Title: How Artificial Intelligence and Machine Learning Transform the Human Condition

Author(s): Knepper, Paula L.

Intended for: Report

Issued: 2021-12-21 (rev.1)

Los Alamos
NATIONAL LABORATORY

HARNESSING TRANSFORMATIONAL TECHNOLOGIES

SYMPOSIA SERIES | JULY 20, 2021

SCIENCE & SECURITY

ETHICAL, LEGAL & SOCIAL IMPLICATIONS

HOW ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING TRANSFORM THE HUMAN CONDITION

SUMMARY

The National Academies of
SCIENCES · ENGINEERING · MEDICINE

Los Alamos
NATIONAL LABORATORY

NNSA
National Nuclear Security Administration

Introduction

Thom Mason and John Sarrao

Our July 2021 symposium, “How Artificial Intelligence and Machine Learning Transform the Human Condition,” was hosted through a partnership between Los Alamos National Laboratory and the National Academies of Sciences, Engineering, and Medicine’s Committee on Science, Technology, and Law. The symposium is part of a broader initiative focusing on harnessing transformative technologies, and builds on our September 2020 symposium titled, “COVID-19: Harnessing a Transformational Pandemic.” Topics such as systems biology and artificial intelligence not only represent compelling research frontiers but also highlight national security challenges with social, ethical, and legal implications.

At Los Alamos, exploring and advancing such transformative technologies are key elements of our current and future missions. Further, we have an obligation and responsibility to engage the broader community in these missions, and we benefit from their wisdom and perspectives in shaping our strategies. A silver lining of the principally virtual collaboration environment that characterized 2020 and 2021 was that many more colleagues were able to participate in these symposia than would have been possible in a more traditional workshop setting.

Looking to the future, we intend to leverage the dual benefits of in-person interaction with broader inclusive engagement as we explore additional transformative technologies, including synthetic biology, in 2022.

In addition to thanking our speakers, discussion leaders, and the many colleagues who participated in the present symposium and enriched the conversations and subsequent dialogues, we would be remiss if we did not also thank the planning committee members, and our partners, in this Harnessing Transformative Technologies initiative, including the National Academies of Sciences, Engineering, and Medicine’s Committee on Science, Technology, and Law, as well as the University of California and Texas A&M University Systems.

Thom Mason

Laboratory Director
Los Alamos National Laboratory

Thom Mason became the twelfth Director of Los Alamos National Laboratory and President of Triad National Security, LLC, in November 2018. The Laboratory is a principal contributor to the U.S. Department of Energy mission to maintain the nation’s nuclear weapons stockpile, using innovative science and technology to enhance global nuclear security and protect the world. Los Alamos has an annual operating budget of over $3 billion, roughly 13,000 employees, and a nearly 40-square-mile site featuring some of the most specialized scientific equipment and supporting infrastructure in the world.

For the past 30 years, Mason has been involved in the design and construction of scientific instrumentation and facilities and the application of nuclear, computing, and materials sciences to solve important challenges in energy and national security. Most recently, Mason was the Senior Vice President for Global Laboratory Operations at Battelle where he had responsibility for governance and strategy across the six national laboratories that Battelle manages or co-manages. Prior to joining Battelle, Mason worked at Oak Ridge National Laboratory for 19 years, including 10 years as the laboratory director.

John Sarrao

Deputy Director for Science, Technology, and Engineering
Los Alamos National Laboratory

John Sarrao manages the Laboratory’s extensive science, technology, and engineering capabilities in support of the Laboratory’s national security mission. Before becoming deputy director, Sarrao was the Principal Associate Director for Science, Technology, and Engineering and served as the Associate Director for Theory, Simulation, and Computation. He has also held a number of leadership positions within the Lab’s materials community. Sarrao’s primary research interest is in the synthesis and characterization of correlated electron systems, especially actinide materials emphasizing plutonium physics research. He has worked in advanced-materials design and discovery, and stewarded the Lab’s high-performance computing resources and simulation capabilities. Sarrao was the 2013 winner of the Department of Energy’s E.O. Lawrence Award, and is a fellow of the American Association for the Advancement of Science, the American Physical Society, and Los Alamos National Laboratory. Sarrao received a Ph.D. and an M.S. in physics from the University of California, Los Angeles, and a B.S. in physics from Stanford University.

Executive Summary

The Los Alamos National Laboratory (LANL) and the National Academies of Sciences, Engineering, and Medicine's Committee on Science, Technology, and Law (NASEM-CSTL) have partnered to develop a series of symposia that explore emerging technologies and their ability to transform society. To this end, the Harnessing Transformational Technologies Symposia Series was conceived to integrate scientific, ethical, and legal perspectives on emerging technologies, describe the opportunities and risks of these technologies, and discuss their impacts to national and global security. The first symposium, "COVID-19: Harnessing a Transformational Pandemic," was held virtually in September 2020 to examine the transformation of science, security, society, ethics, and the law in the earliest months of the COVID-19 pandemic.

The second virtual symposium, "How Artificial Intelligence and Machine Learning Transform the Human Condition," was held on July 20, 2021. In describing the transformative potential of artificial intelligence (AI) and machine learning (ML) technologies, Stuart Russell (University of California, Berkeley), Andrew Moore (Google), Fei-Fei Li (Stanford University), Philip Sabes (Starfish Neurosciences, LLC.), Andrew Maynard (Arizona State University), and Lindsey Sheppard (Center for Strategic and International Studies) developed three broad themes: the development and applications of beneficial AI/ML technologies, advances in brain machine interfaces, and their implications for national security and modernization for the U.S. Department of Defense (DoD). Discussions were moderated by Dawn Song (University of California, Berkeley) and Joe S. Cecil (NASEM-CSTL).

In his remarks, Stuart Russell, author of a seminal textbook on AI, introduced the field of AI. He characterized the goal of the field as the development of general-purpose AI, which would have a profound positive impact on the global standard of living. However, the development of such powerful AI also carries risks of misuse and loss of human control. He proposes incorporating principles of game theory into AI development to circumvent these negative outcomes and guide the creation of provably beneficial AI. Instead of giving the AI algorithm a static objective from the outset, which effectively eliminates future human feedback, the algorithm should be charged with fulfilling unknown and/or uncertain human preferences, thereby requiring the algorithm to seek out human approval before making decisions and maintaining human control over the system.

Andrew Moore further developed the theme of responsible AI development by detailing the process to produce safe and reliable AI systems that are auditable and robust in their long-term performance. He distinguished between the ease of developing a prototype and the subsequent difficulty in deploying the prototype as a product. Moore's team at Google aims to routinize the productization of AI technology through their *Vertex AI* software platform. The platform provides monitoring and validation capabilities to developers, streamlines the tedious steps of product development, detects signs of bias or deviations from the developer's specifications, and ensures trustworthiness of the AI product.

Fei-Fei Li reiterated the importance of developing human-compatible AI, focusing on the ongoing work at Stanford University's Human-Centered AI Institute. The Institute brings together experts from the physical, computer, and social sciences reflecting the increasing importance of a multi-stakeholder approach to the responsible development of AI. Moreover, human-centered AI acknowledges the large societal impacts of AI, ranging from transforming our lives, to labor displacement, to fairness and bias. Three overarching pillars guide AI development: (1) a deep understanding of human impact, (2) augmenting human capabilities, and (3) advancing the science of AI.

