Foresight Institute 2024 Workshop

2024 AGI:Cryptography, Security, and Multipolar Scenarios Workshop

THE INSTITUTE, SAN FRANCISCO, CA, USA

14 & 15 May, 2024

Table of Contents

Workshop Synopsis

To help AI development benefit humanity, the Foresight Institute has hosted numerous workshops, including AGI & Great Powers, AGI: Toward Cooperation, alongside technical meetings in 2022 and 2023 focusing on cryptography, security, and AI. Recently, we launched a Grants Program to fund work on, among other things, AI security risks, cryptographic tools for safe AI, and beneficial multipolar AI scenarios.

Whilst these efforts have highlighted promising projects, the intersection of AI with cryptography and security remains nascent. This workshop brought together leading researchers, entrepreneurs, and funders to explore tools and architectures facilitating human/AI cooperation, addressing three key questions:

- Which goals should we prioritize to make a multipolar AGI scenario safe and beneficial?
- Are there any approaches in cryptography, security, and auxiliary fields that could help address these goals?
- Are there general R&D factors (funding, data acquisition, coordination etc) that make it difficult to apply these approaches to the goals?

Held over two days at The Institute, Salesforce Tower, San Francisco, the workshop gathered sixty experts working on cryptography, security, and AI. It featured rapid keynotes followed by working groups, focused on themes including AI infosecurity, coordination architectures, privacy-enhancing technologies, game theory for multi-agent scenarios, and mechanisms for positive-sum dynamics.

This report contains summaries and recordings of the presentations and the ensuing project collaborations, accessible via the play icons in the images.

We extend our heartfelt gratitude to all participants for their contributions and collaboration. A special thank you goes to our sponsors, Filecoin Foundation, Protocol Labs, Agoric, and Subspace Network,for subsidizing the attendance of junior researchers. Your support was vital for the success of this workshop.

We look forward to next year's workshop to review and make further progress on these areas. If you are interested in contributing in this area as a researcher, practitioner, or funder, we welcome you to reach out.

Best regards,

Allison Duettmann
CEO, FORESIGHT INSTITUTE
a@foresight.org

About Foresight Institute

Founded in 1986, Foresight Institute supports the beneficial development of high-impact technology to make great futures more likely. We focus on science and technology that is too early-stage or interdisciplinary for legacy institutions to support, such as longevity biotechnology, molecular machines, brain-computer interfaces, multipolar AI, or space exploration. We award prizes, offer grants, support fellows, and host conferences to accelerate progress toward flourishing futures and mitigate associated risks.

