DOI:10.1145/2770869

Thomas G. Dietterich and Eric J. Horvitz

Viewpoint
Rise of Concerns about AI: Reflections and Directions

Research, leadership, and communication about AI futures.

Discussions about artificial intelligence (AI) have jumped into the public eye over the past year, with several luminaries speaking about the threat of AI to the future of humanity. Over the last several decades, AI—automated perception, learning, reasoning, and decision making—has become commonplace in our lives. We plan trips using GPS systems that rely on the A* algorithm to optimize the route. Our smartphones understand our speech, and Siri, Cortana, and Google Now are getting better at understanding our intentions. Machine vision detects faces as we take pictures with our phones and recognizes the faces of individual people when we post those pictures to Facebook. Internet search engines rely on a fabric of AI subsystems. On any day, AI provides hundreds of millions of people with search results, traffic predictions, and recommendations about books and movies. AI translates among languages in real time and speeds up the operation of our laptops by guessing what we will do next. Several companies are working on cars that can drive themselves—either with partial human oversight or entirely autonomously. Beyond the influences in our daily lives, AI techniques are playing roles in science and medicine. AI is already at work in some hospitals helping physicians understand which patients are at highest risk for complications, and AI algorithms are finding important needles in massive data haystacks, such as identifying rare but devastating side effects of medications.

AI has been in the headlines with such notable advances as self-driving vehicles, now under development at several companies; Google's self-driving car is shown here.

IMAGE COURTESY OF GOOGLE.COM/SELFDRIVINGCAR/

The AI in our lives today provides a small glimpse of more profound contributions to come. For example, the fielding of currently available technologies could save many thousands of lives, including those lost to accidents on our roadways and to errors made in medicine. Over the longer-term, advances in machine intelligence will have deeply beneficial influences on healthcare, education, transportation, commerce, and the overall march of science. Beyond the creation of new applications and services, the pursuit of insights about the computational

foundations of intelligence promises to reveal new principles about cognition that can help provide answers to longstanding questions in neurobiology, psychology, and philosophy.

On the research front, we have been making slow, yet steady progress on "wedges" of intelligence, including work in machine learning, speech recognition, language understanding, computer vision, search, optimization, and planning. However, we have made surprisingly little progress to date on building the kinds of general intelligence that experts and the lay public envision when they think about "Artificial Intelligence." Nonetheless, advances in AI—and the prospect of new AI-based autonomous systems—have stimulated thinking about the potential risks associated with AI.

A number of prominent people, mostly from outside of computer science, have shared their concerns that AI systems could threaten the survival of humanity.[1] Some have raised concerns that machines will become superintelligent and thus be difficult to control. Several of these speculations envision an "intelligence chain reaction," in which an AI system is charged with the task of recursively designing progressively more intelligent versions of itself and this produces an "intelligence explosion."[4] While formal work has not been undertaken to deeply explore this possibility, such a process runs counter to our current understandings of the limitations that computational complexity places on algorithms for learning and reasoning. However, processes of self-design and optimization might still lead to significant jumps in competencies.

Other scenarios can be imagined in which an autonomous computer system is given access to potentially dangerous resources (for example, devices capable of synthesizing billons of biologically active molecules, major portions of world financial markets, large weapons systems, or generalized task markets[9]). The reliance on any computing systems for control in these areas is fraught with risk, but an autonomous system operating without careful human oversight and failsafe mechanisms could be especially dangerous. Such a system would not need to be particularly intelligent to pose risks.

The AI in our lives today provides a small glimpse of more profound contributions to come.

We believe computer scientists must continue to investigate and address concerns about the possibilities of the loss of control of machine intelligence via any pathway, even if we judge the risks to be very small and far in the future. More importantly, we urge the computer science research community to focus intensively on a second class of near-term challenges for AI. These risks are becoming salient as our society comes to rely on autonomous or semiautonomous computer systems to make high-stakes decisions. In particular, we call out five classes of risk: bugs, cybersecurity, the "Sorcerer's Apprentice," shared autonomy, and socioeconomic impacts.

