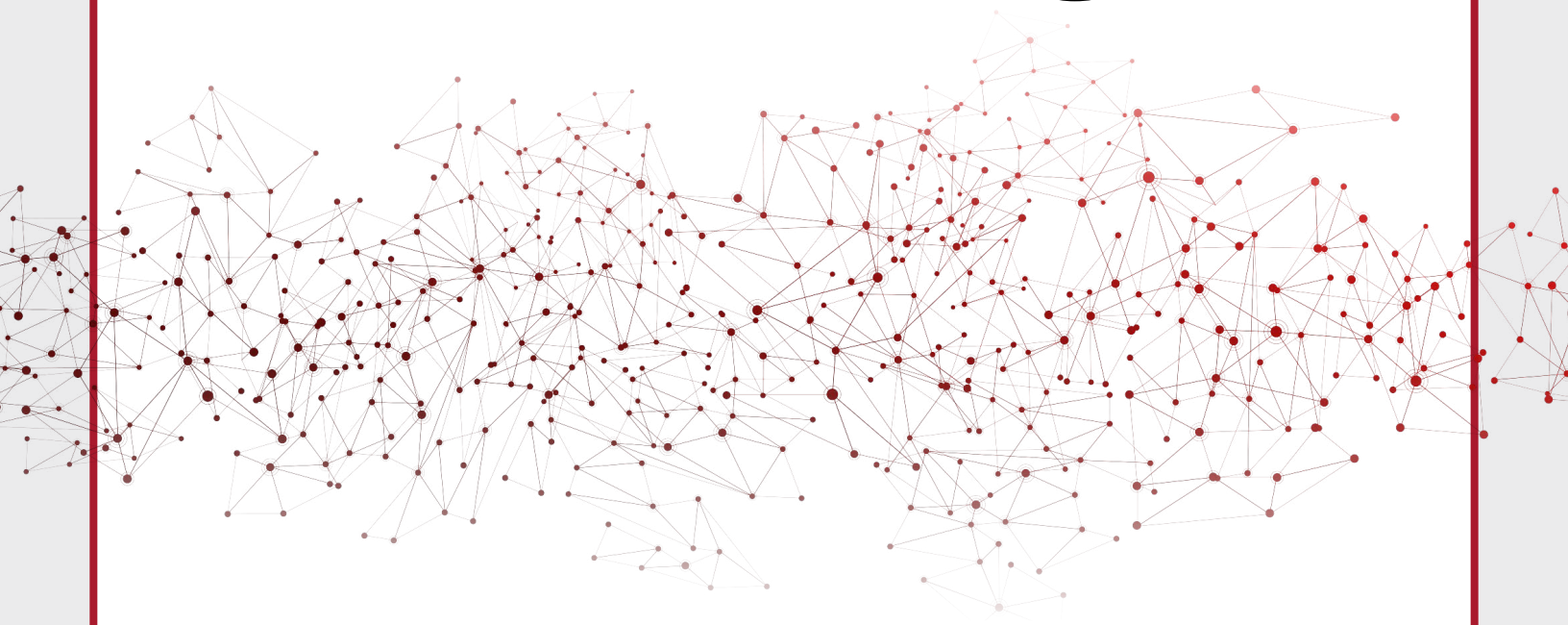
TECHNOLOGY FACTSHEET SERIES

Artificial Intelligence



CONTRIBUTORS:

Enrico Santus, MIT

Nicolas Christin, Carnegie Mellon

Harshini Jayaram, Harvard

EDITOR:

Amritha Jayanti

HARVARD Kennedy School
BELFER CENTER
for Science and International Affairs
Technology & Public Purpose Project

CRCS Center for Research on Computation and Society
at Harvard John A. Paulson School of Engineering and Applied Sciences

The Technology Factsheet 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

CRCS Center for Research on Computation and Society

Harvard John A. Paulson School of Engineering and Applied Sciences
29 Oxford Street, Cambridge, MA 02138

crcs.seas.harvard.edu

Statements and views expressed in this publication are solely those of the authors and do not imply endorsement by Harvard University, Harvard Kennedy School, the Belfer Center for Science and International Affairs, or the Center for Research on Computation and Applied Sciences.

