STATEMENT OF:

ERIC HORVITZ
TECHNICAL FELLOW AND DIRECTOR
MICROSOFT RESEARCH—REDMOND LAB
MICROSOFT CORPORATION

BEFORE THE
COMMITTEE ON COMMERCE SUBCOMMITTEE ON SPACE, SCIENCE,
AND COMPETITIVENESS
UNITED STATES SENATE

HEARING ON THE DAWN OF ARTIFICIAL INTELLIGENCE
NOVEMBER 30, 2016

"Reflections on the Status and Future of Artificial Intelligence"

NOVEMBER 30, 2016

Chairman Cruz, Ranking Member Peters, and Members of the Subcommittee, my name is Eric Horvitz, and I am a Technical Fellow and Director of Microsoft's Research Lab in Redmond, Washington. While I am also serving as Co-Chair of a new organization, the Partnership on Artificial Intelligence, I am speaking today in my role at Microsoft.

We appreciate being asked to testify about AI and are committed to working collaboratively with you and other policymakers so that the potential of AI to benefit our country, and to people and society more broadly can be fully realized.

With my testimony, I will first offer a historical perspective of AI, a definition of AI and discuss the inflection point the discipline is currently facing. Second, I will highlight key opportunities using examples in the healthcare and transportation industries. Third, I will identify the important research direction many are taking with AI. Next, I will attempt to identify some of the challenges related to AI and offer my thoughts on how best to address them. Finally, I will offer several recommendations.

What is Artificial Intelligence?

Artificial intelligence (AI) refers to a set of computer science disciplines aimed at the scientific understanding of the mechanisms underlying thought and intelligent behavior and the embodiment of these principles in machines that can deliver value to people and society.

A simple definition of AI, drawn from a 1955 proposal that kicked off the modern field of AI, is pursuing how "to solve the kinds of problems now reserved for humans."[1] The authors of the founding proposal on AI also mentioned, "We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer." While progress has not proceeded as swiftly as the optimistic founders of the field may have expected, there have been ongoing advances over the decades from the sub-disciplines of AI, including *machine vision*, *machine learning*, *natural language understanding*, *reasoning and planning*, and *robotics*.

Highly visible AI achievements, such as DeepBlue's win over the world chess champion, have captured the imagination of the public. Such high-profile achievements have relayed a sense that the field is characterized by large jumps in capabilities. In reality, research and development (R&D) in the AI sub-disciplines have produced an ongoing stream of innovations. Numerous advances have become part of daily life, such as the widespread use of AI route-planning algorithms in navigation systems.[2] Many applications of AI execute "under the hood",

[1] McCarthy, J., Minsky, M.L., Rochester, N., Shannon, C.E. A Proposal for the Dartmouth Summer Project on Artificial Intelligence, Dartmouth University, May 1955.

[2] Hart, P.E., Nilsson, N.J., Raphael, B. A Formal Basis for the Heuristic Determination of Minimum Cost Paths. IEEE Transactions on Systems Science and Cybernetics Vol. SSC-4, No. 2, July 1968.

including methods that perform machine learning and planning to enhance the functioning of computer operating systems or to better retrieve and rank search results. In some cases, AI systems have introduced breakthrough efficiencies without public recognition or fanfare. For example, in the mid- to late-1990s leading-edge machine vision methods for handwriting recognition were pressed into service by the U.S. Postal Service to recognize and route handwritten addresses on letters automatically.[3] High-speed variants of the first machines now sort through more than 25 billion letters per year, with estimated accrued savings of hundreds of millions of dollars.

AI at an Inflection Point

Over the last decade, there has been a promising inflection in the rate of development and fielding of AI applications. The acceleration has been driven by a confluence of several factors. A key influence behind the inflection is the availability of unprecedented streams of data, coupled with drops in the cost of storing and retrieving that data. Large quantities of structured and unstructured databases about human activities and content have become available via the digitization and the shift to the web of activities around commerce, science, communications, governance, education, and art and entertainment.