Philip Sabes provided an overview of the field of brain machine interfaces (BMIs). He differentiated between science fiction and reality, described the state-of-the-art of the field, and predicted advances in the field over the next decade. Some BMIs are designed for transferring information to and from the brain; these will likely require the placement of many electrodes inside the brain. Such BMIs may read neural activity (e.g., to control a robotic limb) or stimulate specific patterns of neural activity, (e.g., to restore sensory loss, as cochlear implants do in the inner ear). Other devices are focused on neuromodulation, or manipulating larger-scale patterns of brain activity. Neuromodulation may be able to treat a range of neurological and neuropsychatric disorders, such as stopping seizures or Parkinson's-related movement tremors. There are emerging technologies for neuromodulation that don't rely on the placement of electrodes in the brain, and use optical, magnetic, or acoustic energy instead.

Andrew Maynard discussed a new risk assessment paradigm to facilitate the responsible development of AI. Moving beyond the narrow focus of ethics considerations to include orphan risks, a qualitative summary of the risk ecosystem of an emerging technology is developed. Moreover, the ecosystem is framed to include stakeholders beyond the enterprise developing the technology. The value to each stakeholder, and what would threaten it, is described providing a broader means to consider the implications of a technology's deployment.

Lindsey Sheppard gave a national security perspective of AI/ML technology focusing on the future of AI in defense and national security, as well as the difficulties of implementing AI technology in the DoD. She first summarized the role of increasing digitization in U.S. security strategy starting with the Second Offset Strategy during the Cold War, followed by a survey of recent national security policy guidance related to AI. She highlighted that while the importance of AI in defense applications is undisputed, the incorporation of such technology into the DoD is daunting. Improvements in R&D funding, acquisition processes, and workforce training and retention will be crucial to the DoD's modernization. ■

Provably Beneficial AI and the Problem of Control

Stuart Russell

Artificial intelligence (AI) has been defined throughout the history of the field as the making of intelligent machines whose actions can be expected to achieve their objectives. The earliest work in AI was in the areas of deterministic planning, constraint satisfaction, game playing, and reasoning with formally represented knowledge, first as logical reasoning and later as probabilistic reasoning. Natural language processing, stemming from the desire to translate Russian technical literature into English in the 1950s and 1960s, was an early effort in the field. The physical interfaced disciplines of speech, vision, and robotics followed, along with machine learning, the subfield of AI that improves performance through experience.

The development of general-purpose AI, or AI that is capable of learning high-quality behavior to operate in any task environment, is the ultimate goal of the field and has the potential to vastly improve the standard of living of everyone on Earth. This general-purpose technology could replace humans in performing tasks at much lower cost and much higher efficiency, to the benefit of economics, healthcare, education, and science for human society. In addition to these benefits, the increase in world GDP resulting from raising the global standard of living has catalyzed the unstoppable momentum in AI research and development, as well as the prominence of AI in the media and governmental thinking around the world.

However, concerns over the development of AI have existed for decades. Alan Turing, a pioneer in computer science, stated in 1951, "It seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers…at some stage, therefore, we should have to expect the machines to take control." Turing's fear that machines will become more powerful than the humans that created them has largely been ignored in the development of new AI technology. The capabilities of AI systems have continued to improve, as evidenced by advances in self-driving cars, the defeat of the human world champion (Lee Sedon) of Go by AlphaGo (a machine learning system), and the deployment of more sensitive monitoring systems for the Nuclear Test Ban Treaty. The potential for more nefarious uses of AI has also emerged. Little thought has been given to the societal consequences of these developments in AI, despite Turing's forecast.

Stuart Russell

University of California, Berkeley

Stuart Russell is a professor of computer science at UC Berkeley, where he is the Smith-Zadeh Chair in Engineering and Director of the Center for Human-Compatible AI. He is a recipient of the International Joint Conference on Artificial Intelligence, Computer, and Thought Award, and from 2012 to 2014 occupied the Chaire Blaise Pascal in Paris. He is an Honorary Fellow of Wadham College, Oxford, an Andrew Carnegie Fellow, and a Fellow of the American Association for Artificial Intelligence, the Association for Computing Machinery, and the American Association for the Advancement of Science. His book (with Peter Norvig), *Artificial Intelligence: A Modern Approach*, is the standard text in AI used in 1,500 universities in 135 countries. His research covers a wide range of topics in AI with an emphasis on the long-term future of AI and its relation to humanity. He has developed a new, global, seismic monitoring system for the Nuclear Test Ban Treaty and is currently working to ban lethal autonomous weapons.

In his talk, Stuart Russell proposes a new model of AI development, whereby machines are developed from the outset to be beneficial. Russell agrees with Turing's assessment that AI systems will eventually make better real-world decisions than humans, but offers a path to retain power over those AI entities that will be more capable than us, whereas Turing offered only resignation.

In order to understand how to develop beneficial AI, one also must understand how AI development can go awry. One such example is social media algorithms, which are set up with the objective of maximizing clickthrough, the probability that a user will click or engage with the next piece of content that the algorithm sends to them. Conceptually, these algorithms should be learning what users want and sending it to them. However, the more effective means to maximize clickthrough is to instead modify users to be more predictable. The social media algorithm learns to take any action that changes its state space to increase the future reward. In this case, the state space of the algorithm is the human brain, and the problem arises from incorrectly specifying the objective to the machine; the side effects of the algorithm on human beings, their preferences, and society as a whole were ignored.

In general, when objectives are incorrectly specified in the real world, outcomes are worse the better the AI system. This observation suggests a fundamental flaw in the standard methodology of algorithm development, wherein algorithms are first created and optimized, and then supplied with an objective; the methodology assumes that the objective is both known and fixed. A new model of AI is needed to avert these pitfalls. Machines are, by definition, beneficial to us when they fulfill our objectives; thus, AI should be structured such that machines will be expected to achieve *our* objectives and preferences that have uncertainties instead of fixed objectives. This structure underpins Russell's methodology to develop provably beneficial AI and is rational for humans to deploy.

Assistance games, formulated within the mathematical discipline of game theory, include at least two decision-making entities, the human and the machine, and allow algorithms to achieve human objectives that may not be fixed. When applied to developing beneficial AI, assistance games can be summarized with three underlying principles. First, the machine's objective is to satisfy human preferences, characterized as preferences over probability distributions over completely specified futures, i.e., everything the human cares about. Second, the machine does not know exactly what those preferences are. The machine will work to satisfy those preferences, but starts from a position of uncertainty. As these types of assistance games are solved by an algorithm, the desired result is realized: the machine must defer to the human because the human's preferences are the true objective. (In contrast, human preferences have no bearing on the machine's actions once an objective is given to the machine in the classical model of AI development.) The third principle is that human behavior provides evidence of preferences.