Design and layout by Andrew Facini

Copyright 2020, President and Fellows of Harvard College

Printed in the United States of America

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 labeling and generation, speech recognition and synthesis, natural language understanding and production, as well as various other perception-action based engagements. AI, in its current technological state, is being applied to various industries and domains, such as online advertising, financial trading, medical diagnostics, and robotics. The lucrative market opportunities offered by AI applications have attracted investments from tech giants like Google, Apple, Amazon, and Microsoft, as well as research universities and startups.

Currently, United States policy with regards to AI often derives from interpretations of various pre-existing legislations and legal precedents. However, there are now policies being crafted around the development and deployment of AI technologies, addressing technology-specific risks and concerns. Acknowledging the trajectory of the technology and its potential applications, there is a pressing need for U.S. legislators and policymakers to remain engaged in the ethical and practical development of artificial intelligence.

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 scoping discussions and research initiatives. As previously stated, AI most often refers to the theory and application of algorithms and computer systems to perform tasks that normally require human intelligence.

The term *artificial intelligence* was coined at a conference at Dartmouth[1] 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. Part of the reason why AI is becoming more prominent is due to heightened sophistication of algorithms and expanding computational capacity; recent increases in technical capabilities have allowed for more use-cases throughout industry and society. This increase in capability can be attributed to:

1 Anyoha, Rockwell. "The History of Artificial Intelligence". http://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/

1. The Internet and its enormous amount of data (i.e. Big Data),[2] which is becoming a precious resource for training, testing, and applying AI

2. Faster and more capable computer hardware, which allows processing of large datasets. 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. Better models and algorithms which allow us to do more with the data and processing power.

Taxonomy

AI systems can be classified according to numerous criteria:

1. **Scope of Intelligence:** Artificial Intelligence is often classified as being either *narrow AI,* also known as *weak AI,* or *artificial general intelligence (AGI),* also known as *strong AI* or *general AI*. 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. They can apply their learning and perception-action frameworks to one task domain, but they cannot apply them effectively beyond that. AGI characterizes AI applications that can move between domains and ultimately apply their intelligence more broadly.

2. **Approaches**: Where learning happens and the technical approaches to implementation vary. There are a few common approaches:

    a. Symbolic Reasoning (or Symbolic AI/Classical AI): A branch of AI research that focuses on explicitly representing human knowledge in a declarative form through axioms and rules. Symbolic AI become less popular after the late 1980s when Machine Learning techniques began to become more prominent.

    b. Machine Learning (ML): A subfield of AI that leverages statistical methods or numerical optimization techniques to construct models from data without explicitly programming every parameter or step taken by the computer system.

2 "The Digital Universe Opportunities: Rich Data and the Increasing Value of the Internet of Things." https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htmArtificial

i. Artificial Neural Networks (NN): a loosely biologically-inspired Machine Learning approach, where artificial neurons are connected through weights, which are adapted during learning to optimize performance in a specific task.

ii. Deep Learning (DL): An extension of neural networks that combines multiple neural layers with nonlinear functions (e.g. sigmoid) to derive more powerful representations from raw data aimed at solving specific tasks.

3. **Learning Paradigms**: Learning is classified as supervised (i.e. through human annotated data), unsupervised (i.e. automatically observing similarities between data points), or reinforcement (i.e. using a "reward function" to provide feedback to systems after attempts of completing a given objective). Approaches based on unsupervised learning are preferred because they do not need to be trained on expensive annotations; however, as of today, unsupervised approaches are generally outperformed by supervised ones. Semi-supervised methods are sometimes used to reduce the cost of data annotation. Reinforcement learning is promising—it gained fame during its use by DeepMind for Alphago Zero, the AI system which beat the world champion at the game of Go—but still far from satisfying performance expectations on open space problems.[3]

4. **Tasks**: Machine Learning can either be discriminative or generative. Discriminative approaches (also called classifiers) model the boundaries between classes, so that they can estimate whether a data point belongs or not to a given class (e.g. whether a picture represents a horse or a dog). Generative approaches model the characteristics of data points in individual classes, so that they can then generate data points belonging to them (for example, they can generate images of horses and dogs).

    c. Adversarial Neural Networks: 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.

