SAFE

STANFORD UNIVERSITY

Center for AI Safety

Executive Overview

Industry Partnership & Membership Program

Prepared for Prospective Industry Partners

Last Updated: February 5, 2026

CONTACT

Kiana Jafari, Executive Director

Stanford Center for AI Safety

Stanford University

kjafari@stanford.edu

For Stanford policies on Industrial Affiliate Programs, please visit: doresearch.stanford.edu

Our Mission

The Stanford Center for AI Safety leads research, education, and policy to ensure AI systems are safe, trustworthy, and beneficial for humanity. Housed within Stanford's School of Engineering, the Center brings together world-class faculty from computer science, aeronautics and astronautics, management science, medicine, psychiatry, and the social sciences to address the most pressing challenges in AI safety.

Our Vision

We envision a future where AI technologies are developed and deployed responsibly, with robust safety guarantees, transparent decision-making, and alignment with human values. Through interdisciplinary collaboration spanning computer science, engineering, law, and social sciences, we aim to shape the policies and practices that govern AI development worldwide.

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.

Research Areas

The Center's research is organized around five core pillars, each addressing a critical dimension of AI safety:

Research Area	Description
Autonomous & Safety-Critical Systems	Designing AI systems capable of making reliable decisions in uncertain and dynamic environments, from self-driving cars and aircraft to surgical robots and power grids.
Verification, Validation & Assurance	Developing rigorous mathematical methods to prove AI systems behave as intended, including formal verification, testing methodologies, and certification frameworks for neural networks and learned components.
Foundation Model Safety	Studying the failure modes, emergent capabilities, and potential risks of large language models and generative AI systems, including alignment, robustness to adversarial inputs, and safe deployment practices.

Human-AI Collaboration & Oversight	Investigating how humans and AI systems can work together safely and effectively, including trust calibration, interpretability, human-in-the-loop control, and mechanisms for meaningful human oversight.
Trustworthy AI, Governance & Policy	Building frameworks for responsible AI development and deployment that address fairness, accountability, transparency, and societal impact, while informing regulation and industry best practices.

Leadership & Faculty

The Center is led by three Co-Directors and an Executive Director, supported by **16+ faculty affiliates** spanning seven departments, **8 postdoctoral and visiting scholars**, and **30+ graduate students**.

Name	Role & Department	Focus
Clark Barrett	Co-Director, Professor (Research) of Computer Science	Formal verification, automated reasoning, SMT solvers for neural network verification
Grace X. Gao	Co-Director, Assoc. Professor of Aeronautics & Astronautics	Navigation systems, autonomous systems safety assurance
Mykel Kochenderfer	Co-Director, Assoc. Professor of Aeronautics & Astronautics	Decision-making under uncertainty, safe autonomous systems
Kiana Jafari	Executive Director	Human-centric artificial intelligence, center operations & industry partnerships

Faculty Affiliates

The Center draws on deep expertise across Stanford's departments. Below are the faculty affiliates and their connections to AI safety research:

Faculty	Department	AI Safety Connection
Somil Bansal	Aero & Astro	Neural reachability methods for verification, safety filters, control barrier functions. Collaborations with Skydio, Google, Boeing, NASA.
Jose Blanchet	MS&E	Distributionally robust optimization (DRO) for ML models that maintain performance under distributional shifts. Erlang Prize recipient.
Emma Brunskill	Computer Science	Reinforcement learning with safety constraints and sample efficiency for real-world deployment.
David Dill	CS	Co-developed Reluplex algorithm for verifying deep neural networks (CAV Award 2024). NAE member. Co-authored SAFE whitepaper.

Charles Eesley	MS&E	AI/ML algorithms' influence on digital platforms; misinformation dynamics (published in Nature). Stanford Technology Ventures Program director.
Carlos Guestrin	Computer Science	AI ethics, responsible AI development and deployment practices.
Sanmi Koyejo	Computer Science	Leads Stanford Trustworthy AI Research (STAIR) lab. Fairness, robustness, privacy. PECASE Award, Sloan Fellow. Work cited in 2024 Economic Report of the President.
Steve Luby	Medicine	Co-Director, Stanford Existential Risks Initiative (SERI). Frames AI catastrophic risks alongside pandemics, nuclear war, and climate change.
Azalia Mirhoseini	Computer Science	Scalable AI systems, AlphaChip for TPU design. Previously at Anthropic working on Claude's reliability. Founded Scaling Intelligence Lab & Ricursive Intelligence.
Marco Pavone	Aero & Astro	Safety assurance for autonomous vehicles: reachability-based safety, control barrier functions, conformal prediction. Also NVIDIA Distinguished Research Scientist.
Dorsa Sadigh	CS & EE	Safe and interactive robotics, human-robot interaction with safety guarantees.
Mac Schwager	Aero & Astro	Multi-robot systems, distributed safe coordination for autonomous aircraft, cars, and collaborative robotics.
Madeleine Udell	MS&E	Interpretable ML, fairness under unawareness, LLMs for optimization (OptiMUS). Applications in healthcare, finance, engineering.
Nina Vasan	Psychiatry	Pioneer in AI psychological safety. First mental health safety benchmarks for LLMs (with ML Commons). Created Stanford GenAI Psychological Safety Plan. Advised Pinterest, TikTok.
Diyi Yang	Computer Science	Socially aware NLP, understanding social dimensions of language for safer human-AI interaction. Forbes 30 Under 30.
James Zou	Biomedical Data Science	AI fairness, bias detection, robustness. Data Shapley for data valuation. LLM safety benchmarks, AI watermarking. Chan-Zuckerberg Investigator.