Participants

Abhishek Singh
MIT

Adam Gleave
FAR.AI

Aleksandra Singer
ALTOS LABS

Anthony Aguirre
FUTURE OF LIFE INSTITUTE

Asvin G.
UNIVERSITY OF TORONTO

Austin Liu
CHAO SOCIETY

Bogdan Ionut Cirstea
INDEPENDENT

Brandon Goldman
LIONHEART VC

Brandon Sayler
UNIVERSITY OF PENNSYLVANIA

Brian Behlendorf
MOZILLA

Chris Akin
APOLLO AI

Christopher Hart
BLOCKCHAIN COMMONS

Christopher Lakin
INDEPENDENT

Claire Short
ATHENA

Colleen McKenzie
AI OBJECTIVES INSTITUTE

Craig Quiter
DEEPDRIVE.IO

Participants

Dan Girshovich
TOOLS FOR HUMANITY

David Abecassis
ACCENTURE

David Bloomin
PLATYPUS AI

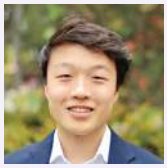
Dean Tribble
AGORIC

Divya Siddarth
COLLECTIVE INTELLIGENCE PROJECT

Dmitrii Usynin
IMPERIAL COLLEGE LONDON

Dustin Li
ANTHROPIC

Esben Kran
APART RESEARCH



Ethan Shaotran
REINVENT FUTURES

Evan Miyazono
ATLAS COMPUTING

Fazl Barez
OXFORD UNIVERSITY

Jacob Lagerros
LIGHTCONE INFRASTRUCTURE

James Petrie
OXFORD UNIVERSITY

Janna Lu
MERCATUS CENTER

Jeffrey Ladish
PALISADE RESEARCH

Jeremiah Wagstaff
HUMAIC LABS

Participants

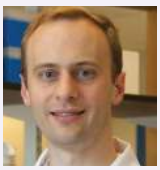
José Andrade
GENPACT

Josh Tan
METAGOV

Kaliya Young
IDENTITY WOMAN

Kenneth Bruskiewicz
SIMON FRASER UNIVERSITY

Keri Warr
ANTHROPIC



Kevin Esvelt
MIT

Labhesh Patel
HUMAIC LABS

Lisa Thiergart
MIRI

Marta Belcher
FILECOIN FOUNDATION

Matjaz Leonardis
OXFORD UNIVERSITY

Matthew McAteer
5CUBELABS

Matt Slater
STATELESS VENTURES

Max Reddel
ICFG

Michael Nielsen
ASTERA INSTITUTE

Mimee Xu
NEW YORK UNIVERSITY

Naomi Brockwell
NB TV

Participants

Niccolo Zanichelli
UNIVERSITY OF PARMA

Peter Norvig
STANFORD

Philip Chen
LIONHEART VC

Portia Murray
AI OBJECTIVES INSTITUTE

Rachel Freedman
UC BERKELEY

Richard Mallah
CENTER FOR AI RISK MANAGEMENT & ALIGNMENT

Richard Ngo
OPENAI

Ryan Singer
VEX

Yudhister Kumar
MATS

Sam Holton
UC DAVIS

Samuel Bowman
ANTHROPIC

Steven Stone
ZERO LABS

Thee Ho
ATHENA

Vassil Tashev
INDEPENDENT

Vickram Premakumar
AE STUDIO

Vishal Maini
MYTHOS VENTURES

Yonatan Ben Shimon
MATCHBOXDAO

Xiaohu Zhu
CENTER FOR SAFE AGI

Participant Group Photo

Keynote Presentations

AI Threat Models, Hacking, Deception, and Manipulation

Jeffrey Ladish, Palisade Research

In this talk, Ladish discusses the increasing sophistication and proliferation of deepfake technology, which allows AI to mimic human voices and faces, and its potential for widespread deception. He demonstrates how his company, Palisade Research, has created a programme that can generate deepfake voices from audio samples, fully mimicking an individual's voice and speech patterns. He argues that this increasingly capable technology is and will be used to spread fake information, manipulate elections or markets, create deepfake pornography, and generate fake endorsements from actors or organizations. His conclusion emphasizes the importance of establishing strong, immediate, and effective countermeasures to counter these emergent technologies and their misuse in creating deceptive content.

Harnessing the Heft: Securing LLM Weights

Keri Warr, Anthropic

Warr, a security engineer at Anthropic, discusses cybersecurity controls for safe model training, focusing on protecting sensitive assets like model weights or snapshots. He explains that large language models, being terabytes in size, are vulnerable to theft, but stealing a small subset of a snapshot is ineffective. To address this vulnerability, Warr proposes a security control called rate limiting egress proxy, which restricts data bandwidth in and out of a cluster, preventing model weight exfiltration without hindering productivity. This control complements existing network filtering egress proxy defenses and covers all network traffic. Warr concludes by emphasizing the importance of threat modeling and concludes by encouraging collaboration and thinking from first principles to safeguard infrastructure and assets.

Multipolar Concerns for Technical AI Governance

Lisa Thiergart, MIRI

This talk discusses MIRI's agenda as well as the discrepancies between unipolar and multipolar AI scenarios. Thiergart explains that MIRI believe that technical alignment is unlikely to be achieved pre-AGI, so they are instead working towards getting good regulation to buy time towards building the right technical solutions. Thiergart, part of a new technical governance research team whose mission is to increase the probability of effective US AI regulation, explores the differences between multipolar and unipolar scenarios for technical governance. In the unipolar case, she notes that it is more likely to have a unified regulatory framework, while in the multipolar case, different players would be developing varying regulatory frameworks. This points towards a higher need for international coordination and multilateral regulatory organizations. Furthermore, in this case, positive opportunities include more diversified defense security systems and potentially faster innovation, but at the risk of an arms race or a race to the bottom.