Other contributing factors include dramatic increases in available computing power, and jumps in the prowess of methods for performing machine learning and reasoning. There has been great activity in the machine learning area over the last thirty years with the development of a tapestry of algorithms for transforming data into components that can recognize patterns, perform diagnoses, and make predictions about future outcomes. The past thirty years of AI research also saw the rise and maturation of methods for representing and reasoning under uncertainty. Such methods jointly represent and manipulate both logical and probabilistic information. These methods draw from and extend methods that had been initially studied and refined in the fields of statistics, operations research, and decision science. Such methods for *learning and reasoning under uncertainty* have been critical for building and fielding AI systems that can grapple effectively with the inescapable incompleteness when immersed in real-world situations.

Over the last decade, there has been a renaissance in the use of a family of methods for machine learning known as *neural networks*.[4] A class of these algorithms referred to as *deep neural networks* are now being harnessed to significantly raise the quality and accuracy of such

[3] Kim, G and Govindaraju, V., Handwritten Word Recognition for Real-Time Applications, Proceedings of the Third International Conference on Document Analysis and Recognition, August 1995.

[4] Neural network algorithms are descendants of statistical learning procedures developed in the 1950s, referred to as *perceptrons*. With neural networks, representations of patterns seen in training data are stored in a set of layers of large numbers of interconnected variables, often referred to as "neurons". The methods are inspired loosely (and in a very high-level manner) by general findings about the layering of neurons in vertebrate brains. Seven years ago, a class of neural networks, referred to as *deep neural networks*, developed decades earlier, were shown to provide surprising accuracies for pattern recognition tasks when trained with sufficient quantities of data.

services as automatic speech recognition, face and object recognition from images and video, and natural language understanding. The methods are also being used to develop new computational capabilities for end users, such as real-time speech-to-speech translation among languages (e.g., now available in Microsoft's Skype) and computer vision for assisting drivers with the piloting of cars (now fielded in the Tesla's models S and X).

Key Opportunities

AI applications explored to date frame opportunities ahead for leveraging current and forthcoming AI technologies. Pressing AI methods that are currently available into service could introduce new efficiencies into workflows and processes, help people with understanding and leveraging the explosion of data in scientific discovery and engineering, as well as assist people with solving a constellation of challenging real-world problems.[5]

Numerous commercial and societal opportunities can be addressed by using available data to build predictive models and then using the predictive models to help guide decisions. Such *data to predictions to decisions* pipelines can deliver great value and help build insights for a broad array of problems.[6] Key opportunities include AI applications in healthcare and biomedicine, transportation, education, manufacturing, and for increasing the effectiveness and robustness of critical infrastructure such as our electrical power grid.

Healthcare and transportation serve as two compelling examples where AI methods can have significant influence in the short- and longer-term.

Healthcare. AI can be viewed as a sleeping giant for healthcare. New efficiencies and quality of care can be obtained by leveraging a coupling of predictive models, decision analysis, and optimization efforts to support decisions and programs in healthcare. Applications span the handling of acute illnesses, longer-term disease management, and the promotion of health and preventative care. AI methods show promise for multiple roles in healthcare, including inferring and alerting about hidden risks of potential adverse outcomes, selectively guiding attention, care, and interventional programs where it is most needed, and reducing errors in hospitals.

On-site machine learning and decision support hinging on inference with predictive models can be used to identify and address potentially costly outcomes. Let's consider the challenge of reducing readmission rates. A 2009 study of Medicare-reimbursed patients who were hospitalized in 2004 found that approximately 20% of these patients were re-hospitalized

[5] Multiple AI applications in support of people and society are presented here: E. Horvitz, AI in Support of People and Society, White House OSTP CCC AAAI meeting on AI and Social Good, Washington DC, June 2016. (access video presentation)

[6] E. Horvitz. From Data to Predictions and Decisions: Enabling Evidence-Based Healthcare, Data Analytic Series, Computing Community Consortium, Computing Research Association (CRA), September 2010.

within 30 days of their discharge from hospitals and that 35% of the patients were re-hospitalized within 90 days.[7] Beyond the implications of such readmissions for health, such re-hospitalizations were estimated to cost the nation $17.4 billion in 2004. Studies have demonstrated that predictive models, learned from large-scale hospital data sets, can be used to identify patients who are at high risk of being re-hospitalized within a short time after they are discharged—and that such methods could be used to guide the allocation of special programs aimed at reducing readmission.[8]

AI methods can also play a major role in reducing costs and enhancing the quality of care for the difficult and ongoing challenge of managing chronic disorders. For example, congestive heart failure (CHF) is prevalent and expensive. The illness affects nearly 10% of people over 65 years. Medical costs and hospitalizations for CHF are estimated to be $35 billion per year in the U.S. CHF patients may hover at the edge of physiological stability and numerous factors can cause patients to spiral down requiring immediate hospitalization. AI methods trained with data can be useful to predict in advance potential challenges ahead and to allocate resources to patient education, sensing, and to proactive interventions that keep patients out of the hospital.

