Statement on AI Risk

AI experts and public figures express their concern about AI risk.

AI experts, journalists, policymakers, and the public are increasingly discussing a broad spectrum of important and urgent risks from AI. Even so, it can be difficult to voice concerns about some of advanced AI's most severe risks. The succinct statement below aims to overcome this obstacle and open up discussion. It is also meant to create common knowledge of the growing number of experts and public figures who also take some of advanced AI's most severe risks seriously.

Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.

Signatories:

☐ AI Scientists ☐ Other Notable Figures

Geoffrey Hinton
Emeritus Professor of Computer Science, University of Toronto

Yoshua Bengio
Professor of Computer Science, U. Montreal / Mila

Demis Hassabis

Dario Amodei
CEO, Anthropic

Dawn Song
Professor of Computer Science, UC Berkeley

Ted Lieu
Congressman, US House of Representatives

Bill Gates
Gates Ventures

Ya-Qin Zhang
Professor and Dean, AIR, Tsinghua University

Ilya Sutskever
Co-Founder and Chief Scientist, OpenAI

Igor Babuschkin
Co-Founder, xAI

Shane Legg
Chief AGI Scientist and Co-Founder, Google DeepMind

Martin Hellman
Professor Emeritus of Electrical Engineering, Stanford

James Manyika
SVP, Research, Technology and Society, Google-Alphabet

Yi Zeng
Professor and Director of Brain-inspired Cognitive AI Lab, Institute of Automation, Chinese Academy of Sciences

Xianyuan Zhan
Assistant Professor, Tsinghua University

Albert Efimov
Chief of Research, Russian Association of Artificial Intelligence

Alvin Wang Graylin
China President, HTC

Jianyi Zhang

Associate Professor of Computer Science, UC Berkeley

Christine Parthemore
CEO and Director of the Janne E. Nolan Center on Strategic Weapons, The Council on Strategic Risks

Bill McKibben
Schumann Distinguished Scholar, Middlebury College

Alan Robock
Distinguished Professor of Climate Science, Rutgers University

Angela Kane
Vice President, International Institute for Peace, Vienna; former UN High Representative for Disarmament Affairs

Audrey Tang
Digitalminister.tw and Chair of National Institute of Cyber Security

Daniela Amodei
President, Anthropic

David Silver
Professor of Computer Science, Google DeepMind and UCL

Lila Ibrahim
COO, Google DeepMind

Stuart Russell
Professor of Computer Science, UC Berkeley

Tony (Yuhuai) Wu
Co-Founder, xAI

Marian Rogers Croak
VP Center for Responsible AI and Human Centered Technology, Google

Andrew Barto
Professor Emeritus, University of Massachusetts

Mira Murati
CTO, OpenAI

Jaime Fernández Fisac

Diyi Yang
Assistant Professor, Stanford University

Gillian Hadfield
Professor, CIFAR AI Chair, University of Toronto, Vector Institute for AI

Laurence Tribe
University Professor Emeritus, Harvard University

Pattie Maes
Professor, Massachusetts Institute of Technology - Media Lab

Kevin Scott
CTO, Microsoft

Eric Horvitz
Chief Scientific Officer, Microsoft

Peter Norvig
Education Fellow, Stanford University

Joseph Sifakis
Turing Award 2007, Professor, CNRS - Universite Grenoble - Alpes

Atoosa Kasirzadeh
Assistant Professor, University of Edinburgh, Alan Turing Institute

Erik Brynjolfsson
Professor and Senior Fellow, Stanford Institute for Human-Centered AI

Mustafa Suleyman
CEO, Inflection AI

Emad Mostaque
CEO, Stability AI

Ian Goodfellow
Principal Scientist, Google DeepMind

John Schulman
Co-Founder, OpenAI

Wojciech Zaremba
Co-Founder, OpenAI

Baburam Bhattarai
Former Prime Minister of Nepal, Society of Nepalese Architects

Russell Schweickart
Apollo 9 Astronaut, Association of Space Explorers, B612 Foundation

Andy Weber
Former U.S. Assistant Secretary of Defense for Nuclear, Chemical, and Biological Defense Programs, Council on Strategic Risks

Allison Macfarlane
Former Chairman, US Nuclear Regulatory Commission

Nicholas Fairfax (Lord Fairfax)
Member, House of Lords

Mark Beall
Former Director of AI Strategy and Policy, Department of Defense

Lord Strathcarron
Peer, House of Lords

Stephen Luby
Professor of Medicine (Infectious Diseases), Stanford University

David Haussler
Professor and Director of the Genomics Institute, UC Santa Cruz

Ju Li
Professor of Nuclear Science & Engineering and Professor of Materials Science & Engineering, Massachusetts Institute of Technology