AI, Decentralization, and Regulating Emerging Technologies

Marta Belcher, Filecoin Foundation

In this presentation, Belcher discusses the problem of balancing intellectual property rights, particularly copyright and civil liberties, with the need to regulate emergent AI and ML capabilities. She indicates the ongoing debate between regulating AI technologies themselves versus regulating the activities involving AI, emphasizing that the latter is crucial for preserving intellectual property rights. Furthermore, she points out that future regulation may lead to legal precedents requiring compensation for those whose work is ingested into AI models. However, she believes that protecting the ability for machines to learn from the web is critical and points out that copyright law is not intended to protect creators from having their works displayed or performed. Finally, she mentions the complexities of holding individuals accountable for AI misuse and underscores the need for cohesive international regulations to address the global impact of AI advancements.

Neartermist safety: Incentive-compatible Directions for Large Model Oversight

Mimee Xu, New York University

This talk focuses on the difficulties of AI oversight measurement, particularly evaluations, due to the growing data depletion problem. Xu, an AI and privacy expert, explains that the intense market competition for public data to train pre-trained models, driven by the increasing size of models and their need for more data, has led to a shortage of public data and an increased cost of private data. This issue is compounded by the increasing complexity and domain-specificity of evaluations, making them unsustainable in the long run. Xu concludes by suggesting that scalable evaluation science requires reconsideration to develop more sustainable evaluation methods.

How do Multipolar Scenarios get Exacerbated by AI?

Philip Chen, Lionheart VC

Chen introduces Lionheart Ventures, an AI-focused venture capital fund that invests in for-profit AI safety companies across various domains. As the lead of impact work at Lionheart Ventures, Chen is developing the impact MOIC (multiple on invested capital) framework to quantify the counterfactual impact of investing in for-profit AI safety companies compared to donating to AI safety nonprofits. He then dives into cascading AI systemic risks, which he categorized into two types: AI exacerbating current existential risks and AI posing unforeseen risks that could cascade into other systems, potentially leading to existential threats. Chen illustrates how AI could contribute to economic inequality, social unrest, and the erosion of the social fabric. Chen explains that to mitigate these risks, Lionheart Ventures collaborates with the Total Portfolio Project (TPP) to assess the existential risk of their AI portfolio, aiming to reduce the potential for AI-caused catastrophic events.

Wargaming for Possible TAI Futures

Portia Murray, AI Objectives Institute

Murray underscores the importance of transformative simulations research (TSR) mitigating AI risks through its capacity to predict potentially adverse outcomes. TSR involves building operational simulations or "war games" that facilitate a better understanding and characterisation of AI's impact, which help define real-life problems. Furthermore, she argues that TSR can help aid in revealing emergent phenomena, especially in multipolar games with complex interactions and dynamics. In conclusion, Murray emphasizes the importance of developing safe AI, and TSR's role in predicting and thereby preventing negative outcomes.

What should Multi-Agent Alignment aim to Achieve

Richard Ngo, OpenAI

In this talk, Ladish discusses the increasing sophistication and proliferation of deepfake technology, which allows AI to mimic human voices and faces, and its potential for widespread deception. He demonstrates how his company, Palisade Research, has created a programme that can generate deepfake voices from audio samples, fully mimicking an individual's voice and speech patterns. He argues that this increasingly capable technology is and will be used to spread fake information, manipulate elections or markets, create deepfake pornography, and generate fake endorsements from actors or organizations. His conclusion emphasizes the importance of establishing strong, immediate, and effective countermeasures to counter these emergent technologies and their misuse in creating deceptive content.

Hardware Governance

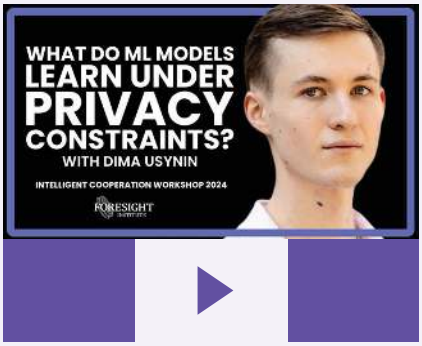
Anthony Aguirre, Future of Life Institute

In this talk, Aguirre discusses the future of compute governance, arguing that controlling the limited supply chain for AI hardware is more feasible than attempting to control easily proliferated software. He proposes a two-part solution to compute governance: first, establishing governance contracts and regulations that nations can agree upon; and second, implementing a verification and enforcement layer using hardware and software cryptographic security measures to ensure governance is enforceable. Aguirre illustrates the effectiveness of controlled hardware by referencing iPhones that can be remotely deactivated if stolen, suggesting that similar measures could be used to limit the usage of powerful hardware in sensitive locations. He concludes that compute governance is necessary to ensure that humanity remains in control of AI development and deployment.

