Letters to the Editor

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 AAAI 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, IJCAI President and Cochair of 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

Michael Wooldridge, Oxford, Head of Dept. of Computer Science, Chair of European Coordinating Committee for Artificial Intelligence

Leslie Pack Kaelbling, MIT, Professor of Computer Science and Engineering, founder of the Journal of Machine Learning Research

Tom Mitchell, CMU, former President of AAAI, chair of Machine Learning Department

Toby Walsh, Univ. of New South Wales & NICTA, Professor of AI and President of the AI Access Foundation

Murray Shanahan, Imperial College, Professor of Cognitive Robotics

Michael Osborne, Oxford, Associate Professor of Machine Learning

David Parkes, Harvard, Professor of Computer Science

Laurent Orseau, Google DeepMind

Ilya Sutskever, Google, AI researcher

Blaise Aguera y Arcas, Google, AI researcher

Joscha Bach, MIT, AI researcher

Bill Hibbard, Madison, AI researcher

Steve Omohundro, AI researcher

Ben Goertzel, OpenCog Foundation

Richard Mallah, Cambridge Semantics, Director of Advanced Analytics, AI researcher

Alexander Wissner-Gross, Harvard, Fellow at the Institute for Applied Computational Science

Adrian Weller, Cambridge, AI researcher

Jacob Steinhardt, Stanford, AI Ph.D. student

Nick Hay, Berkeley, AI Ph.D. student

Jaan Tallinn, cofounder of Skype, CSER, and FLI

Elon Musk, SpaceX, Tesla Motors

Steve Wozniak, cofounder of Apple

Luke Nosek, Founders Fund

Aaron VanDevender, Founders Fund

Erik Brynjolfsson, MIT, Professor at and director of MIT Initiative on the Digital Economy

Margaret Boden, U. Sussex, Professor of Cognitive Science

Martin Rees, Cambridge, Professor Emeritus of Cosmology and Astrophysics, Gruber & Crafoord laureate

Huw Price, Cambridge, Bertrand Russell Professor of Philosophy

Nick Bostrom, Oxford, Professor of Philosophy, Director of Future of Humanity Institute (Oxford Martin School)

Stephen Hawking, Director of research at the Department of Applied Mathematics and Theoretical Physics at Cambridge, 2012 Fundamental Physics Prize laureate for his work on quantum gravity

Luke Muehlhauser, Executive Director of Machine Intelligence Research Institute (MIRI)

Eliezer Yudkowsky, MIRI researcher, cofounder of MIRI (then known as SIAI)

Katja Grace, MIRI researcher

Benja Fallenstein, MIRI researcher

Nate Soares, MIRI researcher

Paul Christiano, Berkeley, Computer Science graduate student

Anders Sandberg, Oxford, Future of Humanity Institute researcher (Oxford Martin School)

Daniel Dewey, Oxford, Future of Humanity Institute researcher (Oxford Martin School)

Stuart Armstrong, Oxford, Future of Humanity Institute researcher (Oxford Martin School)

Toby Ord, Oxford, Future of Humanity Institute researcher (Oxford Martin School), Founder of Giving What We Can

Neil Jacobstein, Singularity University

Dominik Grewe, Google DeepMind

Roman V. Yampolskiy, University of Louisville

Vincent C. Müller, ACT/Anatolia College

Amnon H Eden, University Essex

Henry Kautz, University of Rochester

Boris Debic, Google, Chief History Officer

Kevin Leyton-Brown, University of British Columbia, Professor of Computer Science

Trevor Back, Google DeepMind

Moshe Vardi, Rice University, editor-in-chief of Communications of the ACM

Peter Sincak, prof. TU Kosice, Slovakia

Tom Schaul, Google DeepMind

Grady Booch, IBM Fellow

Alan Mackworth, Professor of Computer Science, University of British Columbia. Ex AAAI President

Andrew Davison, Professor of Robot Vision, Director of the Dyson Robotics Lab at Imperial College London

Daniel Weld, WRF / TJ Cable Professor of Computer Science & Engineering, University of Washington

Michael Witbrock, Cycorp Inc & AI4Good .org

Stephen L. Reed, ai-coin.com

Thomas Stone, Cofounder of PredictionIO

Dan Roth, University of Illinois, Editor in Chief of *The Journal of AI Research (JAIR)*

Babak Hodjat, Sentient Technologies

Vincent Vanhoucke, Google, AI researcher

Itamar Arel, Stanford University, Prof. of Computer Science

Ramon Lopez de Mantaras, Director of the Artificial Intelligence Research Institute, Spanish National Research Council

Antoine Blondeau, Sentient Technologies

George Dvorsky, Contributing Editor, io9; Chair of the Board, Institute for Ethics and Emerging Technologies

George Church, Harvard & MIT

Klaus-Dieter Althoff, University of Hildesheim, Professor of Artificial Intelligence; Head of Competence Center Case-Based Reasoning, German Research Center for Artificial Intelligence, Kaiserslautern; Editor-in-Chief *German Journal on Artificial Intelligence*

Christopher Bishop, Distinguished Scientist, Microsoft Research

Vernor Vinge, San Diego, Professor Emeritus of Computer Science

Steve Crossan, Google

Charina Choi, Google

Matthew Putman, CEO of Nanotronics Imaging

Owain Evans, MIT, Ph.D. student in probabilistic computing

Viktoriya Krakovna, Harvard, Statistics Ph.D. student, FLI cofounder

Janos Kramar, FLI researcher

Ryan Calo, U. Washington, Assistant Professor of Law

Heather Roff Perkins, U. Denver, visiting professor

Tomaso Poggio, Director, Center for Brains, Minds and Machines

Joshua Greene, Harvard, Associate Professor of Psychology

Anthony Aguirre, Santa Cruz, Professor of Physics, cofounder of FLI

Frank Wilczek, MIT, Professor of Physics, Nobel Laureate for his work on the strong nuclear force

Marin Soljacic, MIT, Professor of Physics, McArthur Fellow, Founder of WiTricity

Max Tegmark, MIT, Professor of Physics, cofounder of FLI and FQXi

Meia Chita-Tegmark, Boston University, cofounder of FLI

Michael Vassar, founder of MetaMed and ex-president of MIRI (then known as SIAI)

Seán Ó Héigeartaigh, University of Cambridge, Executive Director, CSER

Andrew Snyder-Beattie, Oxford, Future of Humanity Institute Project Manager (Oxford Martin School)

Cecilia Tilli, Oxford, Future of Humanity Institute researcher (Oxford Martin School)

Geoff Anders, founder of Leverage Research

JB Straubel, cofounder of Tesla

Sam Harris, Project Reason

Ajay Agrawal, U. Toronto

James Manyika, McKinsey

James Moor, Dartmouth

Wendell Wallach, Yale