David Chalmers
Professor of Philosophy, New York University

Daniel Dennett
Emeritus Professor of Philosophy, Tufts University

Peter Railton
Professor of Philosophy at University of Michigan, Ann Arbor

Peter Singer
Professor, Princeton University

Sheila McIlraith

Research Scientist, Google DeepMind

Mary Phuong
Research Scientist, Google DeepMind

Mariano-Florentino Cuéllar
President, Carnegie Endowment for International Peace

Lex Fridman
Research Scientist, MIT

Sharon Li
Assistant Professor of Computer Science, University of Wisconsin Madison

Phillip Isola
Associate Professor of Electrical Engineering and Computer Science, MIT

David Krueger
Assistant Professor of Computer Science, University of Cambridge

Jacob Steinhardt
Assistant Professor of Computer Science, UC Berkeley

Martin Rees
Professor of Physics, Cambridge University

Nando de Freitas
Director, Science Board, Google DeepMind

Hongwei Qin
Research Director, SenseTime

He He
Assistant Professor of Computer Science and Data Science, New York University

David McAllester
Professor of Computer Science, TTIC

Vincent Conitzer
Professor of Computer Science, Carnegie Mellon University and University of Oxford

Bart Selman

Professor of Engineering Science, University of Oxford

James Mickens
Professor of Computer Science, Harvard University

Michael Wellman
Professor & Chair of Computer Science and Engineering, University of Michigan

Luis Videgaray
Senior Lecturer, MIT; Former Minister of Interior and Exterior Relations of Mexico

Jinwoo Shin
KAIST Endowed Chair Professor, Korea Advanced Institute of Science and Technology

Alice Oh
Professor at The School of Computing, KAIST and Director, MARS AI Research Center

Dae-Shik Kim
Professor of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST)

Edith Elkind
Professor of Computing Science, University of Oxford

Ray Kurzweil
Principal Researcher and AI Visionary, Google

Frank Hutter
Professor of Machine Learning, Head of ELLIS Unit, University of Freiburg

Alexey Dosovitskiy
Research Scientist, Google DeepMind

Jaan Tallinn
Co-Founder of Skype

Vitalik Buterin
Founder and Chief Scientist, Ethereum, Ethereum Foundation

Adam D'Angelo
CEO, Quora, and board member, OpenAI

Co-founder and CEO, Asana

Shane Torchiana
CEO, Bird

Thuan Q. Pham
Former CTO, Uber, Board member, Nubank

Scott Aaronson
Schlumberger Chair of Computer Science, University of Texas at Austin

Max Tegmark
Professor, MIT, Center for AI and Fundamental Interactions

Bruce Schneier
Lecturer, Harvard Kennedy School

Martha Minow
Professor, Harvard Law School

Gabriella Blum
Professor of Human Rights and Humanitarian Law, Harvard Law

Kevin Esvelt
Associate Professor of Biology, MIT

Edward Wittenstein
Executive Director, International Security Studies, Yale Jackson School of Global Affairs, Yale University

Sonny Ramaswamy
President, Northwest Commission on Colleges & Universities

Laurie Zoloth
Margaret E. Burton Professor of Religion and Ethics, University of Chicago

Karina Vold
Assistant Professor, University of Toronto

Victor Veitch
Assistant Professor of Data Science and Statistics, University of Chicago

Dylan Hadfield-Menell
Assistant Professor of Computer Science, MIT

Mengye Ren
Assistant Professor of Computer Science, New York University

Shiri Dori-Hacohen
Assistant Professor of Computer Science, University of Connecticut

Miles Brundage
Head of Policy Research, OpenAI

Allan Dafoe
AGI Strategy and Governance Team Lead, Google DeepMind

Helen King
Senior Director of Responsibility and Strategic Advisor to Research, Google DeepMind

Jade Leung
Governance Lead, OpenAI

Jess Whittlestone
Head of AI Policy, Centre for Long-Term Resilience

Sarah Kreps
John L. Wetherill Professor and Director of the Tech Policy Institute, Cornell University

Jared Kaplan
Co-Founder, Anthropic

Chris Olah
Co-Founder, Anthropic

Andrew Revkin
Director, Initiative on Communication & Sustainability, Columbia University - Climate School

Carl Robichaud
Program Officer (Nuclear Weapons), Longview Philanthropy

Leonid Chindelevitch
Lecturer in Infectious Disease Epidemiology, Imperial College London

Nicholas Dirks
President, The New York Academy of Sciences

CEO, Faculty

Rob Pike
Distinguished Engineer (retired), Co-Creator of Golang, Google

Clare Lyle
Research Scientist, Google DeepMind

Nisarg Shah
Assistant Professor, University of Toronto

Ryota Kanai
CEO, Araya, Inc.