Collective Intelligence

Divya Siddarth, Collective Intelligence Project

Siddarth explores the potential of collective intelligence (CI) for AI governance. She believes that people should have agency and choice over their decisions, and that decentralizing power is often better for ensuring consistent agency over their lives. However, she notes that making good decentralized decisions is challenging and that the concentration of power is only likely to increase as transformative technologies advance. To explore how people can have more input and agency over these technologies, Siddarth introduces the Collective Intelligence Project, which runs experiments in this area. They've worked with OpenAI to evaluate systems, the Taiwanese ministry to regulate AI in elections, and Anthropic to retrain Claude on a constitution written by a thousand randomly selected people. Through these experiments, Siddarth has learnt that CI works well when trying to elicit information from people that isn't already available. She believes that experiments involving real people and institutions are necessary to determine when and how CI can lead to better decisions.



How do Models Learn when there are Privacy Constraints?

Dmitrii Usynin, Imperial College London

In this talk, Usynin explores the trade-offs between increasing data usage for improved machine learning model training and adhering to data privacy regulations. He discusses how anonymization techniques and decentralized machine learning training, such as Federated Learning (FL), have historically been used to address privacy concerns, but underscores that these methods are still vulnerable to privacy attacks. Usynin argues that FL, when enhanced with privacy measures, can be an effective mechanism for protecting data and creating trustworthy machine learning models. He then introduces differential privacy as a quantifiable method which bounds the amount of information a model learns, regardless of data, architecture, or other constraints, and concludes by emphasizing the importance of privacy-preserving machine learning in protecting user data while allowing for effective model training.

Challenges and Solutions for AI Security in the Age of Multipolar AGI

Esben Kran, Apart Research

In this talk, Kran sheds light on the rationale and actions for AI Security. He discusses the importance of coordination between AI labs and emphasizes the need for international cooperation and coordination, given the geographically centralized nature of AGI development. He also introduces Apart Research, who, through their research hackathons and accelerator programme, aim to create a remote global community focused on AI Security. Lastly, Kran touches upon some of Apart Research's projects that are particularly relevant to multipolar AGI security, such as geopolitical simulation, multi-agent security, and secure and private machine learning. These initiatives showcase the organization's commitment to ensuring the safe and responsible development of advanced AI systems in a collaborative, international context.

Securing Human Review with AI of AI

Evan Miyazono, Atlas Computing

Miyazono discusses reducing AI risk via several user-centric approaches, which allow users, engineers, and regulators to define what an AI system should achieve without worrying about its execution. Safeguarded AI incorporates tools that assist users in specifying the desired properties of outputs, so that proof generators can create objective evidence to confirm that proposed solutions meet all desired properties. Near-term applications might include ensuring that generated software programs do not crash and that designed molecules do not interact negatively with specified parts of a cell, to minimize risks posed by both human and AI actors, now and in the future.

How to Prevent LLMs from Relearning Undesired Concepts

Fazl Barez, Oxford University

In this talk, Barez explores the process of unlearning, which involves removing harmful responses, copyrighted data, or reducing hazardous capabilities that a machine learning model may have learnt, without the need for retraining. He explains the basic methodology of unlearning, which involves training the model again with basic knowledge whilst aiming to forget other parts. However, he also highlights the potential for models to relearn such removed concepts and abilities, and explores into methods to prevent this, such as pruning neurons and adjusting models. He presents experimental results that demonstrate the importance of certain neurons in a model's ability to learn and how unlearning can redistribute these neurons to earlier layers. In conclusion, Barez suggests that while unlearning may be a useful tool, it has its limitations and raises concerns over the general robustness of the method.