The first set of risks stems from programming errors in AI software. We are all familiar with errors in ordinary software; bugs frequently arise in the development and fielding of software applications and services. Some software errors have been linked to extremely costly outcomes and deaths. The verification of software systems is challenging and critical, and much progress has been made—some relying on AI advances in theorem proving. Many non-AI software systems have been developed and validated to achieve high degrees of quality assurance. For example, the software in autopilot and spacecraft systems is carefully tested and validated. Similar practices must be applied to AI systems. One technical challenge is to guarantee that systems built via machine learning methods behave properly. Another challenge is to ensure good behavior when an AI system encounters unforeseen situations. Our automated vehicles, home robots, and intelligent cloud services must perform well even when they receive surprising or confusing inputs. Achieving such robustness may require self-monitoring architectures in which a meta-level process continually observes the actions of the system, checks that its behavior is consistent with the core intentions of the designer, and intervenes or alerts if problems are identified. Research on real-time verification and monitoring of systems is already exploring such layers of reflection, and these methods could be employed to ensure the safe operation of autonomous systems.[3,6]

A second set of risks is cyberattacks: criminals and adversaries are continually attacking our computers with viruses and other forms of malware. AI algorithms are as vulnerable as any other software to cyberattack. As we roll out AI systems, we need to consider the new attack surfaces that these expose. For example, by manipulating training data or preferences and trade-offs encoded in utility models, adversaries could alter the behavior of these systems. We need to consider the implications of cyberattacks on AI systems, especially when AI methods are charged with making high-stakes decisions. U.S. funding agencies and corporations are supporting a wide range of cybersecurity research projects, and artificial intelligence techniques will themselves provide novel methods for detecting and defending against cyberattacks. For example, machine learning can be employed to learn the fingerprints of malware, and new layers of reflection can be employed to detect abnormal internal behaviors, which can reveal cyberattacks. Before we put AI algorithms in control of high-stakes decisions, we must be confident these systems can survive large-scale cyberattacks.

A third set of risks echo the tale of the Sorcerer's Apprentice. Suppose we tell a self-driving car to "get us to the airport as quickly as possible!" Would the autonomous driving system put the pedal to the metal and drive at 125 mph, putting pedestrians and other drivers at risk? Troubling scenarios of this form have appeared recently in the press. Many of the dystopian scenarios of out-of-control superintelligences are variations on this theme. All of these examples refer to cases where humans have failed to correctly instruct the AI system on how it should behave. This is not a new problem. An important aspect of any AI system that interacts with people

is that it must reason about what people *intend* rather than carrying out commands literally. An AI system must analyze and understand whether the behavior that a human is requesting is likely to be judged as "normal" or "reasonable" by most people. In addition to relying on internal mechanisms to ensure proper behavior, AI systems need to have the capability—and responsibility—of working with people to obtain feedback and guidance. They must know when to stop and "ask for directions"—and always be open for feedback.

Some of the most exciting opportunities for deploying AI bring together the complementary talents of people and computers.[5] AI-enabled devices are allowing the blind to see, the deaf to hear, and the disabled and elderly to walk, run, and even dance. AI methods are also being developed to augment human cognition. As an example, prototypes have been aimed at predicting what people will forget and helping them to remember and plan. Moving to the realm of scientific discovery, people working together with the Foldit online game[8] were able to discover the structure of the virus that causes AIDS in only three weeks, a feat that neither people nor computers working alone could match. Other studies have shown how the massive space of galaxies can be explored hand-in-hand by people and machines, where the tireless AI astronomer understands when it needs to reach out and tap the expertise of human astronomers.[7] There are many opportunities ahead for developing real-time systems that involve a rich interleaving of problem solving by people and machines.

However, building these collaborative systems raises a fourth set of risks stemming from challenges with fluidity of engagement and clarity about states and goals. Creating real-time systems where control needs to shift rapidly between people and AI systems is difficult. For example, airline accidents have been linked to misunderstandings arising when pilots took over from autopilots.[a] The problem is that unless the human operator has been paying very close attention, he or she will lack a detailed understanding of the current situation and can make poor decisions. Here again, AI methods can help solve these problems by anticipating when human control will be required and providing people with the critical information that they need.

A fifth set of risks concern the broad influences of increasingly competent automation on socioeconomics and the distribution of wealth.[2] Several lines of evidence suggest AI-based automation is at least partially responsible for the growing gap between per capita GDP and median wages. We need to understand the influences of AI on the distribution of jobs and on the economy more broadly. These questions move beyond computer science into the realm of economic policies and programs that might ensure that the benefits of AI-based productivity increases are broadly shared.