3 Golstein, Bernard. "A Brief Taxonomy of AI." SharperAI, November 1, 2018. https://www.sharper.ai/taxonomy-ai/.

Applied Fields of AI

Several fields of applied AI have become more prominent and will likely come up in any conversation that touches AI. Some of the most important fields today are:

1. **Computer Vision (CV)**: A field of applied AI focused on image and video processing and understanding.

2. **Natural Language Processing (NLP)**: This field of applied AI focuses on assisting computer systems in understanding and utilizing human speech or writing.

3. **Robotics:** The field studying the development of robots, with all their functions (e.g. movements, interaction, etc.). While it may include the application of AI algorithms to robots, robotics and AI are not necessarily dependent technologies.

Product Applications and Market Development

Artificial Intelligence has applications in both physical and digital space. Examples of applications with a physical presence are autonomous vehicles, automated manufacturing, precision farming, robotics, and, to some extent, personal assistant devices (ex: Amazon Echo). Applications that are primarily in the digital realm are Siri, Cortana, Google Assistant, recommendation algorithms such as those we find on Amazon and Netflix, customer service chatbots, and Google Duplex.

Most technology companies and many defense agencies are investing in AI. AI is also at the core of development plans in the healthcare (e.g. precision medicine, health insurance, etc.) and pharma industry (e.g. drug discovery, etc.). Over the years, much of the funding for AI research flowed from DARPA (the Defense Advanced Research Projects Agency). Siri is actually a consumer application that began as a Defense Department research effort.[4]

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

For productized AI solutions, some of the top players are:[5]

- Incumbent tech companies: Google, Amazon, Microsoft, Salesforce, Baidu, Apple, Intel, Facebook, Nvidia, Tencent, and IBM. Many of these companies have AI Principles that guide their work and engagement with the space.

- Government actors:

    - The U.S. DoD's Joint AI Center (JAIC) is an effort to combine various AI efforts throughout the defense ecosystem;

    - China has pledged over $10B for AI initiatives in the private sector and has established two major research organizations focused on AI, Unmanned Systems Research Center (USRC) and the Artificial Intelligence Research Center (AIRC).

    - Russia established a National Center for Artificial Intelligence, echoing the JAIC approach to streamline national initiatives for AI research and development;

    - The European Union, through its Artificial Intelligence for Europe, pledged to put forward $1.5B through 2020 to strengthen research and innovation.

There are also countless companies that are investing in AI and are growing to be among the dominant players. Some of the front-runners noted most frequently are: CrowdStrike, AIBrain, CloudMinds, SenseTime, and Twitter. It is important to note that many leaders in the space, including 14 start-ups valued at least $1 billion, are Chinese companies.[6] Over time, AI will likely become embedded in all computer applications, at least to some extent.

It is also important to note that many universities, nonprofits and research organizations are pioneering technical progress for AI, as well as catalyzing important conversations around what the implications of AI mean for humanity. Some of these include:

- OpenAI, a San Francisco-based company founded by Sam Altman and Elon Musk, focuses on building safe artificial general intelligence (AGI), and ensuring that AGI's benefits are as widely and evenly distributed as possible.[7]

---

5 Divine, John. "Artificial Intelligence Stocks: 10 of the Best AI Stocks to Buy." U.S. News & World Report. 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.

6 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.

7 "OpenAI," n.d. https://openai.com

- DeepMind, an acquired subsidiary of Alphabet located in the United Kingdom, leads Google's focus on the research, discovery, and development of safe AI (less of a consumer-product approach).[8]

- The Partnership on AI pulls together over 80 partners ranging from the private sector to civil society representatives to shape the dialog and practices around Artificial Intelligence.[9]

- The Future of Life Institute focuses on ways to create positive AI applications while working heavily on risk mitigation.[10]

- Oxford Future of Humanity Institute at the University of Oxford[11], and Leverhulme Centre for the Future of Intelligence at the University of Cambridge[12] look at big-picture questions regarding AI and technology-specific risks.

