

Stanford Center for AI Safety Affiliates Program

AI Safety Vision The mission of the Stanford Center for AI Safety is to develop rigorous techniques for building safe and trustworthy AI systems and establishing confidence in their behavior and robustness, thereby facilitating their successful adoption in society.

AI Safety Team The AI Safety team consists of Stanford professors and their research groups:

Clark Barrett, Computer Science: Satisfiability, formal software and hardware verification, and automated reasoning.

Emma Brunskill, Computer Science: Safe reinforcement learning, robustness, performance, fairness and safety.

Mariano-Florentino Cuellar, Law: Administrative law and legislation, cyberlaw and artificial intelligence, criminal justice, public health, immigration, international law and security, and the history of institutions.

David Dill, Computer Science: Theory and application of formal verification techniques to system designs.

Charles Easley, Management Science and Engineering: Artificial intelligence applications in financial systems.

Grace Gao, Aeronautics and Astronautics: Robust and secure positioning, navigation and timing with applications to manned and unmanned aerial vehicles, robotics and power systems.

Mykel Kochenderfer, Aeronautics and Astronautics and (by courtesy) Computer Science: Advanced algorithms and analytical methods for robust decision making in the presence of uncertainty.

Sanmi Koyejo, Management Science and Engineering: Principles and practice of trustworthy machine learning, including fairness and robustness..

Steve Luby, School of Medicine: Epidemiology and existential risks.

Azalea Mirhoseini, Management Science and Engineering: Scalable and self-improving AI Systems.

Marco Pavone, Aeronautics and Astronautics: Analysis, design, and control of autonomous systems, with a particular emphasis on large-scale robotic networks and autonomous aerospace vehicles.

Stanford Center for AI Safety Affiliates Program

Dorsa Sadigh, Computer Science and Electrical Engineering: Design of algorithms for autonomous systems that safely and reliably interact with people.

Mac Schwager, Aeronautics and Astronautics: Algorithms for control, estimation, and learning in groups of autonomous aircraft, autonomous cars, and robots.

Diyi Yang, Computer Science: Socially aware natural language processing for improving human-human and human-computer interaction.

James Zou, Biomedical Data Science and (by courtesy) Computer Science and Electrical Engineering: Accountable, reliable, robust machine learning with applications to biomedicine and healthcare.

Philosophy

AI-based systems increasingly play important roles in many areas of modern life including manufacturing, transportation, aerospace, and healthcare. These complex systems are expected to be smart and reliable, and although the algorithms used for training these systems are quite well understood, it is very hard for humans to reason about machine-learned systems. This situation is unsatisfactory if we are to use AI components in safety-critical systems. The research thrusts of the Center for AI Safety seek to address this situation by developing methods for verification, robustness, safe learning, explainability, fairness, and more.

Topics

The Center for AI Safety covers three broad topics:

- 1) Formal Techniques for AI Safety
 - a) Formal specifications for systems with AI components
 - b) Verification of systems with AI components
 - c) Analysis of adversarial robustness
 - d) Automatic test-case generation
- 2) Learning and Control for AI Safety
 - a) Safe exploration and learning for better perception of AI systems
 - b) Safe planning and control for AI agents
- 3) Transparency in AI Safety
 - a) Fairness in AI
 - b) Explainable and accountable AI
 - c) Diagnosis and repair for systems with AI components

Engagement

Corporate members are a vital and integral part of the Center for AI Safety. They provide insight on real-world use cases, valuable financial support for research, and a path to large-scale impact.

Corporate engagement includes the following elements:



Stanford Center for AI Safety Affiliates Program

- Opportunity to contribute to the definition of a flagship research project involving multiple faculty and students (Core Members)
- Opportunity to send a Visiting Scholar to Stanford, subject to satisfying university requirements (Core Members)
- Faculty visits to the member company (Core Members)
- Participation on the Center for AI Safety Board of Advisors (Core Members)
- Invitations to semiannual research retreats (Core and Associate Members)
- Slack channel invitation (Core and Associate Members)
- Research seminar announcements (Core and Associate Members)
- Student resume book (Core and Associate Members)

Stanford Center for AI Safety Affiliates Program

Funding

Corporate members have a choice of two membership levels. Associate Members contribute \$100,000 per year and receive full access to all Center for AI Safety research, faculty, and students. Core Members contribute \$300,000 per year with the expectation of at least three years of membership and receive all benefits of Associate Members plus additional opportunities to help define the research agenda, participate in the leadership of the Center, and engage even more deeply with the faculty and students. The Center for AI Safety is a Stanford University industrial affiliates program and is subject to university policies for such programs including openness in research, publication and broad sharing of results, and faculty freedom to pursue research topics and methodology of their choice. See <https://industrialaffiliates.stanford.edu/>.

IP

Researchers in the Center for AI Safety will use and develop open-source software, and it is the intention of all Center researchers that any software released will be released under an open source model, such as MIT and BSD.

Information

For further information please contact any of the professors listed above or Duncan Eddy, Executive Director, at deddy@stanford.edu.