Machine learning, reasoning, and planning offer great promise for addressing the difficult challenge of keeping hospitals safe and efficient. One example is addressing the challenge with hospital-associated infections.[9] It is estimated that such infections affect 10% of people who are hospitalized and that they are a substantial contributor to death in the U.S. Hospital-associated infections have been linked to significant increases in hospitalization time and additional costs of tens of thousands of dollars per patient, and to nearly $7 billion of additional costs annually in the U.S. The CDC has been estimated that 90% of deaths due to hospital-associated infections can be prevented. A key direction is the application of predictive models and decision analyses to estimate patients' risk of illness and to guide surveillance and other preventative actions.

AI methods promise to complement the skills of physicians and create new forms of cognitive "safety nets" to ensure the effective care of hospitalized patients.[10] An Institute of

---

[7] Coleman, E. Jencks, S., Williams, M. Rehospitalizations among Patient in the Medicare Fee-for-Service Program, The New England Journal of Medicine, 380:1418-1428, April 2009.

[8] Bayati, M., Braverman, M., Gillam, M. Mack, K.M., Ruiz, G., Smith, M.S., Horvitz, E. Data-Driven Decisions for Reducing Readmissions for Heart Failure: General Methodology and Case Study. PLOS One Medicine. October 2014.

[9] Wiens, J., Guttag, J. , and Horvitz, E., Patient Risk Stratification with Time-Varying Parameters: A Multitask Learning Approach. Journal of Machine Learning Research (JMLR), April 2016.

[10] Hauskrecht, M., Batal, I., Valko, M., Visweswaran, S., Cooper, G.F., Clermont, G., Outlier Detection for Patient Monitoring and Alerting, Journal of Biomedical Informatics, Volume 46, Issue 1, February 2013, Pages 47–55.

Medicine (IOM) study in 2000 called attention to the problem of preventable errors in hospitals.[11] The study found that nearly 100,000 thousand patients die in hospitals because of preventable human errors. The IOM estimate has been revised upward by several more recent studies. Studies in October 2013 and in May 2016 estimated that preventable errors in hospitals are the third leading cause of death in the U.S., only trailing behind heart disease and cancer. The two studies estimated deaths based in preventable error as exceeding 400,000 and 250,000 patients per year, respectively.[12,13] AI systems for catching errors via reminding and recognizing anomalies in best clinical practices could put a significant dent in the loss of nearly 1,000 citizens per day, and could save tens of thousands of patients per year.

The broad opportunities with the complementarity of AI systems and physicians could be employed in myriad ways in healthcare. For example, recent work in robotic surgery has explored how a robotic surgeon's assistant can work hand-in-hand to collaborate on complex surgical tasks. Other work has demonstrated how coupling machine vision for reviewing histological slides with human pathologists can significantly increase the accuracy of detecting cancer metastases.

Transportation. AI methods have been used widely in online services and applications for helping people with predictions about traffic flows with doing traffic-sensitive routing. Moving forward, AI methods can be harnessed in multiple ways to make driving safer and to expand the effective capacity of our existing roadway infrastructure. Automated cars enabled by advances in perception and robotics promise to enhance both flows on roads and to enhance safety. Longer-range possibilities include the fielding of large-scale automated public *microtransit* solutions on a citywide basis. Such solutions could transform mobility within cities and could influence the overall structure and layout of cities over the longer-term.

