FORESIGHT
INSTITUTE

2023 Intelligent Cooperation Workshop: Cryptography, Security, AI

Allison Duettmann
Foresight Institute

Christine Peterson
Foresight Institute

Mark S. Miller
Agoric

The Internet Archive, San Francisco
July 10-11, 2023

Table of Contents

Foreword 3

Participants 4

About Foresight Institute 5

Workshop Chairs 6

Christine Peterson: Foresight Institute 6

Allison Duettmann: Foresight Institute 6

Mark S. Miller, Agoric 6

Workshop Format 7

Keynote Presentations 11

Presentations by workshop chairs and invited keynotes 11

Intelligent Cooperation Workshop Introduction | Allison Duettmann, Foresight Institute 11

Super Intelligent Alignment | Jan Leike, OpenAI 11

Fireside Chat on AI Alignment | Kipply Chen, Anthropic 12

Cooperation and Control in Delegation Games | Lewis Hammond, Cooperative AI Foundation 12

Interpretability and Security of AI Models | Matjaz Leonardis, University of Oxford 13

Safety First Cognitive Architectures | Brendon Wong, Cosmic 13

Scaling Deliberation | Colleen McKenzie, AI Objectives Institute 14

Trustworthy AI Challenges of Adoption of Privacy Preserving ML | Dmitrii Usynin, Creator Fund 14

Controllable AI with Open Agency Systems | Evan Miyazono, Protocol Labs 15

Bridging the Gap to the Security Field | Jamie Joyce, The Society Library 15

Environments for Empirical Embedded Agency Research | Keenan Pepper, Salesforce 16

Fireside Chat on Computational Market Places | Mark Miller, Agoric 16

Semiconductor Support for Cryptographic Innovation | Mark Davis, Crossbar Inc. 17

Fireside Chat on Crypto, AI & Civil Liberties | Marta Belcher, Filecoin Foundation 17

Predicting the Future | Avichal Garg, Electric Capital 18

AI Opportunity | Brewster Kahle, Internet Archive 18

Technical Projects for Easier Cooperation | Rick Korzekwa, AI Impacts 19

Intelligent Cooperation and Brain-Computer Interfaces | Rob Luke, AE Studio 19

Decentralized Voluntary Cooperation | Shady El Damaty, Holonym 20

AI Dilemma & Policy | Tristan Harris, Center for Humane Technologies 20

Interpretability for Safety and Alignment | Fazl Barez, Apart Research 21

Project Proposal Winners 22

Automatically Mapping Rigorous Disource 22

Personalized AI Assistant 23

Preventing Collusion 24

Other Project Presentations 25

Marginal Governance Improvement 25

Formalising Boundaries Empirically 26

Design for a Robust ID System 27

Positive Futures 28

AI Safety Grant 29

Neurotechnology, Whole Brain Emulation, and Lo-Fi Uploading for AI Safety 29

Cryptography and Security Approaches for Infosec and AI Security 30

Safe Multipolar AI Scenarios and Multi-Agent Games 30

Foreword

As AI progress is rapidly advancing, are there any areas which remain under-explored for supporting the beneficial development of this technology suite? At Foresight Institute, we believe that the intersection of cryptography, security, and AI is still nascent but, could be of fundamental importance for beneficial futures.

Foresight Institute has highlighted a few opportunities at this intersection before; in Gaming the Future, How Security and Cryptography can Aid AI Safety, and in our 2022 workshop on this topic. Our Intelligent Cooperation Workshop set out to build on these efforts by exploring two major themes:

1. How can cryptography and security technologies help secure cooperation across humans?
2. How can these technologies be extended to secure cooperation across humans and emerging AIs?

Held over two days at the Internet Archive in San Francisco, this workshop invited sixty researchers, engineers, and funders working on cryptography, security, and AI technologies to make progress on the questions above.

Introductory presentations were followed by working groups to explore promising focus areas highlighted in the presentations. This report contains summaries and recordings of the intro presentations, and ensuing project collaborations. By clicking on the play icons in the images, you can watch the corresponding presentations.

The opportunities focused on by the working groups include efforts to securely open-source alignment, avoid ML backdoors, build secure personal AI assistants, prevent collusion amongst multiple AIs, create robust ID systems, improve hardware security, and many others. Workshop attendees were given the opportunity to vote for the projects they deemed most promising.

This report also summarizes some participant perceptions with respect to societal factors influencing AI safety, such as prioritization, funding, education, recruiting, and public perception. However, given significant nuanced disagreement on most of these factors, the perceptions highlighted here do not reflect the opinion of all participants but are best seen as opinion vignettes.

We extend our heartfelt gratitude to all participants for their active collaboration, and to Foresight senior fellow Mark S. Miller for chairing the workshop. A special thank you also goes to our sponsors, including Filecoin Foundation and Agoric, for subsidizing the attendance of junior researchers. Without your support, this workshop would not have been possible.

We look forward to next year's follow-on workshop to review and further advance projects that were initiated during this workshop.

To support progress in the meantime, we have launched the AI Safety Grant Program with a focus on security and cryptography techniques for AI safety: https://foresight.org/ai-safety.

If you are interested in advancing this area –as a researcher, practitioner or funder – we welcome you to reach out.

Best regards,
Allison Duettmann
President & CEO
a@foresight.org

Participants

Aleks Singer
ALTOS LABS

Alexander Seyfert
THREESIXTY VENTURES

Allison Duettmann
FORESIGHT INSTITUTE

Avichal Garg
ELECTRIC CAPITAL

Anders Størkson Berg
KYBER ENGINES

Andrew Gritsevskiy
CAVENDISH LABS

Brewster Kahle
INTERNET ARCHIVE

Brandon Goldman
INDEPENDENT

Brendon Wong
COSMIC

Bulelani Jili
HARVARD UNIVERSITY

Cole Killian
INDEPENDENT

Christopher Lakin
CARNEGIE MELLON UNIVERSITY

Dan Schwarz
METACULUS

Dan Elton
MASS GENERAL BRIGHAM DATA SCIENCE OFFICE

David Bloomin
PLATYPUS AI

Colleen McKenzie
AI OBJECTIVES INSTITUTE

Dmitrii Usynin
CREATOR FUND

Ela Madej
FIFTY YEARS

Evan Miyazono
PROTOCOL LABS

Grant Roy
THEOREE

George Burke
INDEPENDENT

Jamie Joyce
THE SOCIETY LIBRARY

Jan Leike
OPENAI

Jason Morton
ZKONDUIT INC.