- Many universities in the U.S. are heavily engaged with artificial intelligence research. Berkeley, Carnegie Melon University, Stanford, MIT, and the University of Washington are generally considered among the most prolific.[13]

8 "DeepMind," n.d. https://deepmind.com/

9 "Partnership on AI," n.d. https://www.partnershiponai.org/

10 "Future of Life Institute," n.d. https://futureoflife.org/?cn-reloaded=1

11 "Future of Humanity Institute," n.d. https://www.fhi.ox.ac.uk/

12 "Leverhulme Centre for the Future of Intelligence," n.d. http://lcfi.ac.uk/

13 Nickelsburg, Monica. "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/.

Current State of Governance and Regulation

Much of the news and attention today focuses on AI research and development strategies rather than governance. Many people see AI development as an "race," though many scholars are encouraging developers and policymakers to reject this narrative.[14]

One 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 the specific application. Some examples of how jurisdiction is already dispersed is that autonomous vehicles are subject to the DoT, the Transportation Safety Administration, and similar agencies, while health diagnoses applications are subject to HIPAA and the HHS.

Below are some of the existing governance and legislation regarding AI broadly.

U.S. Domestic, Existing but 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, et al.*[15]

U.S. Domestic, Specifically Created for AI and Applications:

- **U.S. Deprtment of Defense Directive 3000.09—Autonomy in Weapon Systems:** Released in 2012, this DoD directive establishes guidelines for autonomous and semi-autonomous weapons systems, stressing the importance of appropriate human judgement over the use of force e.g. robust human-in-the-loop systems and human-machine teaming.[16]

- **Guidance on Autonomous Vehicles:** Published by the National Highway and Transportation Safety Administration in 2016, this provides an outline for the evolution of driving from today to full automation, and what the roles of humans and vehicles will be at each stage.[17]

---

14 Barnes, Julian E., and Josh Chin. "The New Arms Race in AI." The Wall Street Journal. Dow Jones & Company, March 2, 2018. https://www.wsj.com/articles/the-new-arms-race-in-ai-1520009261.

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

16 Department of Defense."Autonomy in Weapon Systems." November 21, 2012. http://fas.org/irp/doddir/dod/d3000_09.pdf

17 "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.

- **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.[18]

- **Fundamentally Understanding the Usability and Realistic Evolution of Artificial Intelligence Act of 2017:** Introduced in Congress in 2017, this bipartisan act did not become law. However, it provides a look at what the legislature has considered in the past around AI.[19]

- **National Artificial Intelligence Research and Development Strategic Plan:** The White House Select Committee on Artificial Intelligence began looking into updating this plan in 2018.[20]

- **American AI Initiative:** Launched in 2019, the executive order supports funneling federal funding and resources towards AI-specific research while also implementing U.S.-led international AI standards. Additionally, the initiative calls for new research into increasingly AI literacy in U.S. workers.[21]

International:

- **General Data Protection Regulation (GDPR)**: While this EU 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. [22]

- **China AI Standardization:** Created by the Ministry for Industry and Information Tech, this outlines a Chinese development plan and broader strategy for AI.[23]

- **AI Strategy Paper**: The European Commission outlines an AI development strategy for the EU. [24]

---

18 "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.

19 Weaver, John. "Everything Is Not Terminator: America's First AI Legislation." mondaq. The Journal of Robotics, Artificial Intelligence & Law, August 3, 2018. http://www.mondaq.com/unitedstates/x/724056/new technology/The content of this article is intended to provide a general guide.

20 "The National Artificial Intelligence Research And Development Strategic Plan." The Networking and Information Technology Research and Development (NITRD) Program, October 2016. https://www.nitrd.gov/news/national_ai_rd_strategic_plan.aspx.

21 Minevich, Mark. "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/.

22 Kaput, Mike. "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.

23 "China's Framework of AI Standards Moves Ahead." The National Law Review, July 16, 2018. https://www.natlawreview.com/article/china-s-framework-ai-standards-moves-ahead.

24 Teffer, Peter. "EU in Race to Set Global Artificial Intelligence Ethics Standards." EUobserver, April 25, 2018. https://euobserver.com/science/141681.

