One-Hundred Year Study on Artificial Intelligence: Reflections and Framing

ERIC HORVITZ | 2014

The One Hundred Year Study on Artificial Intelligence (AI) has its roots in a one-year study on **Long-term AI Futures** that we commissioned during my term of service as president of the Association for the Advancement of Artificial Intelligence (AAAI) in 2008-2009. The presidential panel was divided into three groups, with one team exploring opportunities and concerns arising in the near term, another team exploring more speculative long-term outcomes, and a third team taking a special focus on legal and ethical challenges that might come to the fore with the construction and fielding of computational systems with one or more abilities to perceive, learn, reason, and act in the world. The panel was asked to reflect about the opportunities ahead, as well as to consider uncertainties and potential "rough edges" that might come to the fore, and to highlight topics that would require additional study. For potential concerns, the group was asked to reflect about the value of proactive efforts to mitigate concerns and to ensure good outcomes.

At a final gathering at the Asilomar conference center in Pacific Grove, CA on February 2009, panel participants overall expressed optimism about the value that AI systems and methods would continue to deliver to people and society. Per the charter of the panel, participants in each group identified potential concerns and directions for additional examination and scholarship.

Many of the opportunities and concerns discussed over the course of the AAAI study would not be easy to address in a single setting. My sense was that it would be valuable to create a long-term program of study on the influences of AI on people and society—one that provide a long gaze into the future and that would continue to leverage and extend a rich archive. A program extending over a hundred years could provide an enduring platform for scholarship, observation, and proactive guidance.

Such a study would not only focus on concerns with potential rough edges and costs that may rise with the advance of AI. Rather, the study would work more broadly to identify where more investigation, effort, and investment might be needed across a wide range of opportunities and concerns. For example, many promising AI technologies and prototypes have been developed with clear demonstrated value, yet have not yet been translated into real-world usage in such critical areas as healthcare, education, and transportation. Delays with the wide-scale fielding of such solutions translate into significant monetary costs and loss of life. Such problems with delays and missed

opportunities for fielding valuable advances in AI will likely continue to outweigh concerns that some have voiced about applications and influences of AI. Observation and reflection should address opportunities and concerns over the short- and longer-terms, including speculation by some about the possibility that powerful AI systems may one day be difficult to control or to have act in accordance with human wishes.

I converged on a university as providing the most stable and nurturing host for an extended study. I approached colleagues at Stanford University with a proposal for a one-hundred year program in early 2014. After a great deal of discussion and reflection, the proposal was accepted the following June. Stanford University has been a leading center of research in AI as part of its excellence in computer science and electrical engineering. The university has also been strong in important related areas, including the closely affiliated domains of decision science, operations research, and biomedical informatics. Stanford also has very strong programs in ethics, philosophy, psychology, law, medicine, and neuroscience—areas of study that will be important in studies of the long-term futures of AI.

The *One Hundred Year Study on Artificial Intelligence* (AI100) at Stanford will be overseen by a Standing Committee with rotating membership. A Faculty Director will provide administrative and programmatic oversight. The Standing Committee is charged with nurturing and extending the mission of the program. A key responsibility of the Standing Committee will be to assemble and track studies undertaken at five-year intervals. The Standing Committee will recruit a Study Panel, which will work over a year's time to assess developments with the advances and influences of AI on people and society, and provide assessments and recommendations in a final report and public presentation. Such a report will include reflections and guidance on scientific, engineering, legal, ethical, economic, and societal fronts. The Standing Committee may work to focus and frame the direction of studies with terms of reference. When reports are completed, the Standing Committee will work on the effective communication of findings to relevant organizations and people, and be available for ongoing discussions on implications. A digital archive at Stanford of reports, deliberations, and related materials will be made available to the public. Beyond the studies, the Standing Committee will have a continuing presence and will serve a role in ongoing communications and assessments.

The Standing Committee and Study Panel will work to formulate topics of attention and the best ways to approach them. A sampling of interrelated topics of interest includes the following:

Technical trends and surprises. What can we expect in terms of the advancing competencies of technology in the near-term and at more distant points in time? What might surprise us on the technical front? How well have we been able to predict advances of AI in the past?

Key opportunities for AI. How can AI advances and implementations help transform the quality of such critical areas as education, healthcare, science, government, and the overall vitality of society? Where might AI be most useful? What problems and bottlenecks are ripe for being solved or addressed with computational systems that can perceive, learn, reason, and plan? What major positive changes in the world will come with advances in AI systems? Where might AI systems unlock great value by disrupting the status quo in significant ways?