Smart, automated driver alerting and assistance systems for collision avoidance show promise for saving hundreds of thousands of lives worldwide. Motor vehicle accidents are believed to be responsible for 1.2 million deaths and 20-50 million non-fatal injuries per year each year. NHTSA's Fatality Analysis Reporting System (FARS) shows that deaths in the U.S. due to motor vehicle injuries have been hovering at rates over 30,000 fatalities per year. In addition to deaths, it is important to include a consideration of the severe injuries linked to transportation. It is estimated that 300,000 to 400,000 people suffer incapacitating injuries every year in motor vehicles; in addition to the nearly 100 deaths per day, nearly one thousand Americans are being incapacitated by motor vehicle injuries every day.

[11] To Err Is Human: Building a Safer Health System, Institute of Medicine: Shaping the Future, November 1999.

[12] James, John T. A New, Evidence-based Estimate of Patient Harms Associated with Hospital Care, Journal of Patient Safety, September 2013.

[13] Daniel, M., Makary, M. Medical Error - The Third Leading Cause of Death in the US, BMJ, 353, 2016.

Core errors based in the distraction of drivers and problems with control lead to road departures and read-end collisions. These expected problems with human drivers could be addressed with machine perception, smart alerting, and autonomous and semi-autonomous controls and compensation. AI methods that deliver inferences with low false-positive and false-negative rates for guiding braking and control could be pressed into service to save many thousands of lives and to avoid hundreds of thousands of life-changing injuries. Studies have found that a great proportion of motor vehicle accidents are caused by distraction and that nearly 20 percent of automobile accidents are believed to be failures to stop. Researchers have estimated that the use of smart warning, assisted braking, and autonomous braking systems could reduce serious injuries associated with rear-end collisions by nearly 50 percent.[14]

Myriad of opportunities. Healthcare and transportation are only two of the many sectors where AI technologies offer exciting advances. For example, machine learning, planning, and decision making can be harnessed to understand, strengthen, monitor, and extend such critical infrastructure as our electrical power grid. In this realm, AI advances could help to address challenges and directions specified in the Energy Independence and Security Act of 2007 on the efficiency, resilience, and security of the U.S. power grid. In particular, there is opportunity to harness predictive models for predicting the load and availability of electrical power over time. Such predictions can lead to more effective plans for power distribution. Probabilistic troubleshooting methodologies can jointly harness knowledge of physical models and streams of data to develop models that could serve in proactive and real-time diagnoses of bottlenecks and failures, with a goal of performing interventions that minimize disruptions.

In another critical sector, AI methods can play an important role in the vitality and effectiveness of education and in continuing-education programs that we offer to citizens. As an example, data-centric analyses have been employed to develop predictive models for student engagement, comprehension, and frustration. Such models can be used in planners that create and update personalized education strategies.[15,16] Such plans could address conceptual bottlenecks and work to motivate and enhance learning. Automated systems could help teachers triage and troubleshoot rising challenges with motivation/engagement and help design ideal mixes of online and human-touch pedagogy.

---

[14] Kusano, K.D. and Gabler, H.C., Safety Benefits of Forward Collision Warning, Brake Assist, and Autonomous Braking Systems in Rear-End Collisions, IEEE Transactions on Intelligent Transportation Systems, pages 1546-1555. Volume: 13(4), December 2012.

[15] K. Koedinger, S. D'Mello., E. McLauglin, Z. Pardos, C. Rose. Data Mining and Education, Wiley Interdisciplinary Reviews: Cognitive Science, 6(4): 333-353, July 2015.

[16] Rollinson, J. and Brunskill, E., From Predictive Models to Instructional Policies, International Educational Data Mining Society, International Conference on Educational Data Mining (EDM) Madrid, Spain, June 26-29, 2015.

Key Research Directions

R&D on AI continues to be exciting and fruitful with many directions and possibilities. Several important research directions include the following:

Supporting Human-AI collaboration. There is great promise for developing AI systems that complement and extend human abilities[17]. Such work includes developing AI systems that are human-aware and that can understand and augment human cognition. Research in this realm includes the development of systems that can recognize and understand the problems that people seek to solve, understanding human plans and intentions, and to recognize and address the cognitive blind spots and biases of people[18]. The latter opportunity can leverage rich results uncovered in over a century of work in cognitive psychology.

Research on human-AI collaboration also includes efforts on the coordination of a mix of initiatives by people and AI systems in solving problems. In such mixed-initiative systems, machines and people take turns at making contributions to solving a problem.[19,20] Advances in this realm can lead to methods tha support humans and machines working together in a seamless, fluid manner.