- **Communication on Artificial Intelligence**: This paper from the EU touches on many elements of Artificial Intelligence from research and development to the workforce and ethics.[25]

- **Ethics Guidelines for Trustworthy AI**: These guidelines, presented in 2019, put forward seven requirements that AI systems should meet in order to be deemed trustworthy: human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental well-being; and accountability.[26]

Over 20 countries have had some traction around AI investments or regulation since 2017. Many of these are strategy announcements, but some also include budget commitments or task force creation.[27] While these existing pieces of legislation and strategies have focused on individual sovereignties, world leaders from French President Emmanuel Macron[28] to Chinese President Xi Jinping[29] have talked about the need for international cooperation on AI regulation.

25 "Communication Artificial Intelligence for Europe." Digital Single Market - European Commission, April 25, 2018. https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe.

26 Lemonne, Eric. "Ethics Guidelines for Trustworthy AI." Futurium - European Commission, April 30, 2019. https://ec.europa.eu/futurium/en/ai-alliance-consultation.

27 Dutton, Tim, June 28, 2018. https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd.

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

29 Knight, Will. "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/.

Public Purpose Considerations

AI applications, in general, are designed to improve human life, not to create problems for people. However, unintended consequences are already appearing. Moreover, as any other tool, AI may be intentionally misused to harm individuals or the society. The idea and field of "AI ethics" has become increasingly popular over the years, as the technology has increasingly entered mainstream light.

Here are some areas of popular concern:

- **Degree of Autonomy**: With machine learning, 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 (human-in-the-loop).

- **Transparency and Interpretability**: For some AI systems, particularly more complex ones, such as deep learning systems, it is hard to know why they take certain decisions. The systems are being trained from extremely large data sets, and often are "black boxed" in their decision-making process (i.e. DL systems have "hidden layers" of neurons between input data and output data and so it is hard to understand how decisions were made within the algorithm). When a decision process is opaque, concerns of meaningful interpretability, responsibility, accountability, and feedback arise. For example, in military applications of AI, such as 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.[30]

- **Discrimination and Bias**: Because artificial intelligence is currently built on algorithms that learn from data, biased datasets or human bias built into the data may create biased artificial intelligence systems.[31] This can result in discrimination. For example, one resume-analyzing system may consistently choose men over women.[32]

- **Privacy:** Because more data yields better AI, there is a strong incentive for companies to collect as much data about individuals as possible.[33] Furthermore, AI can sometimes let companies determine extremely personal information, such as someone's bipolar diagnosis or likelihood to commit

30 Marquart, Sarah. "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/.

31 AI and bias. IBM Research - US. Accessed January 23, 2020. https://www.research.ibm.com/5-in-5/ai-and-bias/.

32 Dastin, J. "Amazon scraps secret AI recruiting Tool that showed bias against women". https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

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

suicide.[34,35] 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.[36]

- **Security:** Many investments in and applications of AI are associated with military and defense organizations.[37] 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.[38,39] For example, there are concerns that targeted interference could get an autonomous vehicle to read a stop sign as an increased speed limit sign.[40]

- **Impact on Labor:** Artificial intelligence is likely to 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").[41] However, several experts predict that AI will displace, but not replace, jobs.[42] There is much uncertainty around the "Future of Work" and it warrants increased public policy attention given the magnitude of potential impact.

- **False Information:** AI generative models can produce fake photos, videos, text, and sounds that appear convincing.[43,44] This may lead to photos, videos, and audio no longer being compelling evidence that something has taken place.

- **Monopolization of data:** AI systems, in their current state, are extremely data hungry.[45] There is, again, 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.

---

34 "Your Tweets Could Show If You Need Help for Bipolar Disorder." MIT Technology Review, January 5, 2018. https://www.technologyreview.com/s/609900/your-tweets-could-show-if-you-need-help-for-bipolar-disorder/.

35 Marks, Mason. "Suicide Prediction Technology Is Revolutionary. It Badly Needs Oversight." The Washington Post, December 28, 2018. https://www.washingtonpost.com/outlook/suicide-prediction-technology-is-revolutionary-it-badly-needs-oversight/2018/12/20/214d2532-fd6b-11e8-ad40-cdfd0e0dd65a_story.html.