Delays with translating AI advances into real-world value. Numerous advances in AI can reduce costs, introduce new efficiencies, and raise the quality of life. For example, machine learning and inference can play a significant role with reducing costs and enhancing the quality of healthcare. However, the methods have not come into wide use. The sluggish translation of these technologies into the world translates into unnecessary deaths and costs. There is an urgent need to better understand how we can more quickly translate valuable existing AI competencies and advances into real-world practice.

Privacy and machine intelligence. What potential challenges to privacy might come to the fore with advances in AI research and development, including efforts in machine learning, pattern recognition, inference, and prediction? What are the implications for privacy of systems that can make inferences about the goals, intentions, identity, location, health, beliefs, preferences, habits, weaknesses, and future actions and activities of people? What are the preferences and levels of comfort of people about machines performing such inferences? How might people be protected from unwanted inferences or uses of such inferences? Are there opportunities for innovation with new forms of insightful (yet lightweight?) regulatory guidance, policies, or laws, including approaches to disclosure about inferences?

Democracy and freedom. Machine learning and inference have been employed to influence the beliefs and actions of people. Such methods have been recently reported as used to influence the outcome of campaigns for political offices in the U.S. What are potential threats to democracy and freedom of thought in a world where systems can be trained, optimized, and applied with the goal of persuading people to believe differently

and to vote differently, especially when such persuasion can be designed to operate in a stealthy manner?

Law. Advances in AI methods may have numerous implications for laws and regulation. What aspects of common, statutory, and regulatory law may need to be revised in light of advances in the power and applications of AI systems? For example, the 2009 AAAI study called out potential challenges with uses of liability law with regard to the actions of autonomous systems or semi-autonomous systems. Will we need to re-think or extend current notions of liability when automated reasoning or robotic systems take on new roles and autonomy?

Ethics. What ethical challenges and questions might come to the fore with advances in the competencies and uses of AI systems for inferences and robotic actions in the world? How might advances in AI frame new questions in ethics? What uses of AI may be considered unethical? What ethical questions might surface with the common use of new kinds of autonomous decision systems in such high-stakes areas as healthcare and transportation? What ethical issues may arise with systems that display human-like qualities and competencies? Some questions may address subtle yet important issues. For example, might systems that sound like people during spoken dialog on the telephone have to disclaim that they are automated?

Economics. What are the economic and societal implications more broadly of automated reasoning and robotic systems that take on increasingly sophisticated jobs, replacing or shifting the distributions and nature of human work? What are the influences and implications for the distribution of human work and professional aspirations of people? How might the nature of work be shifted by new kinds of platforms that combine machine intelligence and human effort? Turning to the fabric of economic systems, how might machine intelligence influence the design and operation of markets? What are opportunities and consequences ahead with uses of automated mechanism design and the fielding of new procedures for determining winners? What are dangers with uses of algorithmic, high-speed trading and analysis for financial markets?

AI and warfare. AI methods have been employed for decades in weapon systems that provide advice to human decision makers, perform targeting and navigation, and act with different levels of autonomy in different situations. What are the implications of harnessing learning, reasoning, pattern recognition, and autonomous decision making in weapon systems? What new kinds of autonomy in targeting and decision making might be developed on all sides, per competitive pressures of innovation? What are

opportunities and challenges with developing new frameworks and conventions for defining policies about autonomous decision making in weaponry and warfare.

Criminal uses of AI. The 2009 AAAI study warned of the potential rise of new forms of malware and its uses in new forms of "criminal AI," including systems constructed by state and non-state actors that perform intelligent persistent threats by gaining access to data and making inferences to peoples' location, intentions, and activities. Today, people often interact and provide data to multiple devices and services. One can imagine intelligent malware that works cross-service and cross-device to make deep inferences about people over time, to perform experiments to hone skills, and then perform stealthy criminal activities, including financial attacks. What new kinds of malevolent software might be developed and how might it be used? What might be done proactively to thwart such criminal AI?

Collaborations with machines. What challenges may come to the fore as collaborations between people and intelligent machines on tasks becomes more commonplace? The need for an interleaving or transfer of control between people and machines is a key consideration in the joint, computer-human operation of systems like cars, planes, and surgical tools. What are opportunities for developing means for allowing an efficient interleaving of contributions from people and machines on a task, or in the transfer of control between people and machines? What are opportunities for mixed-control systems that rely on guidance from both people and AI systems? Troubling scenarios have come to the fore even with some traditional control systems when the baton of control is passed between automated solutions and people (e.g., China Airlines flight 006). There are great opportunities in this realm for developing, testing, and deploying methods that enable people and machines to work together in a fluid manner, each contributing to the overall operation of a system or to the solution of problems. Work in this realm includes developing methods that endow machines with the skill to explain their reasoning in an understandable manner, so that people can understand their inferential steps, conclusions, and recommendations.