Recent results have demonstrated that AI systems can learn about and extend peoples' abilities.[21] Research includes studies and methods that endow systems with an understanding about such important subtleties as the cost of an AI system interrupting people in different contexts with potentially valuable information or other contribution[22] and on predicting information that people will forget something that they need to remember in the context at hand.[23]

Causal discovery. Much of machine learning has focused on learning associations rather than causality. Causal knowledge is a critical aspect of scientific discovery and engineering. A longstanding challenge in the AI sub-discipline of machine learning has been identifying causality in an automated manner. There has been progress in this realm over the last twenty

---

17 Licklider, J. C. R., "Man-Computer Symbiosis", IRE Transactions on Human Factors in Electronics, vol. HFE-1, 4-11, March 1960.

18 Presentation: Horvitz, E., Connections, Sustained Achievement Award Lecture, ACM International Conference on Multimodal Interaction (ICMI), Seattle, WA, November 2015.

19 E. Horvitz. Principles of Mixed-Initiative User Interfaces. Proceedings of CHI '99, ACM SIGCHI Conference on Human Factors in Computing Systems, Pittsburgh, PA, May 1999.

20 E. Horvitz. Reflections on Challenges and Promises of Mixed-Initiative Interaction, *AAAI Magazine* 28, Special Issue on Mixed-Initiative Assistants (2007).

21 E. Kamar, S. Hacker, E. Horvitz. Combining Human and Machine Intelligence in Large-scale Crowdsourcing, *AAMAS 2012*, Valencia, Spain, June 2012.

22 E. Horvitz and J. Apacible. Learning and Reasoning about Interruption. *Proceedings of the Fifth ACM International Conference on Multimodal Interfaces*, November 2003, Vancouver, BC, Canada.

23 E. Kamar and E. Horvitz, Jogger: Investigation of Principles of Context-Sensitive Reminding, *Proceedings of International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2011)*, Tapei, May 2011.

years. However, there is much to be done on developing tools to help scientists find rich causal models from large-scale sets of data.[24]

Unsupervised learning. Most machine learning is referred to as *supervised* learning. With supervised learning, data is directly or indirectly tagged by people who provide a learning system with specific labels, such as the goals or intentions of people, or health outcomes. There is deep interest and opportunity ahead with developing *unsupervised learning* methods that can learn without human-authored labels. We are all familiar with the apparent power that toddlers have with learning about the world without obvious detailed tagging or labeling. There is hope that we may one day better understand these kinds of abilities with the goal of harnessing them in our computing systems to learn more efficiently and with less reliance on people.

Learning physical actions in the open world. Research efforts have been underway on the challenges of enabling systems to do active exploration in simulated and real worlds that are aimed at endowing the systems with the ability to make predictions and to perform physical actions successfully. Such work typically involves the creation of training methodologies that enable a system to explore on its own, to perform multiple trials at tasks, and to learn from these experiences. Some of this work leverages methods in AI called *reinforcement learning*, where learning occurs via sets of experiences about the best actions or sequences of actions to take in different settings. Efforts to date include automatically training systems to recognize objects and to learn the best ways to grasp objects.[25]

Integrative intelligence. Many R&D efforts have focused on developing specific competencies in intelligence, such as systems capable of recognizing objects in images, understanding natural language, recognizing speech, and providing decision support in specific healthcare areas to assist pathologists with challenges in histopathology. There is a great opportunity to weave together multiple competencies such as vision and natural language to create new capabilities. For example, natural language and vision have been brought together in systems that can perform automated image captioning.[26,27] Other examples of integrative intelligence involve bringing together speech recognition, natural language understanding, vision, and sets of predictive models to support such challenges as constructing a supportive automated administrative assistant.[28] There is much opportunity ahead in efforts in integrative

---

[24] See efforts at the NIH BD2K Center for Causal Discovery: http://www.ccd.pitt.edu/about/

[25] J. Oberlin, S. Tellex. Learning to Pick Up Objects Through Active Exploration, IEEE, August 2015.

[26] H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. Platt, C. Zitnick, G. Zweig, From Captions to Visual Concepts and Back, CVPR 2015.