36 Kuo, L. "Chinese surveillance company tracking 2.5m Xinjiang residents." The Guardian, February 18, 2019. https://www.theguardian.com/world/2019/feb/18/chinese-surveillance-company-tracking-25m-xinjiang-residents

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

38 Athalye, Anish, 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/.

39 https://spectrum.ieee.org/cars-that-think/transportation/sensors/slight-street-sign-modifications-can-fool-machine-learning-algorithms

40 Eykholt, K. et all. "Robust Physical-World Attacks on Deep Learning Visual Classification." https://arxiv.org/pdf/1707.08945.pdf

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

42 Relihan, Tom. "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.

43 Karras, Tero. "This Person Does Not Exist." This Person Does Not Exist. Accessed January 23, 2020. https://thispersondoesnotexist.com/.

44 Radford, Alec. "Better Language Models and Their Implications." OpenAI, December 13, 2019. https://openai.com/blog/better-language-models/.

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

Long-term Concerns

While nearer-term issues require our immediate attention, it is important to consider the potential long-term threats as well. Most common is the idea of the "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")? (2) If achieving some form of the Singularity is technologically possible, should we be building towards it at all given the risks and uncertainty around what it could mean for humanity?

APPENDIX A:

Key Questions for Regulation and Governance of Artificial Intelligence

Artificial intelligence is predicted to transform society—from business to labor to transportation to weapons systems, and beyond—as we know it. We are already experiencing some transformations from AI in present day, and so the need to consider meaningful regulation and governance of the technology is becoming increasingly important. Based on the above public purpose considerations, there are several questions to consider asking about various AI systems:

Nature of Autonomy

- What specific decisions and actions are being automated by the AI system, and in what context?
- Are there any components of the product that are fully owned by human agents? If so, how do these components integrate and interact with the AI system(s)?
- Can a human agent override actions and decisions of the AI system(s)?
    - What situational context is provided to human agents (or "human operators") in order to know when it would, or would not, be appropriate to override the system?
    - What kind of human-computer training is provided, if any at all, to make sure that human operators can team well with the automated system?
        - » If there is training, is the training required?
- What kind of performance and testing standards are considered?
    - What performance threshold will make/has made the product market-ready?
- Why is automating tasks in this context perceived to be necessary and/or beneficial?
    - What are technical or non-technical alternatives to the automated systems being built? What decision process encourages/incentivizes the current technical approach?

Data

- How is the data being used to train the AI system sourced? How much data is being used?
- What methods are you using to clean, update, and expand your data sets?

Transparency, Interpretability, and Accountability

- What measures have been taken to interpret and understand the system's algorithms?
- Are product users aware that they are interacting with AI and not a human?
- What methods of performance feedback are being leveraged to ensure the algorithms are being updated appropriately (both in development and in deployment)?

Fairness, Bias, and Applications

- What measures are taken to protect against algorithmic bias within the product?
- Have there been checks for discrimination against protected classes? If yes, what about checks that other less recognizable groups were not discriminated against?
- Could someone repurpose and use the system for discrimination, and how? Can it be "weaponized" in other ways?
    - What actions are being taken to protect against these dual-use risks?

Security and Privacy

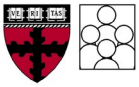
- Are best practices for data storage and cybersecurity being followed?
- Does the AI system interact with minors? If so, what data is being collected and stored, and does this make the minors vulnerable?
- Does the system interact directly with consumers? If so, what data is collected and stored, and is it transparent to the user?
- Does system infer and/or store any sensitive information, such as medical diagnoses? If so, how is that data used and who is it shared with?

- Is the AI foreseeably capable of taking human life? If so, what protections exist against the AI harming humans unintentionally?

Economic Impact

- What are the immediate impacts of the AI system on company/industry employment?
- What are the scaled, long-term impacts of the AI system on employment?
- What protection mechanisms are in place against the monopolization of data and computational resources? Is there a competitive market landscape for AI research, development, and deployment in this domain?

HARVARD Kennedy School
BELFER CENTER
for Science and International Affairs
Technology & Public Purpose Project



CRCS Center for Research on Computation and Society
at Harvard John A. Paulson School of Engineering and Applied Sciences