Research Priorities for Robust and Beneficial Artificial Intelligence: An Open Letter

Artificial intelligence (AI) research has explored a variety of problems and approaches since its inception, but for the last 20 years or so has been focused on the problems surrounding the construction of intelligent agents — systems that perceive and act in some environment. In this context, “intelligence” is related to statistical and economic notions of rationality — colloquially, the ability to make good decisions, plans, or inferences. The adoption of probabilistic and decision-theoretic representations and statistical learning methods has led to a large degree of integration and cross-fertilization among AI, machine learning, statistics, control theory, neuroscience, and other fields. The establishment of shared theoretical frameworks, combined with the availability of data and processing power, has yielded remarkable successes in various component tasks such as speech recognition, image classification, autonomous vehicles, machine translation, legged locomotion, and question-answering systems.

As capabilities in these areas and others cross the threshold from laboratory research to economically valuable technologies, a virtuous cycle takes hold whereby even small improvements in performance are worth large sums of money, prompting greater investments in research. There is now a broad consensus that AI research is progressing steadily, and that its impact on society is likely to increase. The potential benefits are huge, since everything that civilization has to offer is a product of human intelligence; we cannot predict what we might achieve when this intelligence is magnified by the tools AI may provide, but the eradication of disease and poverty are not unfathomable. Because of the great potential of AI, it is important to research how to reap its benefits while avoiding potential pitfalls.

The progress in AI research makes it timely to focus research not only on making AI more capable, but also on maximizing the societal benefit of AI. Such considerations motivated the AAAI 2008–09 Presidential Panel on Long-Term AI Futures and other projects on AI impacts, and constitute a significant expansion of the field of AI itself, which up to now has focused largely on techniques that are neutral with respect to purpose. We recommend expanded research aimed at ensuring that increasingly capable AI systems are robust and beneficial: our AI systems must do what we want them to do. The attached research priorities document [see page 105 of this issue of AI Magazine] gives many examples of such research directions that can help maximize the societal benefit of AI. This research is by necessity interdisciplinary, because it involves both society and AI. It ranges from economics, law, and philosophy to computer security, formal methods, and, of course, various branches of AI itself.

In summary, we believe that research on how to make AI systems robust and beneficial is both important and timely, and that there are concrete research directions that can be pursued today.

Editor’s Note: This letter has been signed, by press time, by nearly 7,000 persons, including many AAAI Fellows. Signatories describe themselves as computer scientists, innovators, entrepreneurs, statisticians, journalists, engineers, authors, professors, teachers, students, CEOs, economists, developers, philosophers, artists, futurists, physicists, filmmakers, health-care professionals, research analysts, and members of many other fields. The earliest signatories follow, reproduced in order and as they signed. For the complete list, see tinyurl.com/ailetter. - ed.

Stuart Russell, Berkeley, Professor of Computer Science, director of the Center for Intelligent Systems, and coauthor of the standard textbook Artificial Intelligence: a Modern Approach

Tom Dietterich, Oregon State, President of AAAI, Professor and Director of Intelligent Systems

Eric Horvitz, Microsoft research director, ex-AAA president, cochair of the AAAI presidential panel on long-term AI futures

Bart Selman, Cornell, Professor of Computer Science, cochair of the AAAI presidential panel on long-term AI futures

Francesca Rossi, Padova & Harvard, Professor of Computer Science, cochair of the AAAI Committee on Impact of AI and Ethical Issues

Demis Hassabis, cofounder of DeepMind

Shane Legg, cofounder of DeepMind

Mustafa Suleyman, cofounder of DeepMind

Dileep George, cofounder of Vicarious

Scott Phoenix, cofounder of Vicarious

Yann LeCun, head of Facebook’s Artificial Intelligence Laboratory

Geoffrey Hinton, University of Toronto and Google Inc.

Yoshua Bengio, Université de Montréal

Peter Norvig, Director of research at Google and coauthor of the standard textbook Artificial Intelligence: a Modern Approach

Oren Etzioni, CEO of Allen Inst. for AI

Guruduth Banavar, VP, Cognitive Computing, IBM Research