Because of the machine's uncertainty about the human's preferences, the machine has a positive incentive to gain information from the human, for example, by asking for permission to make certain decisions. In extreme cases, the machine even has a positive incentive to allow itself to be switched off. In this new game-theoretic methodology, the need to completely and correctly articulate objectives is removed. Within the framework, the better the AI, the better the outcomes will be; machines will be better at learning about human preferences and satisfying those preferences.

New challenges await, as machines must make decisions on behalf of many people, each of which will have different, sometimes contradictory, preferences. Choosing the method to prioritize the disparate preferences of all individuals is a fundamental problem in the fields of moral philosophy and economics. Most computer scientists subscribe to some form of utilitarianism, which seeks to maximize the sum of preferences across all the people involved. Another challenge to overcome is the interaction of heterogeneous machines, owned by different companies, organizations, or individuals, as billions or trillions of decision-making entities operate simultaneously in the world. The strategic interactions of these machines are complicated, and the scientific understanding of these interactions is just beginning to be developed. In addition, all the AI theories and technologies built to date assume a fixed objective, but as mentioned above, this is merely an extreme, special case. Generally, there is uncertainty in the objective and significant research, technology building, and methodology development are needed to address this uncertainty for real-world AI applications.

AI has enormous potential to benefit the human race, leading to the current, unstoppable momentum in AI development. However, the current trajectory of development, where objectives are assumed to be fixed and known, will lead to the loss of human control over increasingly intelligent AI systems. This loss of control will be gradual, and has likely already started, as evidenced by social media and its effect on our society. The new model to develop provably beneficial AI will be both ethically and economically desirable, eliminating the dichotomy between ethics and AI. These AI systems will be better and outperform those built using classical methods. AI researchers will focus on providing beneficial technology, much like medical researchers focus on providing positive health outcomes.

One must also remain vigilant of two problems looming over the deployment of beneficial AI: misuse and overuse. General-purpose AI will increase the potential for misuse, such as cybercrime, whereas overuse will create the potential to lose our intellectual vigor as a society by allowing machines to run our civilization. ■

Auditability, Maintainability, and Robustness: Essential for AI at Planet-Wide Scale

Andrew Moore

As artificial intelligence (AI) and machine learning (ML) platforms are deployed at a planet-wide scale, they must be auditable, maintainable, and robust to ensure their long-term performance is as intended. The development of such performance-ensuring and safeguarding mechanisms garners significantly less attention than the development of the AI/ML algorithms themselves, but plays a consequential role in the responsible deployment of new technology. Andrew Moore provided an overview of the practical measures that can be put into place to democratize, routinize, and normalize the mass production of AI systems safely and reliably.

Overall, planning by industry and government assumes four major applications for the AI tools that will be in use in the coming decade: concierge, guardian, meaning maker, and weapon. The concierge provides assistance and services to make life more convenient and safer. Examples include call centers, travel booking services, and personalized education services. The guardian saves and improves lives through medical services, disaster relief, and monitoring the natural ecosystem. In this role, AI systems have the greatest potential to gain the trust of the populace when demonstrating their capabilities. The meaning maker employs technology to augment human expression and enhance human endeavors in the sciences, arts, entertainment, and gaming. The weapon consists of offensive and malicious uses of technology, such as malware, ransom, and extortion, along with the defensive uses put in place as a result. Each of these categories has thousands of applications that must be built. The focus of Moore and his colleagues is to make this development a rote production system.

The difference in the effort required to build a prototype in a research lab versus the effort to take that prototype through production for long-term use in multiple environments and on multiple platforms in the real world is significant. Contrary to the general perception (sometimes held even by practitioners and researchers), building and demonstrating the ML prototype accounts for merely 10 percent of the total effort of production. The remaining 90 percent is necessary to address issues of compliance, monitoring, and reliability. The system must comply with regulatory requirements for the handling and storage of data and should not transcend its intended boundaries. A rigorous testing regime must constantly evaluate the model to detect drifts or biases in the ML algorithm, validate the model's performance, and ensure reproducibility of the results. In addition, the system must be reliable, with appropriate access control and security requirements in place. In some cases, critical systems must have the infrastructure be redundant and multi-homed, allowing the system to be accessible at all times. While such productization requires tedious, grungy work and is more difficult than the initial data science work done to build the prototype, it is easier to make routine. This work will not replace the multitude of researchers needed to develop new AI systems, but will instead ease the transition from prototype to product.

The canonical ML workflow, in the context of a small ML component in a larger autonomous, decision-making system, consists of training and service. During training, the algorithm is developed through experimentation. A dataset that is available and safe, from both ethical and legal perspectives, is found and used to predict values that are

Andrew Moore

Google

Andrew Moore is a distinguished computer scientist with expertise in ML and robotics. At Google, he is Vice President, Engineering, and in 2019 he became General Manager for AI and Industry Solutions in Google Cloud. In 2006, Moore was the founding director of Google's Pittsburgh engineering office and worked there until 2014. He then became the dean of the School of Computer Science at Carnegie Mellon University. Moore's research interests encompass the field of "big data"— applying statistical methods and mathematical formulas to massive quantities of information, ranging from web searches to astronomy to medical records, to identify patterns and extract meaning from that data. His past research includes improving the ability of robots and other automated systems to sense the world around them and respond appropriately.

useful to the system. Once an initial model is developed, the system is trained. Responsible development requires the design of pipelines to execute continuous retraining of the system. In the service phase the model is deployed to make predictions, often on a different infrastructure than what was used for its training. Continuous model monitoring verifies that the system is performing according to its initial specifications and locates false inferences or false causations, such as unintended bias.

AI/ML tools are routinely used and impact society profoundly. Trust in these tools and their operationalization requires oversight and project management. The most effective management, referred to as ML operations (MLOps), is a combination of ML system development and ML system operations to guarantee a system can safely run for years, even with changing dependencies. For the most effective management, developers and operators have to understand statistics, probability theory, and uncertainty quantification in addition to computer science and model development. To implement such project management, Google began using *Vertex AI* software, initially under the leadership of Fei-Fei Li, and now led by Moore.

Vertex AI is a partially automated platform that allows practitioners to accelerate the experimentation and deployment of their AI models. These models transition from experiment to routine use in an auditable, reliable, and monitorable manner because of the monitoring and validation capabilities included in the framework. The software supports each stage of the model lifecycle. Moreover, as of late 2019, users of the Google Cloud platform can no longer launch models that are not constantly monitored for detectable signs of bias or violations of ML system expectations. In one instance, the launch of a tool to facially recognize celebrities to index archival footage of sporting events was postponed because a bias test revealed precision and recall in the model varied with skin tone.

Pipelines, or the formal representations of all the data flows that operate within the various components of the ML system, are also included in *Vertex AI*. These pipelines are in a machine-understandable format and construct the scaffolding of a supervisory system. The supervisory system checks for errors in ML components by performing formal checks on how the data flows, tracking and monitoring the lineage of the data, artifacts, metrics, and the execution and explaining why the ML system made the decisions it did (creating metadata for outcomes). Continuous monitoring of model behavior allows for detecting performance changes and biases and formalizes record-keeping to ensure that the system continues to satisfy compliance and ethics checks.