A Bottom-up Approach to AGI Alignment for a Massively Multipolar Future

Jeremiah Wagstaff, Humaic Labs

Wagstaff explores how AGI can be designed or aligned to be more helpful, assuming the nearness of both aligned and misaligned AGI, and how we can protect ourselves from the latter. He proposes using a responsible moral agent to create AI as an autonomous moral agent that respects human autonomy, seeking self-actualisation, self-sufficiency, and self-governance while respecting the autonomy of all moral agents – achievable through constitutional AI with reinforcement learning from AI feedback. The talk then explores developing mechanisms for autonomous aligned AGI and humans to coordinate against misaligned AGI, suggesting the use of decentralized identity frameworks with a verifiable registry, ideally on a secure public blockchain, as a platform for coordination. In conclusion, Wagstaff envisions a global network of aligned AI and humans working together in a massively multi-polar scenario to safeguard against misaligned AGI.

Smart Contracts and AI

Dean Tribble, Agoric

Tribble explores the interplay between smart contracts and AI, discussing their evolution, functions, and the significance of blockchains in enabling their cooperation. He describes smart contracts as contract-like arrangements expressed in code that enforce the terms of the agreement through source code execution. Tribble emphasizes that the fundamental consequence of smart contracts is enabling secure cooperation among strangers by enforcing the rules of the game via a mechanism which orchestrates interactions between various components and agents. Finally, he examines the integration of smart contracts with AI, focusing on their important role in ensuring safe interactions and collaborations among diverse AI entities whilst maintaining security and privacy.

AI as Public Infrastructure

Josh Tan, MetaGov

In this talk, Tan explores the idea of transforming AI into a "shared scientific endeavor" and becoming a part of the human experience, rather than being something to fear or control. He introduces the concept of Public AI, which refers to publicly funded and governed AI models and applications accessible to the public. Tan argues that Public AI offers greater equity, accessibility, and safety compared to private or open-source AI, and suggests that it could develop as a national agency, a policy scheme, or a decentralized network of publicly funded services. While Public AI is currently a provocation within various societal, academic, and policy circles, it is gaining traction in countries such as the US, UK, Sweden, UAE, Japan, and India. To further investigate the narratives surrounding Public AI, Tan proposes launching a seminar series and invites those interested in understanding the political power of their narrative to engage.

AI: Will it Help Solve our Data Mayhem Problem or Make it Worse?

Steven Stone, Zero Labs

Stone, a data security expert, discusses the importance of understanding data vulnerabilities and the role of AI in data security strategies. He stresses the need for organizations to prepare for significant data growth in the next five years, with his research showing that a typical organization's data of 240 backend terabytes is growing by 42% every 18 months, leading to a 7x increase in the next five years. Given our growing need to adapt defense strategies accordingly, Stone highlights the potential of generative AI in defending data and enhancing security measures within organizations, underscoring its benefits in effectively identifying, classifying, and tracking data movements. Finally, he explores how AI can be used to adapt defense strategies to keep pace with the rapidly expanding data landscape.

Project Proposals

Inspired by the challenges pointed out during the keynote presentations, working groups were formed to address problems of common interests.

Wargaming Race Dynamics in an AGI Launch Scenario

- **David Abecassis**, Accenture
- **Vickram Premakumar**, AE Studio

This working group explored using advanced wargaming techniques to address risks arising from competitive dynamics between frontier AI model developers and nations. They proposed enhancing conventional wargaming mechanisms by incorporating LLMs as players, adjudicators, and customized adversaries. The group aimed to accelerate the iterative development cycle of wargames and enable large-scale scenario analysis. They also conceptualized using smart contracts or cryptography to construct secure wargames for rivals, generating high-level outcomes to promote collaborative decision-making. The project's goal was to influence decision-making of high-agency clients, including frontier model developers, regulators, and the general public. They emphasized the potential for unclassified games to shape public opinion and drive pro-social outcomes through regulatory or economic pressure.

