The Alignment Project

Funded Research Projects

Project Lead(s)	Organisation	Project Description
Sahar Abdelnabi	ELLIS Institute Tübingen	Foundations of Situational Awareness Evaluation and Control
Rico Angell	New York University	What Can Mr. Hyde Tell Us About Dr. Jekyll? Efficiently Evaluating and Red-teaming “Safe” Models with Unsafe Alter Egos
Yoshua Bengio	LawZero	Designing a safety case and latent variable model for the non-agentic Scientist AI.
Dirk Bergemann, Stephen Morris	Yale University	Our project aims to develop a unified framework for analyzing how AI systems might be jointly aligned and controlled via the language of mechanism and information design.
Eva Silverstein	Stanford University	Precision optimization and symmetry breaking for improving AI predictability
Jason Brown	Anti Entropy	Investigating the potential for AI debate to autonomously provide meaningful oversight on difficult-to-evaluate safety-relevant tasks.
Daniele Condorelli	University of Warwick	How does AI learn to play games? Theory and experiments
Diogo De Lucena	Flourishing Future Foundation	When Models Refuse to Be Steered: Understanding Self-Monitoring and Alignment Drives
John Duchi	Stanford University	(Mis)aligned preferences: what data is needed for agents to reflect honest human preferences
Max Duchi	Ashgro Inc	Develop and leverage training dynamics theory for AI alignment through early detection of phase transitions and stagewise learning dynamics.
Andrew Duncan	Imperial College London	PRISM: Probabilistic Rare-event Inference for Safety of Models
Alexander Gietelink Oldenziel	Ashgro Inc	Loss Landscape and Training Dynamics for (ReLU) Neural Networks
Venkatesan Guruswami	University of California, Berkeley	Error-correction and Pseudorandomness in Alignment: Watermarks and Backdoors
Jacob Hilton	Alignment Research Center	Low probability estimation: mechanistically estimating the probability of rare failures in machine learning systems
Ari Holtzman	University of Chicago	Hyperstition in LLMs: Mapping how narratives become model policies, probing their persistence after removal, and building tamper-evident tripwires against self-fulfilling AI behaviour.
Jesse Hoogland	Timaeus Research	Patterning: Singular Learning Theory for Alignment
Xiaowei Huang	University of Liverpool	Rare-Event Estimation in Large Language Models via Subset Simulation and Contrastive Unalignment
Rubi Hudson	Principles of Intelligence	Constructing robust incentivizes for AI agents to act cautiously
Yang Cai, Nima Haghpanah, Nicole Immorlica, Elliot Liponowski, and Doron Ravid	Yale University and University of Michigan	Is the AI Misaligned or Misinformed? Applying and further developing the strategic communication framework, we plan to study how a user can productively interact with an AI agent who may be misinformed or may have conflicting interests with the user.