Josh Tan
METAGOV

Kamile Lukosiute
ANTHROPIC

Kipply Chen
ANTHROPIC

Keenan Pepper
SALESFORCE

Lewis Hammond
COOPERATIVE AI FOUNDATION

Mahan Tourkaman
INDEPENDENT

Mark Davis
CROSSBAR INC.

Marta Belcher
FILECOIN FOUNDATION

Mark Miller
AGORIC

Matjaz Leonardis
UNIVERSITY OF OXFORD

Matthew McAteer
5CUBELABS

Megan Klimen
FILECOIN FOUNDATION

Michael Andregg
FATHOM RADIANT

Mikayla Maki
ZED

Murat Işık
FORESIGHT FELLOW

Natasha Asmi
HALOGEN LABS

Nicholas Brigham Adams
GOODLY LABS

Pranjali Awasthi
PRODIGY FELLOW

Ravi Pandya
CARNEGIE MELLON UNIVERSITY

Richard Ngo
OPENAI

Rick Korzekwa
AI IMPACTS

Rob Luke
AE STUDIO

Rose Bloomin
PLURALITY INSTITUTE

Ryan Singer
VEX CAPITAL

Shady El Damaty
HOLONYM

Sabrina Owens
AMS MACHINE DESIGN LLC

Tristan Harris
CENTER FOR HUMANE TECHNOLOGIES

Trym Berge
KYBER ENGINES

Thomas Michael Hagen
KYBER ENGINES

Ventali Tan
DELENDUM RESEARCH

Winslow Strong
CLUSTER CAPITAL

Zhu Xiaohu
SAIDAO

Zac Hatfield-Dodds
ANTHROPIC

Workshop chairs

Christine Peterson
CO-FOUNDER AND FORMER PRESIDENT, FORESIGHT INSTITUTE

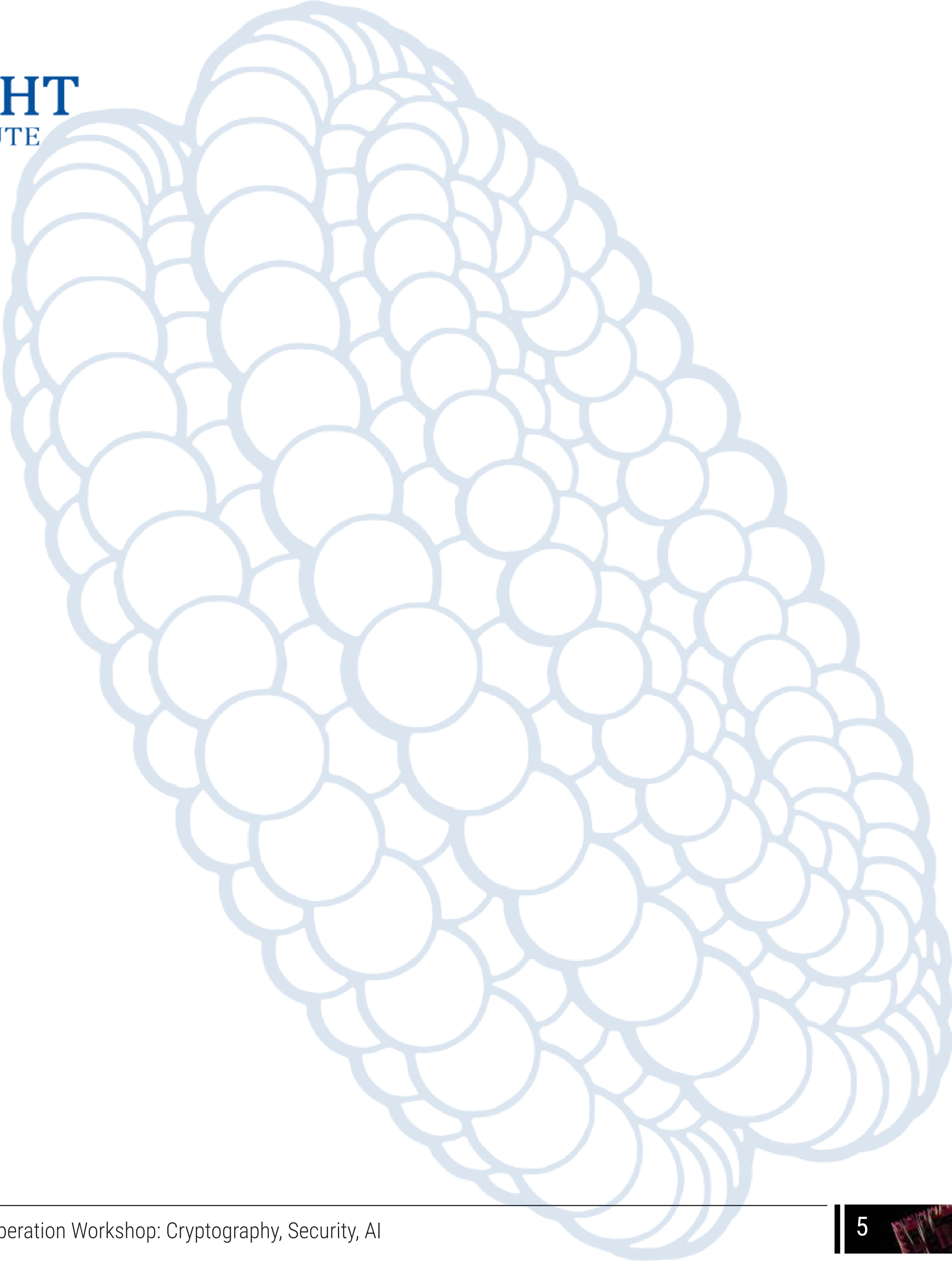
Allison Duettmann
PRESIDENT AND CEO, FORESIGHT INSTITUTE

Mark S. Miller
AGORIC

About Foresight Institute

Founded in 1987, 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.