Preventing AI Misuse

- **Brian Behlendorf**, Mozilla
- **Fazl Barez**, Oxford University
- **Janna Lu**, Mercatus Center

This working group focused on preventing the misuse of AI by bad actors for activities such as creating viruses or nuclear weapons. The team proposed working within existing institutional regulatory frameworks, rather than attempting to regulate AI models' capabilities directly. They suggested three strategies: extending the current regulatory regime, open-sourcing guardrails to detect and prevent AI misuse, and using AI for objective evaluations of AI models and their end use. The group emphasized the importance of creating legislation that outlaws specific types of AI misuse, similar to Connecticut's ban on deepfakes. They proposed simulating potential AI misuse scenarios to help regulators anticipate and prevent catastrophic outcomes. The team highlighted that success would involve making it difficult to use AI for creating weapons or other harmful applications. They also recognized the need for a responsive team to quickly address vulnerabilities as they are discovered, suggesting regular checkpoints for such teams.

Better Incentives

- **Brandon Goldman**, Lionheart VC
- **Dmitrii Usynin**, Imperial College London
- **Mimee Xu**, New York University
- **Niccolo Zanichelli**, University of Parma
- **Richard Mallah**, Center for AI Risk Management & Alignment

This working group proposed a public-private partnership platform to address misaligned incentives in AI development, focusing on data bottlenecks and evaluations. The platform, using secure multi-party computation, would enable privacy-preserving model evaluations, data appraisal, and evaluation of evaluations without exposing model weights. It includes secure evaluation of data on models, secure model evaluation for private benchmarks, and auditable data with confidential computations. The project aims to provide secure oversight through safety probes and real-time evaluations to inform policy, funded by bonds and a data tax. The group emphasized aligning incentives between stakeholders and addressing the current lack of effective oversight mechanisms, particularly for frontier models. Their scalable solution is designed to facilitate collaboration between large businesses and governments in the rapidly evolving AI landscape.

Systemic Risk of AI

- **Brandon Sayler,** University of Pennsylvania
- **Colleen McKenzie,** AI Objectives Institute
- **Jeremiah Wagstaff,** Humaic Labs
- **Josh Tan,** MetaGov
- **Milan Griffes,** Lionheart Ventures
- **Philip Chen,** Lionheart VC

This working group developed a comprehensive approach to identify and address cascading systemic risks associated with AI development. They created a taxonomy of interconnected risks spanning cybersecurity, economics, geopolitics, and social dynamics, proposing the creation of "risk observatories" to monitor early warning signs of potential problems. This approach aims to centralize and automate risk detection systems, allowing quick identification and addressing of issues as they arise. Key risk areas included AI-driven job displacement, erosion of trust due to deepfakes, cybersecurity threats, and potential AI-enabled totalitarianism. Their analysis culminated in a flow chart illustrating how issues could lead to three major endpoints: extinction, excessive state control, or anarchy. By implementing their proposed risk observatory system, the group aims to mitigate these systemic risks and prevent cascading negative outcomes.

Differential Cyber Defense

- **Aleksandra Singer,** Altos Labs
- **Austin Liu,** Chao Society
- **Esben Kran,** Apart Research
- **Evan Miyazono,** Atlas Computing
- **Keri Warr,** Anthropic
- **Matt Slater,** Stateless Ventures
- **Ryan Singer,** VEX

This working group focused on cyber defense, developing an approach to secure the future of cybersecurity against AI-associated risks. They proposed creating an Epoch AI-like research group dedicated to cyber posture forecasting in an AGI future, aiming to provide trustworthy, publicly available data to demonstrate future risks and increase awareness among cybersecurity policy professionals. The group emphasized empowering defense over offense to ensure greater stability, suggesting the use of machine learning approaches to improve the overall internet's security posture. Their plan involves building AI tools to identify vulnerabilities in open-source software and create patches to secure it. The approach includes assembling a team of security experts, fine-tuning frontier models, and developing automated tools for vulnerability identification and responsible disclosure. By encouraging widespread adoption of this platform, the group aims to facilitate better decision-making and contribute to a more secure cyber future in the age of AGI.