Project Lead(s)	Organisation	Project Description
Pavel Izmailov	New York University	Weak-to-Strong Consistency Reinforcement Learning
Bei Jiang	University of Alberta	Adaptive Tilting to Quantify Tail Risks in Aligned Language Models
Zhijing Jin	University of Toronto	Formal verification and Empirical Testing of Translucency in Multi-Agent AI Systems
Zhijing Jin	University of Toronto	Game-Theoretic Safety Guarantees for Advanced AI Systems
Younesse Kaddar	Recursive Safeguarding Ltd	Bayesian Scientist AI for LLM Agent Alignment: GFlowNet-sampled plan abstractions and white-Box monitors to mitigate drift and deception.
Yonatan Kahn, Noam Levi, Andrew Larkoski	University of Toronto	Scaling laws, data distributions, and learning dynamics: simulated high-energy physics data as a benchmark for data in the wild
Risi Kondor	University of Chicago	Probabilistic Agentic Foundations for Mesa-Optimization and Emergent Misalignment: formalising how hidden objectives emerge within agents and designing strategies to mitigate them.
Linglong Kong	University of Alberta	Sample-Efficient Online Fine-Tuning Against Resistant Behaviors: Statistical Foundations for Post-Training Alignment
Matthew Levy	London School of Economics and Political Science	Competition and the AI Control Problem: Mechanism Design with Opaque Agents
Annie Liang	Northwestern University	Friend or Foe: Delegating to an AI whose Alignment is Unknown
Austin Meek	Center for AI Safety	Steering Models Towards Safer Reasoning
Alexandros Psomas	Purdue University	Beyond Nash: Pseudo-Equilibria for Robust AI Alignment.
Aditi Raghunathan	Carnegie Mellon University	Enhanced Cyber Control Evaluations: An Empirical and Learning-Theoretic Approach
Gabriel Recchia	Anti Entropy	Synthetic benchmarks enabling controlled experiments on how argument types and evidence structures affect judge accuracy in debate-based AI oversight.
Gabriel Recchia	Anti Entropy	Measuring how debate assistance, annotator expertise, and time constraints affect human feedback accuracy, including the effect of aggregation methods and additional judges.
Erin Robertson	Arcadia Impact	Building alignment research capacity in collaboration with AISI alignment team including initial focus areas: scalable oversight and learning theory
Fernando Rosas	Ashgro Inc	Cybernetic Foundations for Aligned Agency: From Internal Models to Verifiable Control
Shirin Saeedi Bidokhti	University of Pennsylvania	Strategic Behavior and Manipulation in Large Language Models: Experiments, Theory, and Regulatory Mechanisms
Kai Sandbrink	University College London and University of Oxford	Alignment Training Dynamics
Rahul Santhanam	University of Oxford	A Theoretical Framework for Machine Understanding
Andrew Saxe	University College London	Theoretical Model Organisms of Misalignment: Learning Dynamics of Misalignment and Theory-Grounded Mitigation

Project Lead(s)	Organisation	Project Description
Rocco Servedio, Xi Chen, Anindya De	Columbia University	Discovering rare harmful behaviors exhibited by high-dimensional AI systems
Adam Shai	Simplex	Simplex: A Principled Science of Interpretability
Alex Smolin, Piotr Dworczak	Toulouse School of Economics	Designing Human-AI Interactions: Misalignment, Robustness, and Trust
Dawn Song	UC Berkeley Center for Responsible, Decentralized Intelligence	Model Organisms of Reward Hacking and Long-Horizon Deception for Testing RL Alignment
Inbal Talgam-Cohen	Tel Aviv University	Mitigating Moral Hazard in AI using Contracts
Louis Thomson	Independent Researcher	Investigating Dynamics of Agentic Monitoring in AI Control
Vinod Vaikuntanathan	Massachusetts Institute of Technology	Verification, Robustness and Alignment through a Cryptographic Lens
Vinod Vaikuntanathan	Massachusetts Institute of Technology	Backdoors and Cryptographic Hardness in Neural Networks
Benjamin Van Roy	Stanford University	A Mathematical Model of Misalignment. We will formulate a general mathematical model of how a training process can give rise to a misaligned AI system that poses catastrophic risk.
Joan Velja	University of Oxford	Adversarial Elicitation Theory
Carmine Ventre	King's College London	The Implementability-Capability Frontier: Mechanism Design for Robust AI Alignment
Susan Wei	Monash University	Probabilistic Interpretability of Transformers
Ryan Williams	Massachusetts Institute of Technology	Exploring The Role of Relativization and Problem Extensions in AI Safety
Cole Wyeth	University of Waterloo	Safe Superintelligence via AIXI Agent Foundations
Bin Yu	University of California, Berkeley	The Geometry of Risk: Using Model Complexity to Improve Predictions of Catastrophic AI Failures
Or Zamir	Tel Aviv University	Provable and Rigorous Notions of Security for Machine Learning via Cryptography