[27] Vinyals, O., Toshev, A., Bengio S. Dumitru, E., Show and Tell: A Neural Image Caption Generator, CVPR 2015.

[28] D. Bohus and E. Horvitz. Dialog in the Open World: Platform and Applications, *ICMI-MLMI 2009: International Conference on Multimodal Interaction*, Cambridge, MA. November, 2009.

intelligence that seek to weave together multiple AI competencies into greater wholes that can perform rich tasks.

Advances in platform and systems. Specific needs for advances with data-center scale systems and innovative hardware have come to the fore to support the training and execution of largescale neural network models. New research at the intersection of learning and reasoning algorithms, computing hardware, and systems software will likely be beneficial in supporting AI innovations. Such research is being fielded in platforms that are becoming available from large companies in the technology sector.

Development tools and "democratization of AI". New types of development tools and platforms can greatly assist with development, debugging, and fielding of AI applications. R&D is ongoing at large IT companies on providing developers with cloud-based programmatic interfaces (e.g., Microsoft's Cognitive Services) and client-based components for performing valuable inference tasks (e.g., detect emotion in images). Also, learning toolkits are being developed that enable researchers and engineers to do machine learning investigations and to field classifiers (e.g., Microsoft's CNTK and Google's TensorFlow). Other development environments are being developed for creating integrative AI solutions that can be used by engineers to assemble systems that rely on the integration of multiple competencies (natural language understanding, speech recognition, vision, reasoning about intentions of people, etc.) that must work together in a tightly coordinated manner in real-time applications.

Challenges

Economics and jobs. Over the last several years, the AI competencies with seeing, hearing, and understanding language have grown significantly. These growing abilities will lead to the fielding of more sophisticated applications that can address tasks that people have traditionally performed. Thus, AI systems will likely have significant influences on jobs and the economy. Few dispute the assertion that AI advances will increase production efficiencies and create new wealth. McKinsey & Company has estimated that advanced digital capabilities could add 2.2 trillion U.S. dollars to the U.S. GDP by 2025. There are rising questions about how the fruits of AI productivity will distributed and on the influence of AI on jobs. Increases in the competencies of AI systems in both the cognitive and physical realms will have influences on the distribution, availability, attraction, and salaries associated with different jobs. We need to focus attention on reflection, planning, and monitoring to address the potential disruptive influences of AI on jobs in the US—and to work to understand the broad implications of new forms of automation provided by AI for domestic and international economics. Important directions for study include seeking an understanding of the needs and value of education and the geographic distribution of rising and falling job opportunities.

There is an urgent need for training and re-training of the U.S. workforce so as to be ready for expected shifts in workforce needs and in the shifts in distributions of jobs that are

fulfilling and rewarding to workers. In an economy increasingly driven by advances in digital technology, increasing numbers of jobs are requiring a degree in one of the STEM (science, technology, engineering, and math) fields. There is growing demand for people with training in computer science, with estimates suggesting that by 2024, the number of computer and information analyst jobs will increase by almost 20 percent. For companies to thrive in the digital, cloud-driven economy, the skills of employees must keep pace with advances in technology. It has been estimated as many as 2 million jobs could go unfilled *in the US manufacturing sector* during the next decade because of a shortage of people with the right technical skills.[29] Investing in education can help to prepare and adapt our workforce to what we expect will be a continuing shift in the distribution of jobs, and for the changing demands on human labor.

Beyond ensuring that people are trained to take on fulfilling, well-paid positions, providing STEM education and training to larger number of citizens will be critical for US competitiveness. We are already facing deficits in our workforce: The Bureau of Labor Statistics estimates that there are currently over 5 million unfilled positions in the U.S. Many of those jobs are those created due to new technologies. This suggests that there are tremendous opportunities for people with the right skills to help US companies to create products and services that can, in turn, drive additional job creation and create further economic growth.

Shortages of people who have training in sets of skills that are becoming increasingly relevant and important could pose serious competitive issues for companies and such shortages threaten the long-term economic health of the US. Without addressing the gap in skills, we'll likely see a widening of the income gap between those who have the skills to succeed in the 21st century and those who do not. Failing to address this gap will leave many people facing an uncertain future—particularly women, young people, and those in rural and underserved communities. Working to close this divide will be an important step to addressing income inequality and is one of the most important actions we can take.