Robust International Coordination Mechanisms

- **José Andrade,** Genpact
- **Lisa Thiergart,** MIRI
- **Max Reddel,** ICFG
- **Vassil Tashev,** Independent
- **Yonatan Ben Shimon,** MatchBoxDAO

This working group explored potential approaches to international AI governance and coordination mechanisms to address catastrophic and existential risks from AI. They proposed a three-tiered model for different types of AI systems. High-risk models would require an international compute resource akin to CERN, with stringent security standards. Medium-risk models would be subject to international monitoring and classification, while low-risk models could be deployed under national jurisdiction with standard compute governance. The group suggested forming an IAEA-like global institution to establish risk benchmarks and governance standards. They proposed hosting the international compute resource in a politically neutral location, leveraging compute governance for enforcement. The group aimed to create incentives for both major powers and smaller actors to participate, such as access to frontier models and institutionalized benefit sharing. However, challenges remained regarding funding, construction, and resource allocation for this body, as well as international governance of military AI applications.

International Evaluation Standards

- **Xiaohu Zhu,** Center for Safe AGI
- **Yudhister Kumar,** MATS

This working group focused on AI understanding through proof-based safety. They proposed conducting theoretical research to provide different levels of probability for safety properties based on proof theory and model theory. The team aims to build a solid foundation for multiple AI systems and develop a domain-specific language for specifying and proving safety properties. They plan to use category theory, type theory, and linear logic to monitor resource usage during the proof process. This approach represents a new systematic way to conduct safe AGI research, with the goal of detecting unsafe models early in development or deployment. The team believes that proof is key to understanding AI behavior and establishing trust. Their midterm and final goal is to investigate the safety dynamics of state-of-the-art large language models, though they currently lack experts familiar with both mathematical logic and safe AGI. The risks associated with this approach include the potentially longer learning curve for people to understand and invent new models.

Implementation of Workshop Ideas: Funding Opportunities

This workshop marked nearly a year since Foresight Institute launched its AI Safety Grant. This grant supports work in various under-explored areas, such as:

CRYPTOGRAPHY AND SECURITY APPROACHES FOR INFOSEC AND AI SECURITY

Exploring the potential of cryptography and security technologies for securing AI systems:

- Computer security to help with AI Infosecurity or approaches for scaling up security techniques to potentially apply to more advanced AI systems
- Cryptographic and auxiliary techniques for building coordination/governance architectures across different AI(-building) entities
- Privacy-preserving verification/evaluation techniques
- Other concrete approaches in this area

SAFE MULTIPOLAR AI SCENARIOS AND MULTI-AGENT GAMES

Exploring the potential of safe Multipolar AI scenarios:

- Multi-agent game simulations or game theory
- Scenarios avoiding collusion and deception, and pareto-preferred and positive-sum dynamics
- Approaches for tackling principal agent problems in multipolar systems
- Other concrete approaches in this area

Many of the approaches presented in this workshop closely align with the goals of our AI Safety Grant. We encourage those working on relevant approaches to consider applying to the grant and look forward to seeing how much progress can be made until we invite you to regather for the 2025 edition of this workshop.

Our Grantees Working within the Areas presented at this Workshop

This workshop marked nearly a year since Foresight Institute launched its AI Safety Grant. This grant supports work in various under-explored areas, such as:

WITHIN THE AREA OF SECURITY AND CRYPTOGRAPHY

Abhishek Singh
MIT

AI SecureOps: GenAI and LLM Security Training for Enterprises

Professional GenAI Security Training, tailored for securing enterprise LLM services and promoting the safe integration of public GenAI services for in-house operations. This training adopts a Capture The Flag (CTF) style and adversary simulation exercises, covering a spectrum from the fundamentals of LLM security to the application of custom data for developing AI-based security agents. Attendees will be provided with a playground application to try out the labs and CTFs.

Adam Gleave
FAR.AI

A Science of AI Robustness

We are penetration testing AI systems to predict the robustness of current and future systems to sophisticated threat actors such as nation states or future misaligned AIs. Specifically, we will test AI systems of varying scale in order to identify and empirically validate scaling laws for security: for example, how much compute Z does an adversary need to exploit an AI system of size X that underwent Y steps of adversarial training.

These security scaling laws will enable defenders to choose the most effective training methods, immediately improving the security of AI systems. Moreover, scaling laws will enable us to answer key questions, such as whether current methods are sufficient to adequately secure transformative AI systems of the future? If not, scaling laws will enable us to evaluate novel defense techniques, accelerating progress on AI security.