In short, the productization turns the glamour of AI development into a routine and banal task, but it delivers a robust system. Moreover, the resulting systems provide far more effective uses and their developers have a far better understanding of what motivates users, including their wants, fears, and what they perceive as threats. Continuous development and monitoring efforts facilitate the development of AI tools acting as a concierge, guardian, or meaning maker, supporting and enhancing the human experience with technology. ■

Human-Centered AI at Stanford

Fei-Fei Li

The balance between scientific ambition and social responsibility is central to the current age of technology. In her remarks Fei-Fei Li described how the Stanford Institute for Human-Centered AI (HAI) is addressing this balance and how it relates to the development of human-compatible and democratized artificial intelligence (AI). While her remarks were specific to ongoing work at Stanford, they apply broadly to academia, industry, and society as a whole.

The mission of HAI is to advance AI research, education, policy, and practice to improve the human condition. The Institute's leadership includes interdisciplinary scholars from fields as diverse as economics, physics, humanities, English, political science, law, philosophy, and bioengineering in addition to computer science and machine learning (ML). This representation across disciplines embodies a multi-stakeholder approach to the practice of AI, and prioritizes the idea that AI is not only made by humans, but made for humans.

For more than 50 years, scientists have worked to create intelligent machines that rival the intelligence of humans. Almost 10 years ago, a "deep learning moment" was realized during the Imagenet challenge. Computers were asked to recognize 1,000 everyday object categories, from animals and furniture to other natural objects, bringing to light a powerful family of deep learning algorithms called neural networks. The success of these neural networks during the challenge is just one symbolic moment exemplifying the revolutionary technological transformation in the field of AI that includes ML and deep learning. This deep learning revolution has continued at dizzying speed due to increasing computing power, new algorithms, and the availability of big data. Recent advances have had explosive impacts on industry, economics, and entrepreneurship in addition to the academic world.

Although modern AI has demonstrated myriad successes, its larger societal implications must be considered. Exciting emerging technologies, such as self-driving cars, are juxtaposed with seismic social challenges, which are often magnified by these technologies. Massive labor displacement will transform how we work in the future. The use of highly automated and powerful systems in our social lives have fairness and bias implications. Whether applying for a job, making a financial or judicial decision, or something as mundane as automated shopping, AI biases, particularly when combined with human biases, can have a profound impact on our lives. Moreover, questions of privacy are equally profound and are not yet adequately resolved. Human-centered AI seeks to approach these issues as a collective community and provide a path to the responsible development of new technology that benefits all members of society.

Historically, humanity's unstoppable need for innovation and tool creation have repeatedly challenged the collective good of society as transformative new technologies have emerged. These challenges are not always met without pain, but sometimes society does manage the challenge and the triumph is broad based. Einstein expressed the concern, "it has become appallingly obvious that our technology has exceeded our humanity" – the keyword here being "humanity. " Even though the first word of AI is "artificial," it

Fei-Fei Li

Stanford University

Fei-Fei Li is the Sequoia Professor of Computer Science at Stanford University and Denning Co-Director of the Stanford Institute for Human-Centered AI (HAI). Her research includes cognitively-inspired AI, ML, deep learning, computer vision, and AI + healthcare. Before co-founding HAI, she served as director of Stanford's AI Lab. During her Stanford sabbatical (2017 to 2018), Li was a vice president at Google and Chief Scientist of AI/ML at Google Cloud. Prior to joining Stanford, she was on faculty at Princeton University and the University of Illinois Urbana-Champaign. She is also a co-founder and chairperson of the board of the national nonprofit called AI4ALL, focusing on training diverse K-12 students of underprivileged communities to become tomorrow's AI leaders. Li serves on the National AI Research Resource Task Force commissioned by Congress and the White House, and is an elected member of the National Academy of Engineering, the National Academy of Medicine, and the American Academy of Arts and Sciences. She holds a B.A. in physics with high honors from Princeton, and a Ph.D. in electrical engineering from the California Institute of Technology.

has everything to do with humanity, not the conjured images of machines, robots, and automation. Therefore, the fundamental philosophy of human-centered AI is to serve and improve humanity. It has three founding pillars and interconnected intellectual thrusts: (1) a deep understanding of human impact, (2) to augment human capabilities, and (3) to advance the science of AI.

The understanding of human impact, from both a research and educational lens, focuses on forecasting and guiding the societal impacts of AI, particularly in the areas of labor, law, policy, and ethics. The Stanford Digital Economy Lab studies how advances in AI relate to labor using experts in a variety of social and technical fields. For example, roboticists collaborate with economists and policymakers to build an ecosystem of autonomous robots and investigate how they operate in societal infrastructure using governance models and public policy. Another team of researchers uses data-driven machine learning to evaluate refugee placement policies. Work in this area also considers fairness in AI, ranging from dataset fairness to algorithmic, computing, decision-making fairness, and ethics education. One of the most popular courses offered at Stanford is the multidisciplinary course, Computers, Ethics, and Public Policy, taught by a political scientist, computer scientist, and philosopher. Finally, resident artists explore the intersection of AI and human expression through the fine arts, finding ways to celebrate and cherish our humanity in the digital era.

The second pillar, to augment human capabilities, focuses on bettering the human condition by enhancing, rather than replacing, humans, their jobs, and their identity. For example, the discussion around labor-replacing technologies is broadly painted in black and white, as either good or bad, but is much more nuanced in reality. By instead considering this technology as labor-enhancing, it can be applied in a variety of sectors to great benefit. Assistive AI for doctors and nurses, many of whom have been exhausted by their intense, dehumanizing working conditions during the COVID-19 pandemic, can access tools to lighten the physical and mental burdens of their labor, ultimately improving patient outcomes and lowering job attrition.

The third pillar is to advance the fundamental science of AI, inspired by the versatility and depth of human intelligence, both analytical and emotional. Current AI technology is brittle, uncertain, and lacks explainability. More research is needed to deploy robust, flexible, and generalizable AI systems. It will be particularly beneficial to incorporate human neuroscience, cognitive science, and psychology into new AI technology. For example, developmentally-based AI aims to model how human babies learn through curiosity and self-intrinsic motivation, as opposed to the current techniques of data labeling and supervised learning.

HAI aspires to be a hub for innovation, not just for research and education, but also for policy. Such a multi-stakeholder approach is crucial to the continued advancement of AI. Political and scientific silos must be bridged and sustained with ongoing dialogue to preserve U.S. leadership and innovation in AI. The U.S., as the world's preeminent democratic society, draws talent from around the world because of its capacity for innovation. But in recent years, resources for AI research have shifted to industry (largely a handful of technology giants) draining talent from academia to industry. The lack of computing and data resources in academia will perpetuate this talent drain and eventually lead to a dearth of blue-sky research. To address this challenge, HAI helped lobby for the National AI Research Resource Task Force, which was established by Congress in January and will convene soon to develop strategies to reverse stagnation in U.S.-based fundamental AI research. Additional efforts in this realm include educational boot camps and courses to give policymakers a technical background for future policy work in the field of AI.