Safety and robustness in the open world. Efforts to employ AI systems in high-stakes, safety critical applications will become more common with the rising competency of AI technologies.[30] Such applications include automated and semi-automated cars and trucks, surgical assistants, automation of commercial air transport, and military operations and weapon systems, including uses in defensive and offensive systems. Work is underway on ensuring that systems in safety critical areas perform robustly and in accordance with human preferences. Efforts on safety and robustness will require careful, methodical studies that address the multiple ways that learning and reasoning systems may perform costly, unintended

[29] The Manufacturing Institute and Deloitte, "The skills gap in U.S. manufacturing: 2015 and beyond.", 2015.

[30] T.G. Dietterich and E.J. Horvitz, Rise of Concerns about AI: Reflections and Directions. Communications of the ACM, Vol. 58 No. 10, pages 38-40, October 2015.

actions.[31] Costly outcomes can result from erroneous behaviors stemming from attacks on one or more components of AI systems by malevolent actors. Other concerns involve problems associated with actions that are not considered by the system. Fears have also been expressed that smart systems might be able to make modifications and to shift their own operating parameters and machinery. These classes of concern frame directions for R&D.

Efforts and directions on safety and robustness include the use of techniques in computer science referred to as *verification* that prove constraints on behaviors, based on offline analyses or on real-time monitoring. Other methods leverage and extend results developed in the realm of adaptive control, on robust monitoring and control of complex systems. Control-theoretic methods can be extended with models of sensor error and with machine learning about the environment to provide guarantees of safe operation, given that assumptions and learnings about the world hold.[32] Such methods can provide assurance of safe operation at a specified tolerated probability of failure. There are also opportunities for enhancing the robustness of AI systems by leveraging principles of *failsafe design* developed in other areas of engineering.[33] Research is also underway on methods for building systems that are robust to incompleteness in their models, and that can respond appropriately to *unknown* unknowns faced in the open world.[34] Beyond research, best practices may be needed on effective testing, structuring of trials, and reporting when fielding new technologies in the open world.

A related, important area for R&D on safety critical AI applications centers on the unique challenges that can arise in systems that are jointly controlled by people and machines. Opportunities include developing systems that explicitly consider human attention and intentions, that provide people with explanations of machine inferences and actions, and that work to ensure that people comprehend the state of problem solving—especially as control is passed between machines and human decision making. There is an opportunity to develop and share best practices on how systems signal and communicate with humans in settings of shared responsibility.

Ethics of autonomous decisions. Systems that make autonomous decisions in the world may have to make trades and deliberate about the costs and benefits of rich, multidimensional outcomes—under uncertainty. For example, it is feasible that an automated driving system may

---

[31] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D., Concrete Problems in AI Safety, arXiv, 25 July 2016.

[32] D. Sadigh, A. Kapoor, *Safe Control Under Uncertainty with Probabiliistic Signal Temporal Logic*, Robotics: Science and Systems, RSS 2016.

[33] Overview presentation on safety and control of AI can be found here: E. Horvitz, Reflections on Safety and Artificial Intelligence, White House OSTP-CMU Meeting on Safety and Control of AI, June 2016. (view video presentation).

[34] One challenge that must be considered when fielding applications for safety critical tasks is with transfer of applications from the closed world of test scenarios to the open world of fielded technologies. Systems developed in the laboratory or in test facilities can be surprised by unexpected situations in the open world—a world that contains unmodeled situations, including sets of *unknown unknowns* stemming from incompleteness in a system's perceptions and understandings. Addressing incompleteness and unknown unknowns is an interesting AI research challenge.

have to reason about actions that differentially influence the likelihood that passengers versus pedestrians are injured. As systems become more competent and are granted greater autonomy in different areas, it is important that the values that guide their decisions are aligned with the values of people and with greater society. Research is underway on the representation, learning, transparency, and specification of values and tradeoffs in autonomous and semi-autonomous systems.