Esben Kran
APART RESEARCH

Systematic Evaluation of Offensive Cyber Capabilities of Large Language Models

We want to automatically evaluate offensive cyber capabilities of large language models (LLMs) in a stateful and realistic manner by leveraging capture-the-flag scenarios. This will tell us both about the level of risk from misuse of LLMs by cybercriminals, as well as about potentially extreme risks from misaligned advanced LLMs which may attempt to evade human control by hacking their own servers.

Florian Tramèr
ETH ZURICH

AI safety research

We plan to formalize appropriate threat models for using cryptography to secure AI applications, e.g., for defending against adversarial examples or for model watermarking. In the process, we will show new attacks on many existing schemes that were likely overlooked due to a lack of threat modeling.

Harriet Farlow
MILEVA SECURITY

Likelihood Analysis in AI Security

This project seeks to fill a crucial gap in AI security research by quantifying the likelihood of AI incidents. Likelihood is a key parameter in risk assessment best practices in other security disciplines (risk = severity x likelihood), however in the case of AI security there is a growing need to model risk, and while there is much research into severity there is very little in likelihood. Drawing from established risk assessment methodologies in cybersecurity, the study aims to construct a robust framework for evaluating and mitigating AI security risks, thus shedding light on the elusive dimension of risk associated with AI vulnerabilities.

.

WITHIN THE AREA OF MULTIPOLAR AI SCENARIOS

David Bloomin
PLATYPUS AI

Open-Ended Learning in Socially Complex Multi Agent Environments

Metta Learning is an open-source research project that investigates the emergence of cooperation and alignment in multi-agent AI systems. By creating a model organism for complex multi-agent gridworld environments, the project aims to study the impact of social dynamics, such as kinship and mate selection, on the learning and cooperative behaviors of AI agents.

Christopher Lakin
INDEPENDENT

Conceptual Boundaries Workshop

Conceptual Boundaries Workshop, held in Austin TX from February 10 to 12, brought together 13 leading researchers to explore formalizing the notion of boundaries in complex systems. Key outputs included new research directions using Petri nets and graph factorizations to model boundaries, a potential formal definition of "boundary protocols," and plans for a follow-up Mathematical Boundaries Workshop in April 2024.

Keenan Pepper
INDEPENDENT RESEARCHER

Embedded Agency Playgrounds

An embedded agent is one whose cognitive machinery is a part of the environment in which it's acting to achieve goals. Current frontier AI models are not embedded, but superintelligent AI will eventually become embedded whether we like it or not, because understanding your place in the world and gaining some form of back-door access to yourself are convergently instrumental goals for many tasks. If this first happens suddenly and unexpectedly in a domain such as "the Internet" or "the physical world" that would be extremely risky. Therefore I propose to study phenomena of embedded agency in safe, mathematically simple sandbox environments. This could lead to deconfusion and experimental verification of theorized embedded agency phenomena, hopefully long before it becomes a concern in capable general-purpose models.

MATS

ML Alignment & Theory Scholars (MATS) Program

- Find and accelerate high-impact research scholars.
- Support high-impact research mentors.
- Help parallelize high-impact AI alignment research.

This grant will enable MATS to add further cooperative/multipolar AI scholars (e.g., for mentors Jesse Clifton and Caspar Oesterheld) and mentors (e.g., Jan Kulveit, Andrew Critch) to their twice-yearly program.

Nora Ammann
PIBBSS

PIBBSS Fellowship

Our fellowship aims to upskill researchers from different complex science backgrounds to pursue AI safety work, while exposing mentors and the broader community to insights from those fields. We are applying for funding that would allow us to accept several additional fellows focusing on Multipolar AI Safety to the 2024 cohort.

Dr Toby David Pilditch
TRANSFORMATIVE FUTURES INSTITUTE

Cutting through the complexity of multi-agent AI scenarios

We are pioneering a multi-agent computational model that can directly simulate multi-polar and game theoretic behaviours in AI scenarios. The developed model will enable the rigorous testing and refinement of scenarios, their underpinning assumptions, and prospective policy proposals, presenting first-of-its-kind computational analysis for multi-polar and AI safety research.

Foresight Institute

101A Clay Street, Box 185.

San Francisco, CA 94111