Education & Courses

Center faculty teach a comprehensive portfolio of AI safety courses at Stanford, organized across four tiers. These courses form a pipeline that trains the next generation of AI safety researchers and practitioners—many of whom are available for industry recruiting.

Current & Upcoming

Course	Title	Term	Instructor
AA275	Navigation for Autonomous Systems	Autumn	Grace Gao
AA228V / CS238V	Validation of Safety-Critical Systems	Winter	Sydney Katz
PSYC 248	AI's Psyche and Psych: Mental Health in the Age of AI	Winter	Nina Vasan

Foundational Courses

Course	Title	Term	Instructor
CS120	Introduction to AI Safety	Autumn	Max Lamparth
CS221	Artificial Intelligence: Principles and Techniques	Spring	Sanmi Koyejo
CS281	Ethics of Artificial Intelligence	Spring	Carlos Guestrin

Advanced & Specialized

Course	Title	Term	Instructor
AA228 / CS238	Decision Making under Uncertainty	Autumn	Mykel Kochenderfer
AA274A / CS237A	Principles of Robot Autonomy I	Autumn	Mac Schwager
AA274B / CS237B	Principles of Robot Autonomy II	Winter	Dorsa Sadigh & Marco Pavone
CS333	Safe and Interactive Robotics	Winter	Dorsa Sadigh
AA212	Advanced Feedback Control Design	Winter	Mac Schwager
AA273	State Estimation and Filtering	Spring	Mac Schwager

Seminars & Special Topics

Course	Title	Term	Instructor
CS521	Seminar on AI Safety	Spring	Faculty rotation

Events & Engagement

The Center hosts regular events that bring together faculty, students, and industry leaders. The flagship **Annual AI Safety Meeting** (next: September 22, 2025, Paul Brest Hall, Stanford) serves as the primary gathering for the community. We also run the **AI Safety Seminar Series** each quarter, **Distinguished Speaker fireside chats**, and specialized workshops on topics such as explainability and responsible AI.

Membership Tiers

Corporate members are a vital and integral part of the Center for AI Safety. They provide insight on real-world use cases, financial support for research, and a path to large-scale impact. Membership is structured across three tiers designed to match different levels of organizational engagement:

Benefit	Supporter $50k/yr	Associate $150k/yr	Strategic $300k/yr
Bi-Annual Distinguished Series Access	✓	✓	✓
Annual AI Safety Summit Invitation	✓	✓	✓
Digital Leadership: Virtual Fireside Chats (Attendee)	✓	✓	✓
Brand Recognition: Logo Display	✓	✓	✓
Executive Education Discount for Selected Courses	✓	✓	✓
Nominate Speakers for Virtual Fireside Chats	—	✓	✓
Spotlight at the Annual Summit	—	✓	✓
Talent & Recruitment Access	—	✓	✓
Annual Curated Faculty Interactions	—	✓	✓
Research Theme Engagement	---	1 Theme	2 Themes
Keynote Recognition at Annual Summit	—	—	✓
Bi-Annual Curated Faculty Interactions	—	—	✓
Named Research/Scholar Support	—	—	✓
Engagement with Multiple Research Groups	—	—	✓
Visiting Scholar Position*	—	Add-on	Add-on

** Visiting Scholar positions are subject to Stanford's Visiting Scholar Policy. The desired membership commitment is for 3 years to ensure continuity of funding and successful completion of research; membership is renewed annually.*

Research Openness

Center researchers develop and use open-source software, and any software released will be released under an open-source model. All research results arising from member funding are shared with all program members and the general public. Members may request that additional funding support a particular area of program research or the work of a named faculty member, with the program director determining the specific allocation.

Current Sponsors & Partners

The Center has an established track record of industry and government partnerships:

- **Core Members:** Torc Robotics, Accenture
- **Associate Members:** NRI
- **Federal Sponsors:** DARPA, National Science Foundation
- **Additional Sponsors:** Ford, Future of Life Institute, Siemens, Open Philanthropy, GE Global Research, IBM
- **Affiliated Organizations:** Stanford HAI, Stanford Existential Risks Initiative (SERI)

Why Partner With Us

Joining the Stanford Center for AI Safety as an industry partner provides a unique opportunity to shape the future of safe AI development while accessing world-class research and talent:

- **Direct influence on research direction.** Strategic members collaborate with faculty to define flagship research projects that address your organization's most pressing AI safety challenges.
- **Access to top-tier talent.** Recruit from a pipeline of 30+ graduate students and postdoctoral scholars trained at the intersection of AI, safety, and responsible deployment.
- **Regulatory readiness.** Our research across all five pillars—from formal verification to governance frameworks—directly supports compliance with emerging AI regulations worldwide.
- **Foundation model safety expertise.** With faculty who have built and evaluated frontier models (including former Anthropic researchers), the Center offers unique insights into LLM safety, alignment, and deployment risks.
- **Brand positioning.** Association with Stanford's AI safety research signals your organization's commitment to responsible AI to regulators, customers, and the public.
- **Cross-industry insights.** Annual meetings, seminars, and curated faculty interactions provide a forum for exchanging ideas with peers from leading technology companies and government agencies.

Next Steps

We welcome the opportunity to discuss how a partnership with the Stanford Center for AI Safety can serve your organization's goals. To begin a conversation:

- Schedule an introductory call with the Executive Director to discuss alignment between your organization's needs and the Center's research agenda.
- Request a detailed research portfolio aligned to your specific use cases and interest areas.

CONTACT

Kiana Jafari, Executive Director

Stanford Center for AI Safety

Stanford University

kjafari@stanford.edu

For Stanford policies on Industrial Affiliate Programs, please visit: doresearch.stanford.edu