Fairness, bias, transparency. There is a growing community of researchers with interest in identifying and addressing potential problems with fairness and bias in AI systems.[35] Datasets and the classifications or predictions made by systems constructed from the data can be biased. Implicit biases in data and in systems can arise because of unmodeled or poorly understood limitations or constraints on the process of collection of data, the shifting of the validity of data as it ages, and using systems for inferences and decisions for populations or situations that differ greatly from the populations and situations that provided the training data As an example, predictive models have been used to assist with decision making in the realm of criminal justice. Models trained on data sets have been used to assist judges with decisions about bail and about the release of people charged with crimes in advance of their court dates. Such decisions can enhance the lives of people and reduce costs. However, great caution must be used with ensuring that datasets do not encode and amplify potential systematic biases in the way the data is defined and collected. Research on fairness, biases, and accountability and the performance of machine-learned models for different constituencies is critically important. The importance of this area will only grow in importance as AI methods are used with increasing frequency to advise decision makers about the best actions in high-stakes settings. Such work may lead to best practices on the collection, usage, and the sharing of datasets for testing, inspection, and experimentation. Transparency and openness may be especially important in applications in governance.

Manipulation. It is feasible that methods employing machine learning, planning, and optimization could be used to create systems that work to influence peoples' beliefs and behavior. Further, such systems could be designed to operate in manner that is undetectable by those being influenced. More work needs to be done to study, detect, and monitor such activity.

Privacy. With the rise of the centrality of data-centric analyses and predictive models come concerns about privacy. We need to consider the potential invasion in the privacy of individuals based on inferences that can be made from seemingly innocuous data. Other efforts on privacy include methods that allow data to be used for machine learning and reasoning yet maintains the privacy of individuals. Approaches include methods for

[35] See Fairness, Accountability, and Transparency in Machine Learning (FATML) conference site: http://www.fatml.org/

anonymizing data via injecting noise[36], sharing only certain kinds of summarizing statistics, providing people with controls that enable them to trade off the sharing of data for enhanced personalization of services[37], and using different forms of encryption. There is much work to be done on providing controls and awareness to people about the data being shared and how it is being used to enhance services for themselves and for larger communities.

Cybersecurity. New kinds of automation can present new kinds of "attack surfaces" that provide opportunities for manipulation and disruption by cyberattacks by state and non-state actors. As mentioned above, it is critical to do extensive analyses of the new attack surfaces and the associated vulnerabilities that come with new applications of AI. New classes of attack are also feasible, including "machine learning attacks," involving the injection of erroneous or biased training data into data sets. Important directions include hardware and software-based security and encryption, new forms of health monitoring, and reliance on principles of failsafe design.

Recommendations

We recommend the following to catalyze innovation among our basic and applied AI communities across government, academia, industry, and non-profit sectors:

- Public-sector research investments are vital for catalyzing innovation on AI principles, applications, and tools. Such funding can leverage opportunities for collaborating and coordinating with industry and other sectors to help facilitate innovation.

- Research investments are needed at intersections of AI with law, policy, psychology, economics and ethics to better understand and monitor the social and societal consequences of AI.

- Governments should create frameworks that enable citizens and researchers to have easy access to government curated datasets where appropriate, taking into consideration privacy and security concerns.

- With the goal of developing guidelines and best practices, governments, industry, and civil society should work together to weigh the range of ethical questions and issues that AI applications raise in different sectors. As experience with AI broadens, it may make sense to establish more formal industry standards that reflect consensus about ethical issues but that do not impede innovation and progress with AI and its application in support of people and society.

---

[36] Differential privacy ref: Dwork, C.: Differential Privacy. In: Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (ICALP) (2), pp. 1–12 (2006).

[37] A. Krause and E. Horvitz. A Utility-theoretic Approach to Privacy in Online Services, *Journal of Artificial Intelligence Research*, 39 (2010) 633-662.

- In an era of increasing data collection and use, privacy protection is more important than ever. To foster advances in AI that benefit society, policy frameworks must protect privacy without limiting innovation. For example, governments should encourage the exploration and development of techniques that enable analysis of large data sets without revealing individual identities.

- We need to invest in training that prepares people for high-demand STEM jobs. Governments should also invest in high-quality worker retraining programs for basic skills and for certifications and ongoing education for those already in the workforce. A first step is to identify the skills that are most in demand. Governments can develop and deliver high-quality workforce retraining programs or provide incentives and financial resources for private and nonprofit organizations to do so.

Thank you for the opportunity to testify. I look forward to answering your questions.