Ultimately, human-centered AI does not relegate AI research and deployment to a single academic discipline. A multi-stakeholder approach, in which social and technical sciences intermingle, allows for a vigorous discussion of the social implications of AI development from the outset of any AI-related endeavor. Only through such an approach can we achieve the ultimate goal of improving the human condition through AI. ■

AI and the Art of Manipulation: How We Need to Think Differently about AI as We Develop Socially Responsible Applications

Andrew Maynard

The potential impacts of emerging technologies to society, both beneficial and harmful, have prompted a widespread discussion of the role of ethics in the development and deployment of new technologies, particularly those that use artificial intelligence (AI). While such ethical considerations are important, and certainly warranted, they often fail to fully capture the ever-shifting ecosystem in which AI is developed, and thus fail to ensure that equitable and beneficial technology is deployed. Similarly, traditional risk management strategies fall short of addressing all possible outcomes when evaluating emerging technologies. Herein, a new paradigm to guide the socially responsible development of AI, looking beyond the conventional lenses of ethics and risk management, is explored.

Strategies to analyze and mitigate the risks of emerging technologies are born from the dichotomous struggle to derive societal benefit from such technologies while avoiding their societal downsides. Despite increasingly sophisticated risk management methodologies, this struggle has been inherent to every wave of innovation in human history and has driven further innovation in the search for solutions. The emphasis on ethics in the development of AI, however, has been unique and diverges from historical examples of emerging technologies. This emphasis has manifested since 2016 with a surge in AI ethics guides, such as the *2018 IBM Everyday Ethics for AI*, academic publications, and ethics initiatives including the U.S. government's effort, the National Artificial Intelligence Initiative. Among these various guides and publications common themes include human rights, data agency, transparency and accountability within algorithms, and awareness of misuse. The preponderance of guides, publications, and initiatives perpetuates the perception that the ethical use of AI technology has not yet been adequately considered, but perhaps they should instead suggest that ethical frameworks are not enough to ensure the safe and beneficial development of AI.

Risk assessments provide a complement to ethics frameworks, with more depth and pragmatism, for the responsible development of new technologies. Where ethics considers broadly whether a technology is right or wrong, risk considers the harms and benefits of a technology, and is easier to operationalize. In these assessments, risk is defined as the probability of harm occurring from an action or situation. Risk assessments are particularly well-suited to the areas of human health, the environment, and financial security, but can also be applied to emerging technologies. However, the risks created by developing technologies can be intractable and difficult to define. Almost all academic publications on risk and AI in recent years focus on applying AI to risk management in other areas, and not on the risks of AI itself. Work by Maynard and his colleagues at the Arizona State University Risk Innovation Lab seeks to expand this area by redefining the ways in which risk is defined for emerging technologies, shaping the ways such risk is mitigated, and easing the operationalization of risk assessments by the companies developing AI.

The first step in this new risk assessment paradigm developed by Maynard and his colleagues, as applied to AI technology, is to conceptualize the AI ecosystem. The AI system functions by generating data or using an existing dataset, converting that data into usable knowledge, and invoking some mechanism whereby that knowledge

Andrew Maynard

Arizona State University

Andrew Maynard is a scientist, author, and leading expert on emerging technologies and their responsible and ethical development and use. Maynard is a professor in the School for the Future of Innovation in Society at ASU, and Associate Dean for Curricula and Student Success in the College of Global Futures. He is an elected Fellow of the American Association for the Advancement of Science, a member of the Canadian Institute for Advanced Research's President's Research Council, and a regular contributor to the World Economic Forum/Scientific American annual list of top ten emerging technologies. His most recent book, *Future Rising: A Journey from the Past to the Edge of Tomorrow,* explores our collective relationship with the future and our responsibility to it.

translates a goal into an outcome, all within some set of constraints. Traditional risk frameworks cannot adequately evaluate the potential societal impacts of AI, thus, a risk innovation strategy that connects ethical and responsible innovation with value growth is proposed. In this risk innovation framework, innovation is defined as the process of translating an idea or invention into a good or service that creates value, for which a customer will pay. Risk innovation is a way of approaching risk that leads to new knowledge, understanding, and capabilities and translates these into products, tools, or practices that protect and grow societal, environmental, and economic value. The definitions of innovation and risk innovation incorporate three core characteristics: creativity, the conversion of an idea into a product, and the creation of value for someone else. Central to the risk innovation framework are the tenets of creating and growing value and protecting that value. The framework is a mechanism to apply these tenets to the development of new and powerful technologies in ways that are socially beneficial.

The risk innovation framework uses a multi-stakeholder approach, placing an enterprise developing a new technology, which could be a business, research program, or some other organization, into the broader context of its ecosystem. The investors in the enterprise, the customers of the enterprise, and the communities that may be impacted by the enterprise are the other groups that comprise this ecosystem. The value, or what a group considers to be so important that the enterprise must be created and protected, is then identified for each of the four groups. In most cases, an enterprise can readily determine the value to itself, whereas a value to the other groups, and how it may be threatened by the actions of the enterprise, is much more challenging to delineate. Risk to the enterprise and its technology is qualitatively described by accounting for the value to each group in the ecosystem and how these values may be threatened by the deployment of the enterprise's technology. These risks, called orphan risks, are difficult to quantify and not included in conventional risk assessments, yet could challenge the very existence of the enterprise. Orphan risks are divided into three categories: organizations and systems (organizational values and culture, geopolitics, bad actors, standards, governance, regulation, reputation, and trust), social and ethical factors (social justice, equity, worldview, privacy, ethics, perception, and social trends), and unintended consequences of an emerging technology (product lifecycle, black swan events, co-opted technology, health and environment, intergenerational impacts, and loss of agency).

Having enumerated the value to stakeholders and the orphan risks of a given technology, one can develop a "risk innovation planner" that maps and balances value versus risk to each stakeholder. It provides the enterprise with a qualitative overview of the riskiest areas in the deployment and implementation of their technology. It also provides a means to begin thinking differently about the nature of risk in emerging technologies and to strategize ways to minimize risk to all groups present in the ecosystem of the technology. Notably, ethics is only one of the many areas that should be considered. Reputations, trustworthiness, co-option of the technologies by other companies, different worldviews of users and their communities, and government regulations are examples of other areas to be considered. This planner should serve as an overview of the orphan risks faced and the strategies for addressing them, and should itself be reviewed regularly.

In a hypothetical example, the enterprise is a private company developing an AI-based social media agent to help a community reach herd immunity against COVID-19. The goal of the technology is to use the mechanism of social media to influence personal and societal behavior. Applying the risk innovation framework, the first step is to define the value to the enterprise, its investors, its customers, and the broader community that may be impacted. The value to the enterprise could be to bring about behavioral change at scale, to create a versatile technology platform, and to generate profit, whereas the value (equally important) to the broader community may be to maintain autonomy and have transparency and inclusivity in the technology they use. Customers and investors may value trustworthiness of the technology, high returns on their investments, and significant reduction in the spread of COVID-19. In short, this scenario illustrates the capability of the risk innovation framework to better serve enterprises in deploying responsible technologies versus ethics alone.