AI and human cognition. There are great opportunities to develop computing systems that augment human abilities by better understanding the competencies and weaknesses of human cognition. Cognitive psychologists have worked over a century on characterizations of the human mind as bounded reasoning system, with well-characterized competencies, biases, and blind spots. Machines could help to augment human cognition by understanding such human weaknesses and biases—in health and illness. For example, personalized systems could help to remind people of things that

they are likely to forget, to make better decisions, and to have better focus of attention when they are likely to be distracted.

Safety and autonomy. What are key technical opportunities with the specification and verification of ranges of desired or safe behaviors of autonomous systems? What methods might be used to ensure expected and appropriate behavior, even when autonomous systems encounter new or unforeseen situations in the open world? Could systems understand when they do not know enough to act in the world and to reach out to people for assistance in understanding situations, goals, and appropriate actions? This important topic area was studied and discussed at the 2009 AAAI study, including review of work to date on methods that might be taken as modern and more formal versions of Asimov's Laws of Robotics. There is much to be done with leveraging logical theorem proving, decision theory, and program verification to develop new approaches to safe behaviors. Advances would have implications for concerns about the possibility of the loss of control of intelligent machines.

Loss of control of AI systems. Concerns have been expressed about the possibility of the loss of control of intelligent machines for over a hundred years. The loss of control of AI systems has been a recurrent theme in science fiction and has been raised as a possibility by scientists. Some have speculated that we could one day lose control of AI systems via the rise of superintelligences that do not act in accordance with human wishes—and that such powerful systems would threaten humanity. Speculations about the rise of such uncontrollable machine intelligences have called out different scenarios, including trajectories that take a slow, insidious course of refinement and faster-paced evolution of systems toward a powerful intelligence "singularity." Are such dystopic outcomes possible? If so, how might these situations arise? What are the paths to these feared outcomes? What might we do proactively to effectively address or lower the likelihood of such outcomes, and thus reduce these concerns? What kind of research would help us to better understand and to address concerns about the rise of a dangerous superintelligence or the occurrence of an "intelligence explosion"? Concerns about the loss of control of AI systems should be addressed via study, dialog, and communication. Anxieties need to be addressed even if they are unwarranted. Studies could reduce anxieties by showing the infeasibility of concerning outcomes or by providing guidance on proactive efforts and policies to reduce the likelihood of the feared outcomes.

Psychology of people and smart machines. We don't have a good understanding about how people will feel about and respond to intelligent applications, services, and robots in their environments and lives. What are the psychological implications of

implementations of intelligence in our various environments? For example, how will our self-conception—our sense of who we are—change with the rise of machines with human-like qualities and competencies? What is the evolving nature of relationships with intelligent software and systems? What is the basis for anxieties about machines over the centuries, and how might anxieties about smart machines stimulate fear and discomfort? What is the basis for fears about “superintelligences”? How do popular media and stories contribute to these fears? We can work to better understand and study potential dependencies, fears, anxieties, and attractions. Psychological studies could be undertaken in a proactive manner (e.g., with “Wizard of Oz” methods that simulate future intelligences) in advance of the availability of such machines.

Communication, understanding, and outreach. What do computer-science non-experts, including experts in other fields, and people in different spheres of life (e.g., political leaders) understand about key issues and developments with AI? How might anxiety (that some consider premature or erroneous) about the loss of control by machines lead to a slow down or shutdown of important AI research that might otherwise better humankind? What are the multiple risks of poor understanding? What education and communication programs would be valuable for informing non-experts about the status of AI research and capabilities—and the challenges and opportunities ahead?

Neuroscience and AI. Despite many decades of research, we have little genuine knowledge about how the human brain performs its computational magic. While people have been using phrases like “brain-inspired” (e.g., with regard to neural network models, including network models referred to as “deep learning”), we have a surprisingly poor understanding about the operation of the nervous system, and the computational methods that it may employ. Allusions to the brain are typically made with a stretch of the imagination—and with confidence powered largely by intuitions. There is a rich opportunity for better understanding the impressive capabilities of the brain via computational models, methods, metaphors, and results. Going in the other direction, findings about the structure and operation of nervous systems can help AI scientists to formulate questions and to guide research about computational approaches to thought and intelligent behavior. Pushing hard on both directions will likely be important, and there is much to be learned.

AI and philosophy of mind. Philosophers have long grappled with interesting and compelling questions about the nature of mind, including questions about the nature and foundations of conscious experience. These discussions often include careful development of a language to bring discussions to a common ground. Questions have

been asked about whether machines that we build might one day be conscious and find themselves “aware” and “experiencing” inner or subjective worlds similar to those experienced by people. Discussions about the computational foundations of consciousness will likely rely on insights from AI, neuroscience, and philosophy