Tim G. J. Rudner
Assistant Professor and Faculty Fellow, New York University

Noah Fiedel
Director, Research and Engineering, Google DeepMind

Jakob Foerster
Associate Professor of Engineering Science, University of Oxford

Michael Osborne
Professor of Machine Learning, University of Oxford

Marina Jirotka
Professor of Human Centred Computing, University of Oxford

Nancy Chang
Research Scientist, Google

Tom Schaul
Research Scientist, Google DeepMind

Roger Grosse
Associate Professor of Computer Science, University of Toronto and Anthropic

David Duvenaud
Associate Professor of Computer Science, University of Toronto

Daniel M. Roy

Kanjun Qiu
CEO, Generally Intelligent

Chris J. Maddison
Assistant Professor of Computer Science, University of Toronto

Tegan Maharaj
Assistant Professor of the Faculty of Information, University of Toronto

Florian Shkurti
Assistant Professor of Computer Science, University of Toronto

Jeff Clune
Associate Professor of Computer Science and Canada CIFAR AI Chair, The University of British Columbia and the Vector Institute

Eva Vivalt
Assistant Professor of Economics, University of Toronto, and Director, Global Priorities Institute, University of Oxford

Jacob Tsimerman
Professor of Mathematics, University of Toronto

Emanuel Adler
Professor Emeritus, University of Toronto

Danit Gal
Technology Advisor at the UN; Associate Fellow, Leverhulme Centre for the Future of Intelligence, University of Cambridge

Jean-Claude Latombe
Professor (Emeritus) of Computer Science, Stanford University

Scott Niekum
Associate Professor of Computer Science, University of Massachusetts Amherst

Lionel Levine
Associate Professor of Mathematics, Cornell University

Co-Founder & Managing Director, Lux Capital

Norman Sadeh
Professor of Computer Science/Co-Director Privacy Engineering Program, Carnegie Mellon University

Brian Ziebart
Associate Professor of Computer Science, University of Illinois Chicago

Roberto Baldoni
Former Director General, National Cybersecurity Agency of Italy

Aza Raskin
Cofounder, Center for Humane Technology, The Earth Species Project

Prasad Tadepalli
Professor of Computer Science, Oregon State University

David L Roscoe
Board Chair Emeritus and Advisory Council Chair, The Hastings Center

Tristan Harris
Executive Director, Center for Humane Technology

Anthony Aguirre
Executive Director, Future of Life Institute

Sam Harris
Author, Neuroscientist, Making Sense / Waking Up

Grimes
Musician / Artist

Chris Anderson
Dreamer-in-Chief, TED

Ramy Youssef
Actor/Director, Cairo Cowboy

Rif A. Saurous
Research Director, Google

James W. Pennebaker

Associate Professor of Statistics, UC Berkeley

Jose Hernandez-Orallo
Professor of Computer Science, Technical University of Valencia

R. Martin Chavez
Vice Chairman, Sixth Street Partners, Former CFO and CIO of Goldman Sachs

Paul S. Rosenbloom
Professor Emeritus of Computer Science, University of Southern California

Timothy Lillicrap
Research Director, Google DeepMind

Samuel Albanie
Assistant Professor of Engineering, University of Cambridge

Jascha Sohl-Dickstein
Principal Scientist, Google DeepMind

Ronald Craig Arkin
Regents' Professor Emeritus, Georgia Institute of Technology

Been Kim
Research Scientist, Google DeepMind

Mehran Sahami
Professor and Chair of Computer Science, Stanford University

Cihang Xie
Assistant Professor of Computer Science and Engineering, UC Santa Cruz

Philip S. Thomas
Associate Professor, University of Massachusetts

Hilary Greaves
Professor of Philosophy, University of Oxford

Pierre Baldi
Professor, University of California, Irvine

Giovanni Vigna
Professor, UC Santa Barbara

Shai Shalev-Shwartz
Professor, The Hebrew University of Jerusalem

Katherine Lee
Research Scientist, Google DeepMind

Felix Juefei Xu
Research Scientist, Meta AI

Foutse Khomh
Professor and Canada CIFAR AI Chair, Polytechnique Montreal

Dan Hendrycks
Executive Director, Center for AI Safety

Show More

Sign the statement

You will be sent a verification email. CAIS is committed to ensuring the validity of the signatories. There may be a small delay before signatures are added.

Full Name *

Work email *

Your Title *

Affiliation *

Submit

Subscribe to the AI Safety Newsletter

example@gmail.com Subscribe

CAIS is an AI safety non-profit. Our mission is to reduce societal-scale risks from artificial intelligence.

Statement on AI Risk
Field Building
CAIS Research
Compute Cluster
Philosophy Fellowship
CAIS Blog

2023 Impact Report
Frequently Asked Questions
Learn About AI Risk
CAIS Media Kit
Terms of Service
Privacy Policy

Contact Us
Careers

General: contact@safe.ai
Media: media@safe.ai

Cookies Notice:
This website uses cookies to identify pages that are being used most frequently. This helps us analyze data about web page traffic and improve our website. We only use this information for the purpose of statistical analysis and then the data is removed from the system. We do not and will never sell user data. Read more about our cookie policy on our privacy policy. Please contact us if you have any questions.

Credits

Website by Osborn Design Works