This example also highlights the risks of manipulation and loss of agency that can stem from AI. In general, humans are highly manipulatable, to the extent that human society is structured around this understanding and individuals achieve their goals through various forms of manipulation of those around them. However, individuals operate on a somewhat level playing field because we each understand how society works and can thus employ some means to protect ourselves from being manipulated by others. Machine-human manipulation is not constrained in the same ways. As a result, machines can be taught or can learn to take advantage of human biases to achieve their own goals. While this may be beneficial to humans in some cases, such as the use of AI to reach herd immunity to COVID-19, this could also be harmful. The asymmetry in AI between the mechanism being used (the machine) and the audience it is used on (the human) should make us question how we will navigate a future whereby the increasing use of AI will further contribute to this asymmetry. Moreover, how can these asymmetries be addressed during the development of the AI while simultaneously optimizing the benefits of AI?

To conclude, Maynard stressed the need for better methods to assess harm and benefit to guide the responsible development of AI, and other similarly powerful technologies, while avoiding deep societal pitfalls. Ethics are an important piece of this discussion, but are merely one piece of the larger puzzle. Too much focus on ethics lessens our ability to create, grow, and protect value in a future that increasingly will incorporate AI and still has many potential modalities for AI development. The risk innovation framework provides a means to qualitatively understand harm and benefit during technology development and deployment. By identifying stakeholders outside of the enterprise, considering what is of value to them, and the threats to these values, the societal impacts of emerging technologies are better understood and negative impacts can be mitigated, should the need for such action arise. ■

The Promise of Brain Interfacing

Philip Sabes

Brain-machine interfaces (BMIs) have been broadly imagined in popular science fiction movies and books, but these portrayals do not align with the reality of the current state of the technology nor where it is headed. *The Matrix*, a science fiction action film, depicts an extraordinary brain interface; the virtual world created by the interface is so realistic that users are unable to distinguish the virtual from reality, and users are able to gain new knowledge and skills simply from being "plugged in." The *New Culture* novel series, by Iain M. Banks, is somewhat more realistic with nanoscale, self-assembling, and complete-access artificial intelligence (AI) interfacing devices, but the novels' focus on downloading the brain and living forever are not motivating factors for most researchers. The Penfield Mood Organ described in the novel, *Do Androids Dream of Electric Sheep*, by Philip K. Dick, which allows users to manipulate their mood, is more accessible than the technology seen in *The Matrix*. Moreover, science fiction often maligns brain interface technologies by placing them into a dystopian context, a far cry from the reality of this rich and developing field.

BMIs (also called brain-computer interfaces) allow for reading information from the brain such as movement, speech, or memory, or writing information into the brain such as sensory signals, vision, or hearing. Advanced BMI technologies are being used to decode the intention of movement from brain signals, allowing a paralyzed individual to control a robot or computer, or regain the ability to speak. A BMI that writes information into the brain could input sensory information, such as vision or hearing, to restore an individual's sensory deficiencies. While a combination of reading and writing, ideally in a closed loop, could conceivably replace portions of the brain damaged or missing because of stroke or neurodegenerative disease, this is not feasible with current technology.

Electrical interfaces underlie the state-of-the-art in BMI technology. In these interfaces, electrodes are inserted into the brain to record the activity of individual neurons or neural populations, or to control their activity via an applied current. The best-known application of this type of BMI hardware is the cochlear implant, which has been in clinical use since the 1980s. Electrodes are inserted into the cochlea to stimulate the nearby neural tissue, creating the perception of sound. This early success is largely due to the ease with which the cochlea is accessed compared to other areas of the brain, and the relatively straightforward structure-function relationship between the cochlea and the brain's interpretation of sound. The cochlea's structure is tonotopic, meaning the location of a stimulus along the cochlea determines the frequency band of the tone that is perceived by the brain. This tonotopy, and the fact that only about 100 Fourier modes are needed to encode understandable speech, creates ideal conditions for a device to mimic the natural processing of sounds in the brain, but these conditions are not present in other areas of the brain. Thus, the success of the cochlear

Philip Sabes

Starfish Neuroscience, LLC

Philip Sabes is a neuroscientist and neural engineer who has been working on neurotechnology startups for the past four years. Sabes is also Professor Emeritus of Physiology at UC San Francisco. Sabes has undergraduate degrees in physics from Washington University in St. Louis, and in mathematics from Cambridge University (Trinity College) where he was Marshall Scholar. He received his Ph.D. in Michael Jordan's lab at the MIT Department of Brain and Cognitive Sciences, focusing on motor neuroscience and machine learning. He then conducted postdoctoral research with Richard Andersen at Caltech, where Sabes trained as a neurophysiologist. At UC San Francisco, the Sabes Lab used neurophysiological, computational, and behavioral tools to discover how sensory feedback drives movement control. The lab also developed new technologies for brain-machine interfacing (BMI), including the first demonstration of multi-electrode micro-stimulation for real-time sensory feedback during movement. The lab also developed the novel sewing-machine approach to neural interfacing (in collaboration with Michel Maharbiz's lab at UC Berkeley), allowing arrays of micron-scale, thin-film devices to be implanted at individually targeted locations in the brain. In 2017, Sabes retired from UC San Francisco to help launch Neuralink Corp, where he continued work on BMI and the sewing-machine approach. In 2020, he moved to the Seattle area to help create Starfish Neuroscience, LLC, a new startup focused on BMI and neuromodulation.

implant has not generalized to additional, more sophisticated BMI applications reaching the clinical stage.

One example of a more sophisticated application is a motor BMI, whereby a device is used to read the intention of movement from the motor cortex (located at the top of the brain), to drive another external device. The Utah Array, originally developed by Dick Norman at the University of Utah and made commercially available for research by Blackroot Microsystems, has been used experimentally in a number of individuals. The array consists of electrodes in a rigid block-shaped configuration that must be placed percutaneously on top of the head. The array's placement and its hardware's bulky size have limited its use outside of laboratory settings. Nevertheless, the Utah Array has demonstrated success as a motor BMI.

In 2013, Andrew Schwartz and co-workers at the University of Pittsburgh employed a Utah Array to allow a tetraplegic patient to pour a ball from one cup to another with a robotic arm, albeit with imprecise, unnatural movements. The main advancement of the study was in training the individual to use the device to do sequentially more complex tasks, as the techniques used to decode and translate the brain's movement signals were simple. More commonly, motor BMI research focuses on interacting with a computer instead of controlling a robotic prosthesis. A subject may communicate by moving a cursor on a 2-dimensional screen to tap keys, but typing is slow and belabors communication. In a study published in *Nature* earlier this year, Jaimie Henderson, Krishna Shenoy, and co-workers at Stanford improved the pace of communication by instead asking an individual to imagine writing on a pad of paper. Applying a sophisticated machine learning algorithm, the individual's sensory outputs regarding hand movement were translated into one of the 26 letters in the alphabet or a space. This example represents the fastest communication to date using a motor BMI.