Achieving the potential tremendous benefits of AI for people and society will require ongoing and vigilant attention to the near- and longer-term challenges to fielding robust and safe computing systems. Each of the first four challenges listed in this Viewpoint (software quality, cyberattacks, "Sorcerer's Apprentice," and shared autonomy) is being addressed by current research, but even greater efforts are needed. We urge our research colleagues and industry and government funding agencies to devote even more attention to software quality, cybersecurity, and human-computer collaboration on tasks as we increasingly rely on AI in safety-critical functions.

At the same time, we believe scholarly work is needed on the longer-term concerns about AI. Working with colleagues in economics, political science, and other disciplines, we must address the potential of automation to disrupt the economic sphere. Deeper study is also needed to understand the potential of superintelligence or other pathways to result in even temporary losses of control of AI systems. If we find there is significant risk, then we must work to develop and adopt safety practices that neutralize or minimize that risk. We should study and address these concerns, and the broader constellation of risks that might come to the fore in the short- and long-term, via focused research, meetings, and special efforts such as the Presidential Panel on Long-Term AI Futures[b] organized by the AAAI in 2008–2009 and the One Hundred Year Study on Artificial Intelligence,[10,c] which is planning centuries of ongoing studies about advances in AI and its influences on people and society.

The computer science community must take a leadership role in exploring and addressing concerns about machine intelligence. We must work to ensure that AI systems responsible for high-stakes decisions will behave safely and properly, and we must also examine and respond to concerns about potential transformational influences of AI. Beyond scholarly studies, computer scientists need to maintain an open, two-way channel for communicating with the public about opportunities, concerns, remedies, and realities of AI. C

a See http://en.wikipedia.org/wiki/China_Airlines_Flight_006.

b See http://www.aaai.org/Organization/presidential-panel.php.

c See https://ai100.stanford.edu.

References

1. Bostrum, N. *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press, 2014.
2. Brynjolfsson, E. and McAfee, A. *The Second Machine Age: Work Progress, and Prosperity in a Time of Brilliant Technologies.* W.W. Norton & Company, New York, 2014.
3. Chen, F. and Rosu, G. Toward monitoring-oriented programming: A paradigm combining specification and implementation. *Electr. Notes Theor. Comput. Sci. 89*, 2 (2003), 108–127.
4. Good, I.J. Speculations concerning the first ultraintelligent machine. In *Advances in Computers, Vol. 6.* F.L. Alt and M. Rubinoff, Eds., Academic Press, 1965, 31–88.
5. Horvitz, E. Principles of mixed-initiative user interfaces. In *Proceedings of CHI '99, ACM SIGCHI Conference on Human Factors in Computing Systems* (Pittsburgh, PA, May 1999); http://bit.ly/1OEyLFW.
6. Huang, J. et al. ROSRV: Runtime verification for robots. *Runtime Verification*, (2014), 247–254.
7. Kamar, E., Hacker, S., and Horvitz, E. Combining human and machine intelligence in large-scale crowdsourcing. *AAMAS 2012* (Valencia, Spain, June 2012); http://bit.ly/1h6gfbU.
8. Khatib, F. et al. Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nature Structural and Molecular Biology 18* (2011), 1175–1177.
9. Shahaf, D. and Horvitz, E. Generalized task markets for human and machine computation. *AAAI 2010*, (Atlanta, GA, July 2010), 986–993; http://bit.ly/1gDIuho.
10. You, J. A 100-year study of artificial intelligence? *Science* (Jan. 9, 2015); http://bit.ly/1w664U5.

Thomas G. Dietterich (tgd@oregonstate.edu) is a Distinguished Professor in the School of Electrical Engineering and Computer at Oregon State University in Corvallis, OR, and president of the Association for the Advancement of Artificial Intelligence (AAAI).

Eric J. Horvitz (horvitz@microsoft.com) is Distinguished Scientist and Director of the Microsoft Research lab in Redmond, Washington. He is the former president of AAAI and continues to serve on AAAI's Strategic Planning Board and Committee on Ethics in AI.

Watch the authors discuss their work in this exclusive *Communications* video. http://cacm.acm.org/videos/rise-of-concerns-about-ai-reflections-and-directions