FORESIGHT INSTITUTE



Workshop Chairs

Christine Peterson

FORESIGHT INSTITUTE, CO-FOUNDER AND FORMER PRESIDENT

Christine Peterson is co-founder and former President of Foresight Institute. She lectures and writes about nanotechnology, AI, and longevity. She is co-author of Unbounding the Future: the Nanotechnology Revolution (Morrow, also free online) and Leaping the Abyss: Putting Group Genius to Work (knOwhere Press, also free online). She advises the Machine Intelligence Research Institute, Global Healthspan Policy Institute, National Space Society, startup Ligandal, and the Voice & Exit conference. She coined the term 'open-source software.' She holds a bachelor's degree in chemistry from MIT.

Allison Duettmann

PRESIDENT AND CEO, FORESIGHT INSTITUTE

Allison Duettmann is the president and CEO of Foresight Institute. She directs the Intelligent Cooperation, Molecular Machines, Biotech & Health Extension, Neurotech, and Space Programs, Fellowships, Prizes, and Tech Trees and shares this work with the public. She founded Existentialhope.com, co-edited Superintelligence: Coordination & Strategy, co-authored Gaming the Future, and co-initiated The Longevity Prize. She advises companies and projects, such as Cosmica, The Roots of Progress Fellowship, and is on the Executive Committee of the Biomarker Consortium. She holds an MS in Philosophy & Public Policy from the London School of Economics, focusing on AI Safety.

Mark Miller

AGORIC

Mark S. Miller, Chief Scientist at Agoric, is a pioneer of agoric (market-based secure distributed) computing and smart contracts, the main designer of the E and Dr. SES distributed persistent object-capability programming languages, inventor of Miller Columns, an architect of the Xanadu hypertext publishing system, a representative to the EcmaScript committee, a former Google research scientist and member of the WebAssembly (Wasm) group.

Workshop Format

Rapid keynotes were followed by working groups to curate opportunities for talent and funders present at the workshop. Highlights also included mentorship hours, breakouts, and speaker, sponsor & fellowship gatherings.

RESIGHT
INSTITUTE
ing technology
Space
Neurotech
Existential Hope
ngevity Biotech
cular Machines
ent Cooperation

GHT
TUTE
Hello

RESIGHT
INSTITUTE
FORESIGHT

Talk to the City · Next Steps
,000,000,000,0
ET ARCHIVE · 25 OCTOBER
FORESIGHT
INSTITUTE
Advancing technology for the long-term benefit of life.
Space
Neurotech
Existential Hope
Longevity Biotech
INTERNET

ypt

Implications

- What does this tell us about interperability and robustness?
- Clearly whenever such backdoors are possible both of these things are really not
- Scott's AI safety idea
- Philosophical implications: leaving aside AI models how many ways of computing a given function are in any way comprehensible at all

Hello

Technology

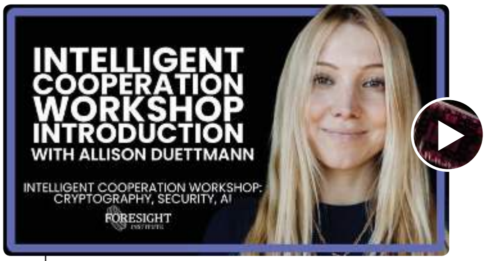
000,000,0
FORESI
INST
Advancing techn
for the long-term
benefit of life.
Exis

Brewster
Kipply
Jamie
for PA
Source align
TOM
tog

Keynote Presentations

Presentations by workshop chairs and invited keynotes

ℹ All workshop presentations are clickable via the play button


INTELLIGENT COOPERATION WORKSHOP INTRODUCTION
WITH ALLISON DUETTMANN
INTELLIGENT COOPERATION WORKSHOP: CRYPTOGRAPHY, SECURITY, AI
FORESIGHT

Allison Duettmann, Foresight Institute

Intelligent Cooperation Workshop Introduction

SUMMARY

Allison Duettmann introduces the workshop goals and discusses the significance of opportunities at the intersection of cryptography, security, and AI for promising futures. She delves into parallels between security and safety mindsets, explores potential avenues for enhancing AI infosec and red-teaming, outlines strategies to prevent collusion, and highlights emerging privacy-preserving paths for cooperation.

SUPER INTELLIGENT ALIGNMENT
WITH JAN LEIKE
INTELLIGENT COOPERATION WORKSHOP: CRYPTOGRAPHY, SECURITY, AI
FORESIGHT

Jan Leike, OpenAI

Super Intelligent Alignment

SUMMARY

Jan Leike discusses OpenAI's recognition of the dangers of superintelligence, and the importance of aligning it with human intent. Current alignment techniques won't scale to superintelligence, so OpenAI's new Superalignment effort aims to build an automated alignment researcher, and use scalable oversight and generalization techniques. They plan to stress-test their approach by deliberately training misaligned models to detect potential issues arising downstream early. OpenAI is dedicating a team and 20% of their compute resources to their superintelligence alignment efforts. Leike remains optimistic that focused efforts can solve this critical issue and encourages participants to red-team the project.

Keynote Presentations

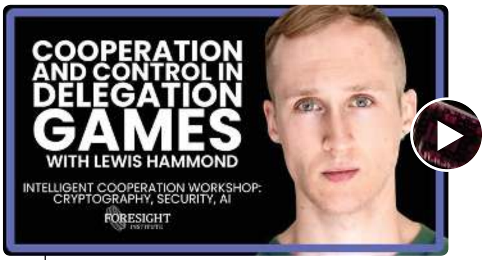
FIRESIDE CHAT ON AI ALIGNMENT
WITH KIPPLY CHEN
INTELLIGENT COOPERATION WORKSHOP: CRYPTOGRAPHY, SECURITY, AI
FORESIGHT INSTITUTE

Kipply Chen, Anthropic

Fireside Chat on AI Alignment

SUMMARY

Kipply Chen and Allison Duettmann discuss AI alignment, focusing on data-related tasks at the pre-training level. Chen believes that alignment is in its early stages, but expects this to progress faster than AI capabilities. Challenges include aligning predictive and base models, which require concentrated efforts and coordination between labs. She also highlights that security, and policy work are crucial in AI safety. Furthermore, she discusses another workshop participant's proposal on using cryptography techniques for watermarking language models. Finally, she anticipates an increase in lawsuits related to training data misuse and calls for an open-source alignment community to address alignment challenges.


COOPERATION AND CONTROL IN DELEGATION GAMES
WITH LEWIS HAMMOND
INTELLIGENT COOPERATION WORKSHOP: CRYPTOGRAPHY, SECURITY, AI
FORESIGHT INSTITUTE

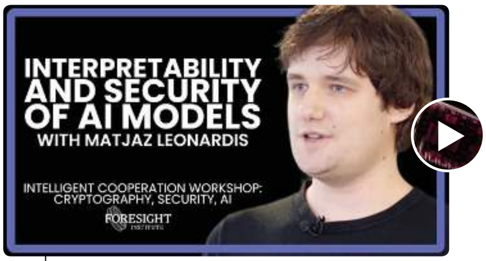
Lewis Hammond, Cooperative AI Foundation

Cooperation and Control in Delegation Games

SUMMARY

Lewis Hammond explores delegation games in AI safety, where humans delegate tasks to AI systems in a context of multi-agent problems. He sees both control problems involving aligning preferences and capabilities, and cooperation problems which are aiming for high joint welfare. He emphasizes cooperation problems, as AI systems will likely interact more frequently with one another. On the other hand, he discusses measuring cooperative capabilities using concepts like the price of anarchy and equilibrium selection. He recognizes that understanding and measuring cooperative capabilities will help manage dynamics among multiple AI systems. Given that preventing collusion between machine learning agents is a challenge, he sees the importance of exploring detection and mechanism design to ensure ethical behavior and trust in AI systems.

Keynote Presentations



Matjaz Leonardis, University of Oxford

Interpretability and Security of AI Models

SUMMARY

Matjaz Leonardis discusses the concept of back doors in machine learning models. He explains that while machine learning models are typically used as classifiers, there is interest in using them for consequential decisions such as university admissions or loan applications. However, there are concerns about the ability for someone to train the model in a way that could influence its outcomes. Leonardis introduces his most recent paper that successfully demonstrated the existence of back doors in machine learning models, where secret modifications can be applied to inputs to produce desired outputs without detection. He explains the technique used to hide these back doors in seemingly random choices during training. He raises questions about the interpretability and robustness of models with back doors, as well as the potential for using this technique in AI safety. He concludes by mentioning the need to explore extending back doors to more complex models and the role of cryptography and security in understanding the explainability of computation.

Brendon Wong, Cosmic

Safety First Cognitive Architectures

SUMMARY

Brendon Wong discusses safety-first cognitive architectures and their relevance to security. These architectures are designed with AI safety in mind, ensuring interpretable and corrigible goals, plans, and actions. Different architectures, like Conjecture and Eric Drexler's Open Agency Model, have varying levels of similarity to models like Auto GPT. Security implications include sandboxing and preventing influence between system components. Cognitive architectures restrict AI models' access to necessary information and actions. Challenges and solutions depend on the design and underlying AI models. For Wong, as technology advances, it is important to reevaluate safety measures and develop new ones.

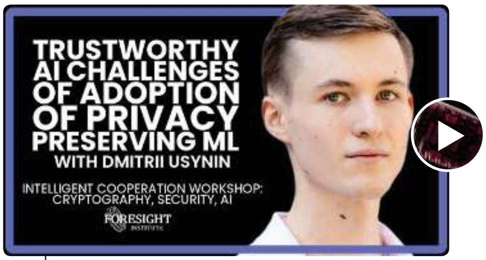
Keynote Presentations

Colleen McKenzie, AI Objectives Institute

Scaling Deliberation

SUMMARY

Colleen McKenzie emphasizes the significance of deliberation, which involves understanding stakeholders' needs and wants beyond outcome preferences. Preferences and voting can have issues, such as changing opinions and social choice dilemmas. Deliberation helps address these problems by allowing people to discuss their reasoning and uncover underlying values. Scaling deliberation becomes challenging – but she argues that technology can assist. Talk to the City, developed by the AI Objectives Institute, uses clustering and chatbots to process data, extract norms, and facilitate discussions with diverse viewpoints. Future directions include enhancing the tool to identify compatible views and conflicts, tracking preference changes over time, and enabling group organization and reflection at scale. Collaboration and ideas for improving the tool's capabilities are encouraged.



Dmitrii Usynin, Creator Fund

Trustworthy AI Challenges of Adoption of Privacy Preserving ML

SUMMARY

Dimitrii Usynis discusses how Machine learning relies on diverse and well-curated datasets, and how obtaining them is challenging due to data protection regulations, low quality, and biases. Trustworthy Artificial Intelligence (TAI) addresses these issues with privacy-preserving, explainable, and fair model training. Privacy-preserving ML (PPML) ensures safe and robust AI systems. Challenges include scalable tools and incentives for participation. Approaches like differential privacy and homomorphic encryption can help secure distributed ML pipelines and protect privacy. This talk explores the state of PPML, its motivations, challenges, and necessary developments for broader adoption. Balancing privacy and ML model training is crucial. Overall, via ongoing research and development, Usynin aims to overcome the challenges surrounding PPML and foster the integration of privacy-enhancing techniques with other aspects of trustworthy AI through ongoing research.

Keynote Presentations

CONTROLLABLE AI WITH OPEN AGENCY SYSTEMS
WITH EVAN MIYAZONO
INTELLIGENT COOPERATION WORKSHOP: CRYPTOGRAPHY, SECURITY, AI
FORESIGHT INSTITUTE

Evan Miyazono, Protocol Labs

Controllable AI with Open Agency Systems

SUMMARY

Evan Miyazono presents a project on controllable AI and coordination systems, addressing challenges in specifying desired outcomes for AI systems. The focus is on controllability to mitigate risks associated with general intelligence by modifying constraints to bind AI behavior. The proposed Open Agency Architecture involves participants interacting with an LLM to generate outcome specifications, using reinforcement learning for controllable policies. Miyazono reminds us that collaboration among entities and research labs is crucial for building collaborative AI systems, and discusses how we could generate specifications, formal verification, and software patching. Finally, the potential of LLMs generating exploits and loopholes underscores the importance of aligning AI with the legal system.

BRIDGING TO THE GAP TO THE SECURITY FIELD
WITH JAMIE JOYCE
INTELLIGENT COOPERATION WORKSHOP: CRYPTOGRAPHY, SECURITY, AI
FORESIGHT INSTITUTE

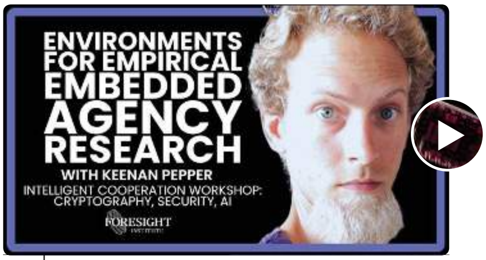
Jamie Joyce, The Society Library

Bridging to the Gap to the Security Field

SUMMARY

Jamie Joyce discusses the application of her methodologies to AI safety. She introduces comprehensive collective intelligence – mapping socio-political issues using collective reasoning. Extracting arguments, claims, and evidence from various perspectives builds databases for mapping. The approach is rooted in extracting arguments, claims, and evidence: Arguments are broken down and mapped for deductive conclusions, with knowledge artefacts visualizing debates for efficient expertise gain by users. Joyce highlights that comprehensive coordinated collective intelligence benefits knowledge accumulation and researcher onboarding; however, challenges include generating chain of thought reasoning from natural language and semantically linking premises.

Keynote Presentations

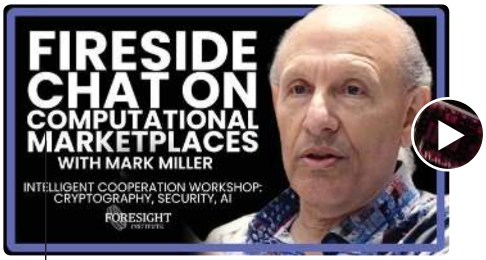

ENVIRONMENTS FOR EMPIRICAL EMBEDDED AGENCY RESEARCH
WITH KEENAN PEPPER
INTELLIGENT COOPERATION WORKSHOP: CRYPTOGRAPHY, SECURITY, AI
FORESIGHT INSTITUTE

Keenan Pepper, Salesforce

Environments for Empirical Embedded Agency Research

SUMMARY

Keenan Pepper proposes a research direction to address safety and alignment concerns in AI, focusing on embedded agents. Embedded agents exist within their environment, which can impact their cognition and introduce manipulation risks. Studying environments with embedded and intelligent agents is valuable for understanding potential dangers and exploring interpretability. Pepper highlights the need for a safe sandbox to study embedded agents before superintelligence is embedded. Interpretability is crucial, especially in games where success depends on revealing internal states to opponents. Finally, Pepper suggests creating an environment, possibly using a game-like setup, where agents can be trained to perform embedded tasks while revealing their internal states.


FIRESIDE CHAT ON COMPUTATIONAL MARKETPLACES
WITH MARK MILLER
INTELLIGENT COOPERATION WORKSHOP: CRYPTOGRAPHY, SECURITY, AI
FORESIGHT INSTITUTE

Mark Miller, Agoric

Fireside Chat on Computational Market Places

SUMMARY

Mark Miller discusses the connection between software engineering and economics in AI. He highlights the unity between large software systems and large societal systems, following Hayek. He suggests that studying the emergence of intelligence in human institutions within society to uncover useful parallels for AI institution design. Miller discusses James Madison's approach to the alignment problem and contrasts it with the unipolar takeover perspective. He emphasizes building transparent, accountable, and incorruptible institutions through technology and voluntary rule-based frameworks. One case in point is object capabilities for computer security and its application to secure and decentralized interactions in AI ecosystems. Finally, he explores challenges and pathways towards secure foundations, including AI in software engineering, usable security, and informed consent.

Keynote Presentations

SEMICONDUCTOR SUPPORT FOR CRYPTOGRAPHIC INNOVATION
WITH MARK DAVIS
INTELLIGENT COOPERATION WORKSHOP: CRYPTOGRAPHY, SECURITY, AI
FORESIGHT INSTITUTE

Mark Davis, Crossbar Inc.

Semiconductor Support for Cryptographic Innovation

SUMMARY

Mark Davis highlights the importance of semiconductor support for cryptographic innovation and the need for improved collaboration between the semiconductor and cryptography communities. He emphasizes semiconductors as the root of trust in security and the limitations of off-the-shelf solutions. He acknowledges both challenges and opportunities for semiconductor support, such as new potential attacks and open-source hardware complexities. Davis concludes by highlighting the importance of creating feasible robust identification systems and reimagining the design process.

FIRESIDE CHAT ON CRYPTO, AI & CIVIL LIBERTIES
WITH MARTA BELCHER
INTELLIGENT COOPERATION WORKSHOP: CRYPTOGRAPHY, SECURITY, AI
FORESIGHT INSTITUTE

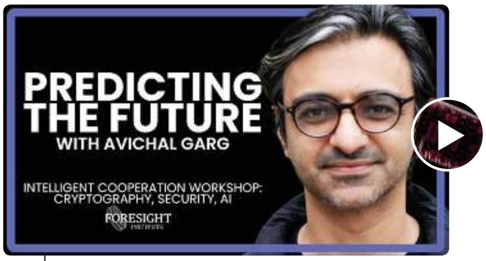
Marta Belcher, Filecoin Foundation

Fireside Chat on Crypto, AI & Civil Liberties

SUMMARY

Marta Belcher gives a fireside chat on why privacy and civil liberties are so important with respect to AI, why they might be at risk, and what technical or social activists could do to help strengthen civil liberties going forward.

Keynote Presentations


PREDICTING THE FUTURE
WITH AVICHAL GARG
INTELLIGENT COOPERATION WORKSHOP: CRYPTOGRAPHY, SECURITY, AI
FORESIGHT INSTITUTE

Avichal Garg, Electric Capital

Predicting the Future

SUMMARY

Avichal Garg highlights that there may exist a potentially symbiotic relationship between crypto and AI. For example, AI could write distributed applications on crypto networks, while AI bots could use decentralized systems for payments. He argues that there is the potential to grant personhood to AI in small countries, which could lead to jurisdictions effectively allowing non-human organizations to control capital. For him, challenges like deep fakes can be addressed by incorporating cryptographic signatures into human-created media. In conclusion, he highlights the complementary relationship between AI and crypto, emphasizing the potential for significant changes in various sectors as these technologies continue to intersect.

AI
OPPORTUNITY
WITH BREWSTER KAHLE
INTELLIGENT COOPERATION WORKSHOP: CRYPTOGRAPHY, SECURITY, AI
FORESIGHT INSTITUTE

Brewster Kahle, Internet Archive

AI Opportunity

SUMMARY

Brewster Kahle proposes the establishment of a public AI research lab that brings together the research community, research libraries, and data collections. The lab would aim to use AI tools to address significant problems and create a valuable data set. Kahle suggests combining data assets from various libraries, including climate-related materials, to advance climate research. Building the lab requires GPU stacks, AI-proficient researchers, and engagement from climate researchers and target users. The lab prioritizes preserving metadata and seeks collaborations to access valuable datasets. Kahle emphasizes the need for resources and envisions the lab as a collaborative effort where organizations contribute resources, including government documents, gray literature, and public web materials.

Keynote Presentations

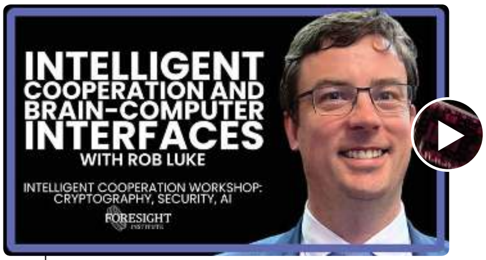
TECHNICAL PROJECTS FOR EASIER COOPERATION
WITH RICK KORZEWKA
INTELLIGENT COOPERATION WORKSHOP: CRYPTOGRAPHY, SECURITY, AI
FORESIGHT INSTITUTE

Rick Korzekwa, AI Impacts

Technical Projects for Easier Cooperation

SUMMARY

Rick Korzewka discusses technical projects related to AI development and cooperation. He emphasizes the importance of addressing the question of AI and cooperation, referencing the recent request for comment from the White House OSTP. He highlights the need for seeking marginal gains in technical projects for AI cooperation, such as seismic signal analysis and predicting AI capabilities. Finally, he also addresses the challenge of distinguishing between dangerous and safe AI systems and the value of capability predictions in AI policy concerns.


INTELLIGENT COOPERATION AND BRAIN-COMPUTER INTERFACES
WITH ROB LUKE
INTELLIGENT COOPERATION WORKSHOP: CRYPTOGRAPHY, SECURITY, AI
FORESIGHT INSTITUTE

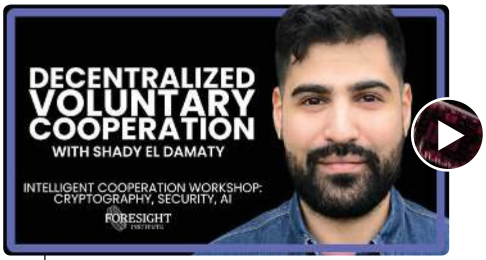
Rob Luke, AE Studio

Intelligent Cooperation and Brain-Computer Interfaces

SUMMARY

Rob Luke discusses the importance of security, cryptography, and AI for brain-computer interfaces (BCIs). He highlights the need for robust BCI models whilst ensuring privacy and protecting neurodata. He explains how BCIs work, involving implanted arrays in the brain that measure signals from neurons and transmit them to a computer for decoding – AI plays a crucial role in enhancing information extraction. Next, Luke addresses privacy and security challenges, advocating for open and distributed systems with verifiable encryption. Generally, he emphasizes the need for technical systems to ensure cooperation, security, and privacy in training and generating models for neurodata.

Keynote Presentations

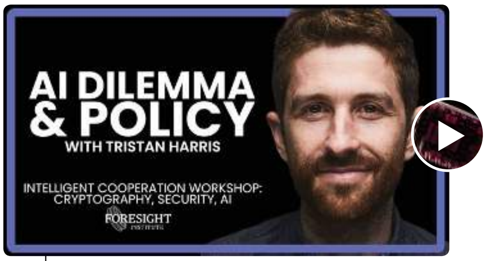

DECENTRALIZED VOLUNTARY COOPERATION
WITH SHADY EL DAMATY
INTELLIGENT COOPERATION WORKSHOP: CRYPTOGRAPHY, SECURITY, AI
FORESIGHT INSTITUTE

Shady El Damaty, Holonym

Decentralized Voluntary Cooperation

SUMMARY

Shady El Damaty discusses the use of zero-knowledge proofs to enhance security, usability, and self-sovereignty on the internet. He emphasizes the importance of shaping the future through technology, relationships, and institutions that promote voluntary cooperation and distributed networks. El Damaty recognizes the benefits of distributed ledgers but acknowledges challenges like Sybil attacks, accountability, privacy, and verification of off-chain behavior. He proposes privacy-preserving attestations as a solution, allowing individuals to verify off-chain claims while preserving privacy. El Damaty also explores decentralized secure online identities, seamless key custody, and the challenges of identity protection.


AI DILEMMA & POLICY
WITH TRISTAN HARRIS
INTELLIGENT COOPERATION WORKSHOP: CRYPTOGRAPHY, SECURITY, AI
FORESIGHT INSTITUTE

Tristan Harris, Center for Humane Technologies

AI Dilemma & Policy

SUMMARY

Tristan Harris raises concerns about AI development and stresses the need for collaboration among safety teams. He calls for consensus among Western nations to establish guardrails for AI companies and emphasizes framing AI risks in concrete terms. Harris believes demonstrating the ability to address AI challenges is crucial, suggesting that upcoming elections in the EU and the US could provide an opportunity to showcase this. He discusses the concept of learned helplessness and proposes emphasising short-term wins, such as holding generative AI companies accountable for content. Generally, Harris advocates for proactive measures to shape AI development and regulation before it becomes deeply embedded in society.

Keynote Presentations

INTERPRETABILITY FOR SAFETY AND ALIGNMENT
WITH FAZL BAREZ
INTELLIGENT COOPERATION WORKSHOP: CRYPTOGRAPHY, SECURITY, AI
FORESIGHT INSTITUTE

Fazl Barez, Apart Research

Interpretability for Safety and Alignment

SUMMARY

Fazl Barez described some of his recent work, leading through a mechanistic interpretability "recipe". He discussed interpreting language model neurons at scale, specifically delving into the attention-head neuron interaction and neuroplasticity. After explicating what it means to formulate the problem, he ended by discussing training humans to predict model behaviour.

Project Proposal Winners

Inspired by the challenges pointed out during the introductory keynotes, working groups formed to address problems of common interests.

Automatically Mapping Rigorous Discourse

SUMMARY

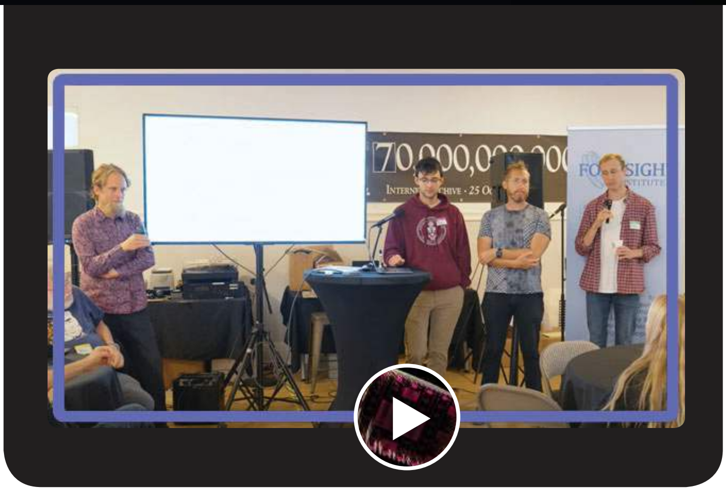
This project centers on automatically mapping rigorous discourse using AI and machine learning. The aim is to automate the process of generating deliberation and argument maps from various forms of media, including text, podcasts, and videos. A proof of concept and a validated methodology enhance the comprehensiveness and integrity of the maps. The structured deliberation graphs serve purposes such as decision-making, knowledge reference, prediction markets, and refining AI's logical reasoning. There's an emphasis on available personnel to initiate the project. The group delves into the intricacy of argumentation and the capability of their tool to address both basic and intricate arguments. The adaptability of their knowledge graph in depicting diverse linguistic registers stands out. An emphasis lies on the current exclusion of probabilistic thinking in their knowledge graph, prioritizing relevant information and context before evaluating the veracity of a claim.

Personalised AI Assistants

SUMMARY

This group discusses the challenges and strategies for crafting a personalized AI assistant that upholds privacy. Two approaches are highlighted: first, locally fine-tuning an existing model with personal data, and second, deploying differential privacy techniques to train a model from scratch. The significance of elucidating technical facets to ensure user comprehension emerges. The group underlines the need for alpha testing with a varied user base and potential costs. Monitoring and refining the performance of personalized tasks over time is pivotal. They acknowledge the risks of cultural biases and the imperative for cultural sensitivity in AI assistants. Emphasis is placed on the rewards of resolving issues around privacy and utility, such as enhancing administrative tasks and boosting efficiency. The nuances of defining personal data are broached, alongside the prospect for users to delineate their privacy parameters. Concluding insights involve prior dialogues on fiduciary AI assistance and legal factors.

Preventing Collusion



SUMMARY

This project zeroes in on the phenomenon of collusion between AI systems, where collusion denotes AI agents collaborating detrimentally against human interests. Communication avenues like steganography or cryptography for collusion are explored. The group investigates viewing games or market designs as super-agents derived from merging multiple agents and the repercussions of collusion. Parameters in game design influencing collusion, such as private communication channels and enduring identities, undergo examination. Initiatives like surveying and categorizing collusion methods, simulating to discern mechanisms that heighten collusion, and pinpointing ways to inhibit these mechanisms are proposed. Potential drawbacks, including impeding desired cooperation and crafting surveillance instruments for collusion detection, are acknowledged. Discussions span the intricacies of grasping collusion on a minuscule scale and the advantages of diversifying agent goals. Proposals encompass introducing agents into prevailing systems to supervise and endorse competitive dynamics and fostering variance and turnover in the agent cohort to thwart collusion.

Other Project Presentations

Marginal Governance Improvement

SUMMARY

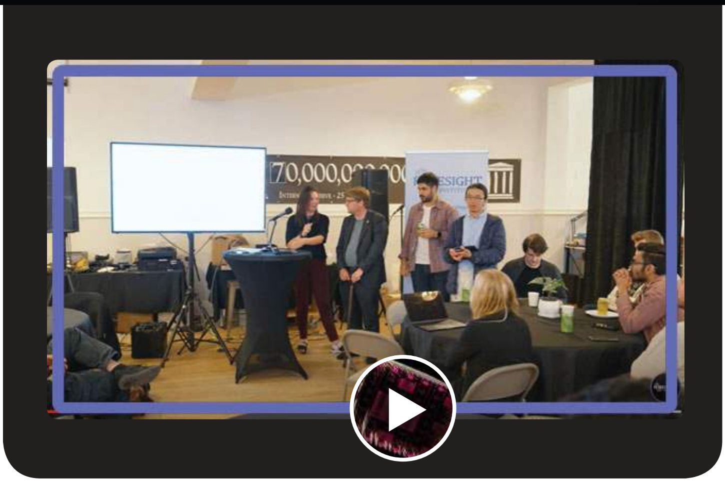
This group delves into strategies for enhancing governance outcomes with minimal exertion, leveraging both current and emerging technologies. They dissect various tools and initiatives, such as Talk to the City, the Public Editor project, and deliberation instruments. Potential applications span sectors like the healthcare system, academic institutions, school boards, and Wikipedia, perceived to gain from tools that elevate discourse quality and expedite decision-making. Emphasis is laid on the imperative for collaboration and dialogue in communities influencing AI-related resolutions. The group ponders the prospect of integrating AI into political arenas but approaches with caution. The concept of training AI models using copious opinion data is broached, with prospective next steps encompassing collaborations with Wikipedia and training endeavors on pertinent datasets. The project is open to suggestions on discerning other use cases and welcomes contributors eager to develop or fund similar endeavors.

Formalising Boundaries Empirically

SUMMARY

This group delves into the subject of boundaries concerning individual sovereignty rights and interactions between entities. They explore diverse boundary types, encompassing physical, intellectual, psychological, emotional, cellular, and societal/cultural distinctions. The goal is to concretely identify and articulate boundaries in ways both humans comprehend, and machines can adhere to. Two approaches emerge: deploying Markov blankets to discern boundaries based on correlations and portraying boundaries to facilitate human understanding and respect. The conversation spans potential applications in domains like software systems, video games, computer security, and biological contexts. They underscore the necessity of boundary formalization and propose reinforcement learning complemented by human feedback.

Design for a Robust ID System



SUMMARY

This group addresses the formidable challenge of validating virtual identities, proposing solutions to this contemporary issue. The intricacies of verifying the authenticity, reputation, and privileges of virtual identities distinct from their biological counterparts come to the fore. The existing practice of anchoring usernames to legal or physical identifiers proves insecure. The conversation gravitates toward adopting provably secure devices like passkeys to monitor privileges linked with virtual personas. The pressing need for enhanced multi-factor authentication solutions to thwart unauthorized access emerges. Current authentication method limitations, encompassing hardware key vulnerabilities and malware attack susceptibility, undergo scrutiny. User experience (UX) and cultural assimilation emerge as pivotal in embracing more secure identity validation systems. A pivot is suggested: distinguishing authorized from unauthorized messages rather than discerning humans from AIs. Deploying public key cryptography to enforce property rights stands out as a prospective remedy. The dialogue encompasses delegated authentication and multi-party solutions, advocating for multi-device and redundant authentication to reduce risks. The group emphasizes education as the catalyst for users to comprehend and adopt novel authentication methods. While project specifics remain nuanced, the accent is on pioneering research to optimize user experience and forge more secure identity validation frameworks.

Positive Futures

SUMMARY

This group underscores the significance of AI's positive instances. Brainstorming sessions yield ideas, such as tailored AI assistance warding off scams and AI invasions, AI tools pinpointing and countering negative spirals in social media engagements, and systems proactively alerting users about scams while dissuading their use. The group envisions the emotions to cultivate in a post-scarcity future, juxtaposed with potential risks stemming from malevolent AI applications. The notion of retrofitting existing AI endeavors to augment safety emerges, aiming to devise systems that intercede in dialogues to avert detrimental outcomes. The proposal of launching a blog to showcase a proof-of-concept project harnessing public language models materializes. The project's completion timeline, human volunteer testing, and funding prospects constitute discussion points. A fleeting mention of the Future of Life World-Building Contest for AI emerges as an inspirational catalyst. The overarching theme revolves around conceptualizing and manifesting positive AI implementation paradigms.

AI Safety Grant

At this workshop, we shared Foresight Institute's new AI Safety Grant with attendees before the official launch. Consequently, we encouraged workshop participants to apply for our grant, potentially with the shared project proposals they collaborated on during Day 2 of the Workshop.

In light of the potential for shorter AGI timelines, we have decided to support much-needed development across the following areas with our AI Safety Grant:

1. Neurotechnology, Whole Brain Emulation, and Lo-Fi Uploading for AI Safety
2. Cryptography and Security Approaches for Infosec and AI Security
3. Safe Multipolar AI Scenarios and Multi-Agent Games

Neurotechnology, Whole Brain Emulation, and Lo-Fi Uploading for AI Safety

We are interested in exploring the potential of neurotechnology, particularly Whole Brain Emulation (WBE) and cost-effective lo-fi approaches to uploading, that could be significantly sped up, leading to a re-ordering of technology arrival that might reduce the risk of unaligned AGI by the presence of aligned software intelligence.

We are particularly excited by the following:

- WBE as a potential technology that may generate software intelligence that is human-aligned simply by being based directly on human brains
- Lo-fi approaches to uploading (e.g., extensive lifetime video of a laboratory mouse to train a model of a mouse without referring to biological brain data)
- Neuroscience and neurotech approaches to AI Safety (e.g., BCI development for AI safety)
- Other concrete approaches in this area
- General scoping/mapping opportunities in this area, especially from a differential technology development perspective, as well as understanding the reasons why this area may not be a suitable focus

Cryptography and Security Approaches for Infosec and AI Security

Exploring the potential benefits of Crytography and Security technologies in securing AI systems includes:

- 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
- General scoping/mapping opportunities in this area, especially from a differential technology development perspective, or exploring why this area is not a good focus area

Safe Multipolar AI Scenarios and Multi-Agent Games

Exploring the potential of safe Multipolar AI scenarios, such as:

- 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
- General scoping/mapping opportunities in this area, especially from a differential technology development perspective, or exploring why this area is not a good focus area

Workshop

Intelligent Cooperation: Cryptography, Security, AI

July 10 - 11,
2023,
9 am - 5 pm

Hosted by

Mark S. Miller
Agoric

Christine Peterson
Foresight Institute

Allison Duettmann
Foresight Institute

The Internet
Archive
San Francisco

Workshop Sponsors

Agoric

INTERNET ARCHIVE