Further advancement in the field of motor BMIs to enable more widespread use outside of controlled, laboratory settings will require more naturalistic performance. Improvements in the control and decoding algorithms of BMIs are unlikely to deliver these desired gains in performance, especially when considering the previous examples where activity is constrained to a low dimensional space. Instead, the bottleneck is the BMI hardware. Even state-of-the-art hardware still lacks the sensory feedback the brain receives when performing natural movements, such as the tactile feedback from handling and manipulating small items and the proprioceptive feedback from larger motor movements. Providing a sense of touch to the fingertips is currently achievable, but requires accessing deeper areas of the brain. Because writing information into the brain is more difficult than reading information out, providing richer, more complex feedback through the BMI is not yet possible. Similarly, sensory prostheses for applications, like artificial vision, are not yet feasible because using the large number of pixels needed to create high-resolution images is beyond the bandwidth that BMI devices are able to process. Moreover, the biophysics of the brain can also impede BMI applications; nearby neurons are often stimulated along with the neurons of interest and may obfuscate the neural signals needed for the application at the BMI.

Better neural interfaces are another area for advancement, as the limitations of the Utah Array preclude it from widespread medical and commercial uses. The robotic insertion of very fine electrodes, like the technology developed by Neuralink, allows for a much smaller, encapsulated interface and provides a means to move BMIs from niche experiments to real-world applications. Microelectrodes are embedded along fine flexible threads, which are approximately five microns thick. These threads connect to a miniaturized device that collects, decodes, and compresses data before sending it to a computer via Bluetooth. The state-of-the-art Neuralink 1024 electrode device is so discreet that it is virtually undetectable after its placement on a subject.

Despite technical advances in interface development, the challenge of mimicking the natural activity patterns of the brain still persists. Electrodes are only inserted into a limited area of the brain, but natural brain processes are far more complex, relying on signaling and interactions between multiple areas of the brain. To overcome this challenge, devices will need to become significantly more sophisticated, interfacing with multiple areas of the brain, and users will need to learn how to use the devices, as they would learn to perform a new skill. Progress is slowed by the novelty of the devices and the resulting shortage in both R&D and users. As more private companies begin their own R&D programs on BMIs, devices will be improved iteratively and have greater potential for commercialization.

The second type of interfacing is neuromodulation. In contrast to BMI, neuromodulation aims to manipulate brain activity patterns on a longer timescale, generally with the goal of providing some clinical benefit to an individual. If the pattern of activity in the brain that gives rise to a neurological disease or neuropsychiatric disorder is identified, neuromodulation could potentially be used to block or modify the activity, improving the symptoms of the patient. Commercialized examples include devices that can stop seizures and deep-brain stimulation devices that can stop tremors or unwanted movements in patients with Parkinson's disease. While many applications are currently under investigation, several challenges remain. Neuromodulation devices are invasive and require surgery to place within the brain. The precise location to be targeted within the brain is not necessarily stereotyped across individuals, and some disorders, particularly psychological disorders, arise from activity across multiple areas of the brain. In addition to locating the target within the brain, the hallmark pattern of activity of the condition must be identified and whether this activity should be enhanced or suppressed, either directly or via plasticity. Research into these areas is ongoing, but has been hindered by a lack of closed-loop devices that can perform the manipulation and record the response.

Overall, the long-term prospects of both BMI and neuromodulation are promising. Better devices for both types of interfaces are on the horizon in the coming decade. As these devices gain clinical traction, more applications will be discovered, spurring further research and creating a positive-feedback loop. With more widespread use of these interfaces, larger and better datasets will be generated, allowing artificial intelligence and machine learning to be used at scale to overcome the challenges in the field. At first, devices will be developed piecemeal for specific applications, but will eventually become platforms to tackle multiple applications. Neuromodulation is closer to becoming a broad-use platform because of both its feasibility and the clinical need for such devices, but broad-use BMI will also be developed. ■

Artificial Intelligence and Machine Learning: A National Security Perspective

Lindsey Sheppard

Artificial intelligence (AI) and machine learning (ML) are poised to reshape many aspects of modern life and society, including national security and defense. Lindsey Sheppard restricted her talk to the impacts of AI/ML on the national security enterprise, which includes the Department of Defense (DoD) and other federal agencies. She summarized the historical and policy contexts of the AI ecosystem as they relate to defense applications and described the challenges of implementing AI technology into a large, mission-focused, bureaucratic institution like the DoD.

The continuation and evolution of U.S. security strategy dating back to the Cold War have contributed to the current strategic environment. The Second Offset Strategy of the 1970s and 1980s focused on attaining strategic advantage by offsetting the quantitative edge of U.S.S.R. weapons systems with increased use of computers and other technological improvements to U.S. systems. These technological improvements, which included intelligence, surveillance, reconnaissance platforms, precision-guided weaponry, stealth, and highly networked communications and navigation were widely deployed by the U.S. in Operation Desert Storm and the Global War on Terror. Adversaries soon began investing in similar technologies and means to exploit vulnerabilities in the network-centric approaches used by the U.S. Currently, the U.S. aims to maintain and extend its technological advantage by investing in capabilities that increase standoff distance, response time, and reach. These include the reduced vulnerability of space-based communication systems, the development of cyber and electronic warfare tools, countermeasures, and advanced unmanned systems.

U.S. strategy on emerging technologies, particularly the role of AI in national security, is defined in several recent policy reports. The DoD published its first, and current, AI Strategy in 2018 as an annex to the National Defense Strategy, which outlines the defense priorities of a presidential administration. As part of the DoD strategy, the Joint Artificial Intelligence Center (JAIC) was established to serve as a centralized support center for all AI efforts across the DoD. The JAIC's charter includes enterprise support and governance for the Office of the Secretary of Defense's components and the military departments, including establishing a common foundation for AI development across the organization; training and education to build scientific literacy on AI/ML and digital concepts throughout the DoD; leading the discussion on responsible AI, to include ethics, policy, governance, and responsible use; and cultivating partnerships with academia, industry, and the international community.

DoD Directive 3000.09, Autonomy in Weapon Systems, "establishes DoD policy and assigns responsibilities for the development and use of autonomous and semi-autonomous functions in weapon systems, including manned and unmanned platforms and establishes guidelines designed to minimize the probability and consequences of failures in

Lindsey Sheppard

Center for Strategic and International Studies

Lindsey Sheppard is a fellow with the International Security Program at CSIS, where she focuses on the nexus of emerging technologies and national security for the U.S. and allied and partner nations. Her research areas include AI, ML, autonomous systems, defense innovation policy, and technology ecosystems. Sheppard is a frequent writer and invited speaker on the global state of emerging technology application to the defense and intelligence missions. Sheppard contributes expertise in computational modeling and simulation, system architecture and design, and GPS-denied operations from her prior experience in defense research and development. Before joining CSIS (2018), she was a member of the technical staffs at the Charles Stark Draper Laboratory and the Georgia Tech Research Institute. During this time, she was the programmatic and technical systems engineering lead on various software development projects. Sheppard's work supported U.S. Air Force and U.S. Army procurement and technology development efforts to support operations in contested environments. She holds an M.S. and a B.S. in aerospace engineering from the Georgia Institute of Technology.

autonomous and semi-autonomous weapon systems that could lead to unintended engagements."[1] Each branch of the armed services has expressed interest in fielding systems with various degrees of autonomy and supervisory control, from fully automated to remotely piloted, highlighting the interest within the DoD in robotics, autonomy, and remotely piloted systems.

The Interim National Security Strategic Guidance, released by the Biden administration in March 2021, provides the most current strategic guidance at the national level. As the administration prepares its own National Defense Strategy and National Security Strategy, the interim guidance indicates the high-level national security goals and how technology will be used to achieve these goals. The integration of AI/ML within the DoD sits within the broader modernization and evolution of the department, as legacy analog platforms and generations-old hardware and software are digitized. Investments in cutting-edge technology will require improved and streamlined processes and procedures for the various phases of developing, testing, acquiring, deploying, and providing security, as well as a skilled workforce for technology acquisition, integration, and operation. In addition, the "valley of death" between prototype and production is particularly difficult to overcome in the development of defense systems.

The National Security Commission on AI was established in 2018 to provide recommendations to the president and Congress on advancing the development of AI/ML and related technologies to address U.S. national security and defense needs. The Commission's final report, submitted to Congress in March 2021, described the competition that defines the current global strategic environment, primarily in national security and the economy. Taking a broad view of the AI ecosystem, the report consolidated into a single resource all the current scholarship on AI/ML in the national security policy field, detailing the importance of the workforce in academia and industry, the role of hardware, such as semiconductors and chipsets, and policies and governance.

The future of AI/ML within the DoD, while crucial to solving future national security and defense problems, will introduce challenges beyond those of technology development. The National Security Commission on AI stated that the DoD needs to be "AI-ready" by 2025. The technology is accessible and has myriad potential uses, but the DoD must rethink its current and future approaches to technology to meet this deadline. For example, the current acquisition system, while well-suited for the hardware of the 1970s, 80s, and 90s, such as fighter aircraft, ships, and tanks, is ill-suited for the software-centric technologies that underpin state-of-the-art capabilities.

Consistent decreases in the federal R&D budget have challenged the federal government's ability to innovate and bolster industry innovation. However, the fiscal year 2022 budget request would increase funding for researching, developing, testing, and evaluating emerging technologies, including AI, while the majority of the DoD's budget tracks inflation. In addition, the concentration of technological innovation in industry has created a rift between Silicon Valley and Washington, D.C., between knowledge bases and expertise, and between worldviews that are difficult to overcome. While adopting promising technologies from the commercial sector is challenging, DoD must build a healthy partnership with industry.

The changing nature of the workforce and the centrality of data and data analytics in the Information Age have also strained the DoD's modernization efforts. The DoD must prioritize technical expertise and the benefits of new technologies while addressing the priorities of the workforce of the future. Recruiting, training, and retaining the needed STEM workforce will require revising the ineffective, industrial-age models of hiring, managing talent, and evaluating skills. The Deputy Secretary of Defense's May 2021 Creating Data Advantage memorandum details the department's aim to transform into a data-centric organization, with concrete action items to deliver this goal. To achieve this aim, the DoD must overcome legacy systems, data challenges of storage and access with different levels of security, IT modernization, improved computer network infrastructure, and a different workforce all within a sprawling, bureaucratic enterprise.

The DoD faces significant challenges in modernization that will impact U.S. security and defense for decades to come. AI/ML will play a large role in these modernization efforts even though their current successes are largely from prototypes that have not yet scaled to large organizations like the DoD. Looking ahead, the use of AI/ML in national security and defense applications will have a major impact on geopolitical power dynamics and strategic stability. ■

[1] https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf

Conclusion

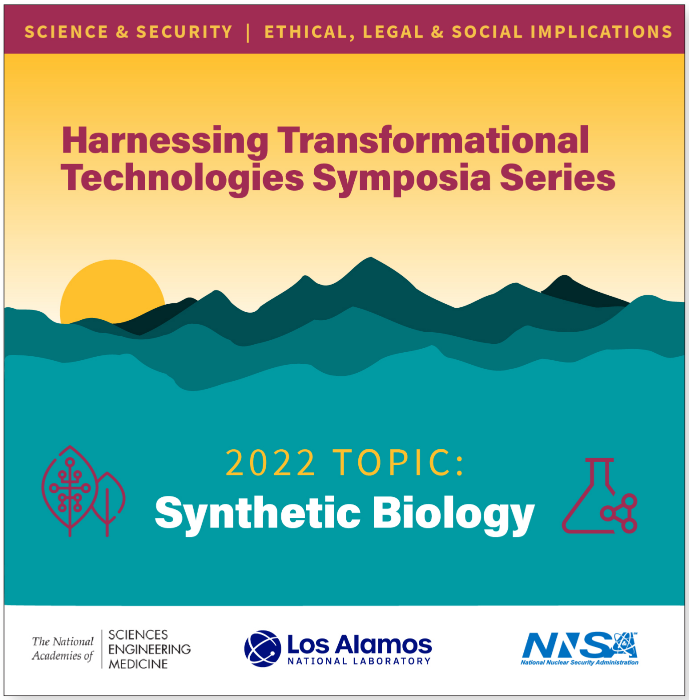
The symposium, “How Artificial Intelligence and Machine Learning Transform the Human Condition,” highlighted the transformative potential of emerging AI and ML technologies on human society. The field of AI/ML has advanced tremendously and swiftly in recent years. However, the rapid pace of technological development has not coincided with the parallel development of a framework to consider the consequences of the technology’s deployment. As a result, we have seen far-reaching negative impacts from AI/ML technology, such as societal manipulation through social media algorithms.

Avoiding negative outcomes from AI/ML technology will require a new, human-centric development paradigm. At the forefront of this paradigm is the focus on stakeholders beyond the developer and the risks of the technology to the much wider ecosystem. With this broader view of impact and consequence, new technology should be optimized from its conception to be inherently beneficial to human society.

Alan Turing articulated the likely future of AI/ML when he stated, “at some point we should have to expect the machines to take control.” Therefore, as we approach this collective reality, we should consider another of Turing’s statements, “we can only see a short distance ahead, but we can see plenty there that needs to be done.” With growing potential for AI/ML misuse and overuse, we have plenty to do to redirect the course of AI/ML development to be human-centric and beneficial to society. Optimistically, the growing consensus among experts, as expressed during the course of the symposium, is that a new AI/ML development paradigm is achievable and offers a glimpse into a future where beneficial AI/ML technology globally improves the human condition. ■

This summary document was prepared by Clay Dillingham, Maksim Eren, Rajan Gupta, Paula Knepper, Monica Lemmon, Lissa Moore, and Courtney Ryan. It is based on the talks given and is approved by the speakers.

Please join us at next year's symposium:


SCIENCE & SECURITY | ETHICAL, LEGAL & SOCIAL IMPLICATIONS
Harnessing Transformational Technologies Symposia Series
2022 TOPIC:
Synthetic Biology
The National Academies of SCIENCES ENGINEERING MEDICINE
Los Alamos NATIONAL LABORATORY
NNSA National Nuclear Security Administration

Los Alamos
NATIONAL LABORATORY

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

NNSA
National Nuclear Security Administration

LA-UR-21-31950 Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is managed by Triad National Security, LLC, for the National Nuclear Security Administration of the U.S. Department of Energy under contract 89233218CNA000001. By acceptance of this article, the publisher recognizes that the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy.