FORESIGHT INSTITUTE

Foresight Crypto, AI, and Security Workshop

Foresight Institute
Allison Duettmann
Aaron King

Oct 4-5, 50 Years HQ
San Francisco, CA

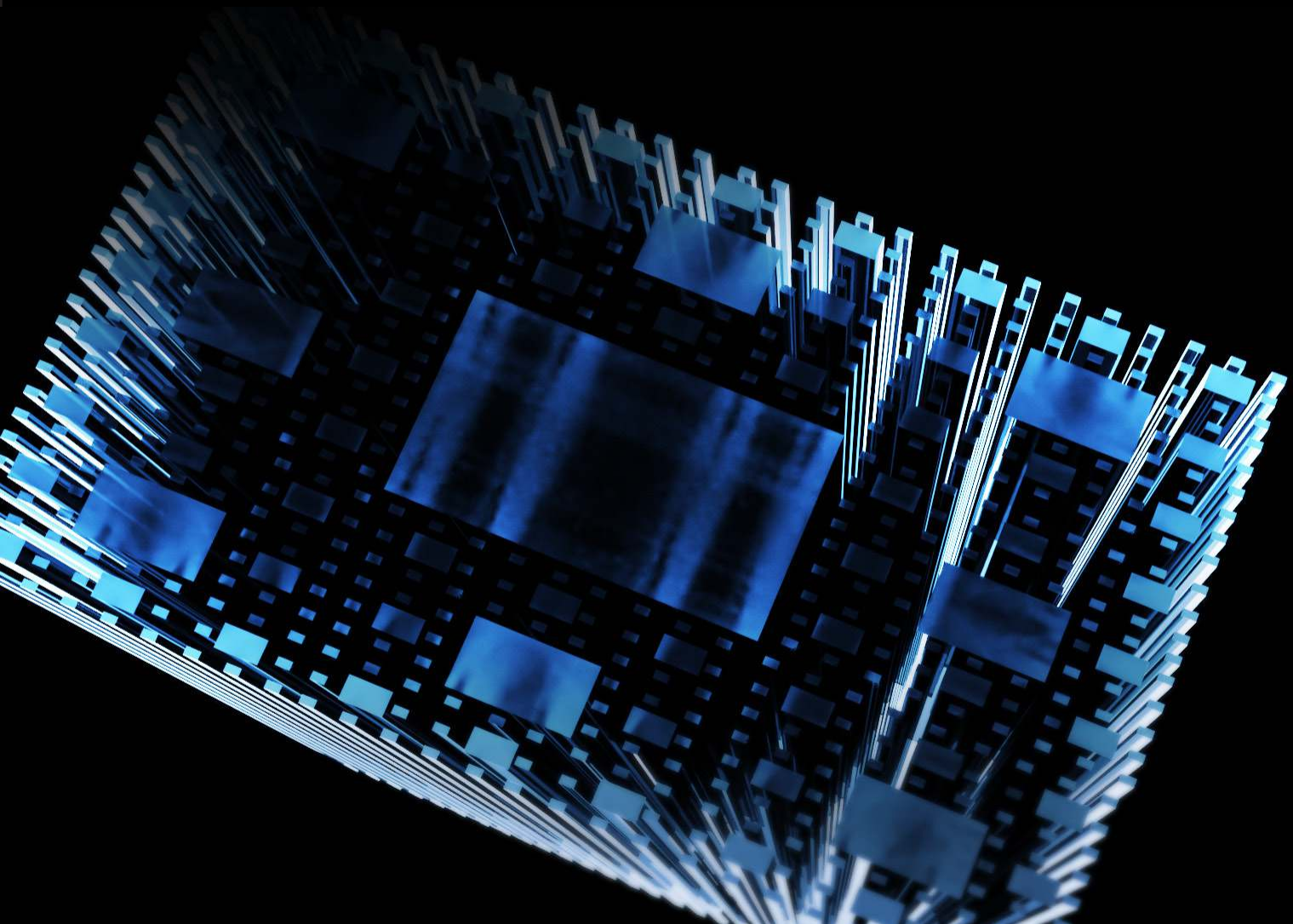


Table of Contents

About Foresight

Since 1986, The Foresight Institute has advanced beneficial use cases of high-impact technologies for long-term futures.Through virtual seminars, fellowships, workshops, and prizes, we support science and technology that is too early-stage, interdisciplinary, or ambitious for legacy institutions to fund.

Our focus areas include:
Molecular Machines to better control matter
Biotechnology to reverse aging
Computer Science to secure human AI cooperation
Neurotechnology to support human flourishing
Spacetechnology to further exploration

FORESIGHT INSTITUTE

Workshop Sponsors

Agoric

Protocol Labs

50 Y Fifty Years

Participants

Ale Borda
Fifty Years

Alex Blania
Worldcoin

Alex Aleksyenko
Independent

Alexander Green
Persona

Allison Duettmann
Foresight Institute

Amanda Ngo
Ought

Anna Yelizarova
Future of Life Institute

Andy Ku
IN-Q-TEL

Andreas Kuehn
ORF

Anthony Aguirre
Future of Life Institute

Austin Liu
Cornell University

Chhi'mèd Künzang
Protocol Labs

Chris Waclawek
Worldcoin

Christine L. Peterson
Foresight Institute

Christopher Bender
Renegade

David Ernst
Secure Internet Voting

Danny O'Brien
Filecoin Foundation

Dean Tribble
Agoric

Deepak Maram
Cornell University

Deger Turan
AI Objectives Institute

Eduardo da Veiga Beltrame
ImYoo

Eugene Leventhal
Smart Contracts Research Institute

Evan Miyazono
Protocol Labs

Gernot Heiser
seL4

Ghada Almashaqbeh
University of Connecticut

Greg Osuri
Akash Network

Henry de Valence
Penumbra Labs

Isabella Duan
University of Chicago

Jazear Brooks
Swish

Jeffrey Ladish
Anthropic

Joel Thorstensson
Ceramic Network

Jonathan Passerat-Palmbach
Imperial College

Juan Benet
Filecoin

Kanjun Qui
Generally Intelligent

Kate Sills
Independent

Kevin Compher
IN-Q-TEL

Kipply Chen
Anthropic

Kydo
Stanford University

Mark S. Miller
Agoric

Martin Karlsson
Discourse Graphs

Michael Freedman
Field's Medalist

Matjaz Leonardis
University of Oxford

Matthew McAteer
Formic Labs

Max Webster
Hivemind Ventures

Mei Z
Lemniscap

Michelle Ritter
Steel Perlot

Morgan Livingston
TechCongress Fellow

Mikayla Maki
Zed

Nicolas Moes
European AI governance

Nick Matthew
Figment Capital

Philipp Koellinger
DeSci

Remco Bloemen
Worldcoin

Robin Hanson
George Mason University

Rosie Campbell
OpenAI

Ryan Singer
Chia

Ryan Grant
Independent

Shaun Conway
iXO

Silke Elrifai
Independent

TJ
Oxford University

Ventali Tan
Delendum

Willem Van Der Schoot
Independent

Winslow Strong
Independent

WhyRUSleeping
Protocol Labs

Ying Tong Lai
Electric Coin Company

Executive Summary

Many civilizational opportunities, from spurring cooperation to supporting defense, lie at the intersection of cryptography, security, and AI. These technologies could make human AI cooperation secure, cheap, automatable, cross-jurisdictional, and incorruptible.

Nevertheless, progress at the nascent intersection of these technologies is still underfunded and flies under the radar of specialists in each field, as well as junior talent seeking promising career paths. To change this, this two-day workshop invited specialists across domains to learn about undervalued opportunities for progress relevant to their field, and catalyze lasting collaboration toward shared long-term goals.

Keynote presentations on progress in cryptography, security, and AI were followed by working groups to determine opportunities for progress. This report summarizes the keynotes and resulting project proposals, some of which are now under development.

These icons, found throughout the report, link to recordings of each presentation

Thank you to our participants for making this workshop into such a success and to our sponsors for making this event possible. Thank you to Mark S. Miller for chairing this event. Mark S. Miller is the chief scientist of Agoric, a pioneer of agoric (market-based secure distributed) computing and smart contracts.

Areas explored are based on technologies highlighted in Gaming the Future, a Foresight Institute book exploring technologies to secure human AI cooperation (available here).

We encourage you to reach out if you are interested in supporting ongoing progress.

Allison Duettmann
President & CEO
Foresight Institute
a@foresight.org

Lightning Presentations

The Long-term Potential of Cryptocommerce

Allison Duettmann | President of Foresight Institute Allison gave opening remarks about the potential for cryptocommerce. Although opportunities for bright futures enabled by bio, nano, and AI technologies are now within our reach, their proliferation also comes with risks and authoritarian attempts at control. Cryptographic tools can help us navigate the traps because they enable decentralized, secure cooperation, thereby unlocking a path of high technology, security, and freedom. Despite their potential, cryptographic tools are still highly undervalued, not just in traditional politics, law and economics, but even in the "crypto" community. She encourages this workshop to explore how cryptographic tools can help civilization to cooperate, defend itself, and do both in light of AI.

Computational Markets & Agoric Systems

Mark Miller | Agoric Humans cooperate through institutions to solve problems that could not be solved individually. Civilization as a whole is the superintelligent ecosystem emergent of those interactions. However, the infrastructure that enables cooperation has extremely bad security. The exploitation of these vulnerabilities are currently still limited by human effort, but AI is already changing the ability to exploit these systems. Securing the foundation for cooperation is necessary if we want civilization to be resistant to emerging technologies, and there are promising projects we can support today.

Crypto Tools for Paretotopia

Juan Benet | Filecoin Crypto can be used as a massive lever to move us toward paratotopia. The concept of paratotopia is based off of an ideal outcome between two competing interests. Juan believes the pie of resources can be grown via crypto technology to produce better cooperation. Altering incentive structures could stimulate large untapped resources to be deployed toward the most impactful projects that further the common good.

Intelligent Cooperation Tech Tree

Ying Tong Lai | Electric Coin The concepts and connections within the field of Intelligent Cooperation are difficult to navigate. Having a map of developmental fields related to AI, distributed computation, and cryptography helps visualize how these fields overlap and what kind of collaborative efforts are possible. Ying Tong walks through the current version of the tech tree to briefly describe the high-level connections between these concepts and how it has helped shape her view of privacy and security.

Tamper-evident logs and Unalterable Docs

Kate Sills | Independent On the one hand we have costly problems of fraud and malicious behavior present in the real world, and on the other we have cryptographic tools. How do we merge these together to create an elegant solution? Unforgeable signatures, timestamped records, and unalterable documents are potential applications of cryptographic tools. However, building appropriate tools means having fluid ideals to recognize the best solutions for a particular problem. Centralization of time-stamping is fine, as long as there are competitors for situations of prolonged down-time. Decentralized blockchains are costly, while cryptographic tools are cheap.

Better Directions for Using Information

Whyrusleeping | Protocol Labs Let's build an exocortex to outcompete the AGI! An exocortex is an external information processing tool to augment your own cognition. Daily planners, computers, and smart phones are crude examples of working exocortex machines. To get to the next level, brain computer interfaces and good software will converge to get us there. Whyrusleeping believes software is the current bottleneck. We have good information storage and search tool, but information access and comprehension are lacking. Beyond these concepts, the hardest and most interesting ability to enhance is our ability to inference, to generate new ideas from pre-existing information.

Creative Computing

Matjaz Leonardis | University of Oxford Creativity and computation - can we go beyond simple CRUD apps? Most of today's apps connect you with information other people have produced. What would it take to have a computer produce novel information and solve problems in a truly open ended way? The ability to solve problems in unique ways is often labeled a cornerstone of human civilization but we don't have a good understanding of it. At least, not a complete enough understanding to convert it to code. If we can change that, we can open the floodgates of computational power as applied to complex problems.

FGoAT: Proof of Location

Deepak Maram | Cornell University Deepak is working on verifiably proving geolocation of files, identities, and arbitrary computation. The current method to determine geographic location is through GPS or IP addresses, which can be spoofed. Having geo-distributed verifiers would be a great starting point, and the millions of geodistributed websites hosted by businesses all over the world seems like an untapped resource for solving this problem. Deepak is leveraging distributed timestamps from these servers to validate locations, and has built a demonstration of the mechanism.

Blockchain Interoperability

Dean Tribble | Agoric Dean is weaving together multiple blockchains to work together in a massive network. The 'interchain' is a collection of 50 blockchains connected by the Inter Blockchain Communication (IBC) Protocol that forms a network of distributed commerce. Usually, to bridge blockchains, there is a pseudo-trusted third party that intermediates communication between two blockchains. IBC is like TCP for blockchains - it handles the bridging securely and automatically without needing third-party parsing of data.

Privacy-preserving Proof of Personhood

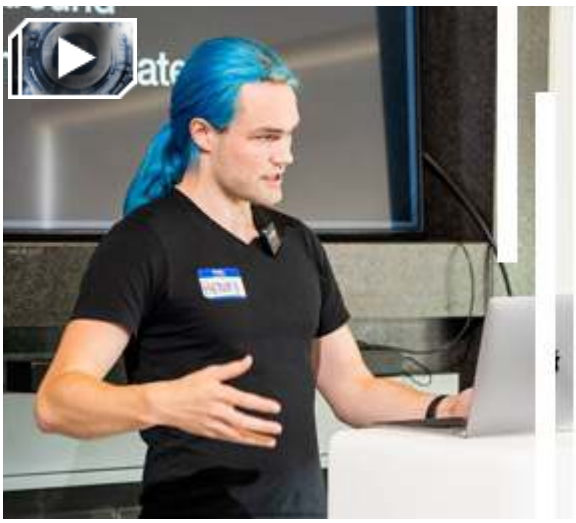
Remco Bloemen | Worldcoin Proof of personhood is one of the most pressing problems on the internet. In an age of bots and account farms, how do we enable online voting or distribution of power when anyone can pretend to be thousands of people all at once? Worldcoin hopes to solve this problem by using biometric data from the eye to validate identity. They want to use this system to enact universal basic income for over one million people without having to deal with welfare fraud.

Turing-Complete Expression Evaluation for Recursive SNARKs

Chhi'mèd Künzang | Protocol Labs Protocol labs created a new breed of zero knowledge proofs - SNARKs (Zero-Knowledge Succinct Non-Interactive Argument of Knowledge) - used for computation of functions using cryptographically secured data. Building upon the previous work to create decentralized data storage via the InterPlanetary File System, they are improving SNARKs by using recursion to have SNARKs act as prerequisites for validating other SNARKs. This new system is being handled by Lurk, a specific implementation of the computer language Lisp, for their own zero knowledge technology protocols. This system can be generalized for other purposes and improves how SNARKs function overall.

Private Decentralized Exchange

Henry de Valence | Penumbra Labs Real world coordination requires control over information. Henry argues that information leaks are value leaks, and that value is lost whenever markets leak information unnecessarily. Privacy, then, is a way of retaining that value and market systems that allow privacy should by default outcompete forced public markets. Zero-knowledge proofs are one way to achieve this but it requires executing off-chain. Penumbra is developing a new methodology for encrypting trading information while maintaining the integrity of the public ledger.



Private ML for Healthcare

Matthew McAteer | Formic Labs Machine learning has amazing potential in the medical field. Automated analysis of large volumes of medical data could save lives and millions of dollars. However, blackmail and other threats creates an imperative to keep these records private. Data theft, model theft, and a host of other vulnerabilities may prevent machine learning from being employed for healthcare. Privacy-preserving ML using federated learning and homomorphic encryption could be the solution. The technology is still emerging but the initial results are promising and may lead to a new era in healthcare data management.

Cryptography in the Cryostat, Topological Phases of Matter

Michael Freedman The DNA of humans are essentially error-correcting codes. Michael goes on to talk about extremely complex condensed matter in the form of crystal interfaces between metal alloys. The ground state for this matter is determined by cryptographic principals. As we move toward quantum computing, we need to fully understand the nature of excitation states and state braiding in these cryptographically protected systems.

Automating Complex Reasoning

Amanda Ngo | Ought Ought is a nonprofit machine learning research lab. They built elicit.org, a machine learning assisted research tool. The way they are using machine learning is by breaking down the steps of research into simple, answerable questions for language models rather than attempting to pass judgement on the validity of data without context. Amanda demonstrated the tool live during the presentation, which is accessible at elecit.org right now.

Deflecting The Sacred

Robin Hanson | George Mason University When attempting to affect change in the world, you will inevitably run up against concepts that others consider sacred. This creates a very tough barrier to change, especially if you are trying to change something like democracy, family, or religion. The essence of sacred is the bond shared between those who consider a particular idea sacred. It's difficult to see things the same when seeing them in high resolution, so sacred things tend to be seen in abstract detail, even when looking at them up close, to allow for concensus.

Challenges on the Path to Decentralized AI

Jonathan Passerat-Palmbach | Imperial College GPT-3 is a foundation model for many AI applications. AI such as GPT-3 are highly centralized due to a number of logistical and security reasons. However, Jonathan believes such AI wastes resources and is inherently unfair. To build a decentralized AI, we will need the right governance, privacy, and incentives. Jon works on building the right models to understand these concepts in the context of cryptography. Apart from the structure of the underlying logic, one of the larger practical problems for decentralized AI is how to finance it as a public good.

Windfall Trust

Anna Yelizarova | Future of Life Institute The windfall clause is a promise for any AI companies to donate much of their windfall if they experience astronomical profits. Such profits would be on the order of 1% of global GDP. The windfall trust is an entity designed to manage this abundance should it ever come to fruition. The assets the trust holds would belong to all of humanity, with an implementation of universal basic income to the poorest first to create a steadily rising income floor. The institute is attempting to answer legal questions about the process as well as figure out governance and distribution methods.

Project Presentations

Decision Auctions

David Ernst, Secure Internet Voting
Kate Sills, Independent

What are you trying to do?
We want to build a way to find group consensus that maximizes happiness for both the people who get their way and the people who don't.

How is it done today? What are the limitations of the current system?
It is often done with many rounds of negotiation that might involve informal guessing and feeling each other out, with the results possibly causing resentment or inferior outcomes.

What is new in your approach and why do you think it will be successful?
We are assigning dollar values to preferred outcomes: "It is worth $6 to me for us to go with X".

If successful, what difference will it make?
People who are choosing in groups, such as couples, families, colleagues, workshop co-participants, friends, housemates will have a new toolset available.

How much will it cost?
It does not require sophisticated technology, and can be used now. We know of one couple who has been successfully using the technique for over a decade. It may increase transaction costs in the beginning and the learning curve may be steep, but we've heard claims that negotiation using this technique may be more efficient than the status quo, after practice.

How long will it take?
The approach is easy to adopt, but it does require a cultural shift to "ask" culture and a rejection of "guess" culture, which may take time to adjust to for some people.

What are the mid-term and final exams to check for completeness?
We need to validate whether the technique is being used - whether people initially choose to adopt it and also whether initial adopters continue to use the technique.

ZK ML

Chhi'mèd Künzang, Protocol Labs
Joel Thorstensson, Ceramic Network
Jonathan Passerat-Palmbach, Imperial College
Deepak Maram, Cornell University
Remco Bloemen, Worldcoin

What are you trying to do?
We want to facilitate and incentivize large-scale cooperative machine learning while preserving individual data and model privacy.

How is it done today? What are the limitations of the current system?
Current federated learning uses privacy-enhancing techniques, but it is centralized and people rightly don't trust it.

What is new in your approach and why do you think it will be successful?
The full realisation of our project would see the creation of a decentralised environment where autonomous and human actors would contribute to improve and leverage on various instances of artificial intelligence.

If successful, what difference will it make?
We have seen a growing number of initiatives deploying various forms of collaborative learning in the scientific community, thus highlighting the appeal of early adopters to mine more than just their own data sources to train more powerful and unbiased models.

How long will it take?
The mid-term will take 6-12 months while the full vision may require multiple years to complete.

What are the mid-term and final exams to check for completeness?
Verifiable evaluation of ML models (succinct zk proofs) will enable innovative use cases such as an Iterated prediction market (to establish a model's value) and a Market for model queries (verified use of known models as a service.)

The final product is fully decentralised model training. Model IP will be protected via either FHE or MPC. The owner obtains strong guarantees that the decentralised training was performed according to their preset conditions. Individual contributors and institutions are rewarded in accordance with the significance of their contribution.

Digital Ghosts

Deger Turan, AI Objectives Institute
Evan Miyazono, Protocol Labs
Niamh Peren, Foresight Institute
Martin Karlsson, Discourse Graphs
Ryan Singer, Chia
WhyRUSleeping, Protocol Labs

What are you trying to do?
We are unblocking the ability to talk to someone who's time-poor, deceased, or the composite of a group.

How is it done today? What are the limitations of the current system?
You can read someone's intentional writing & notes and generate group composites using consensus mechanisms.

What is new in your approach and why do you think it will be successful?
This will succeed because disk space is cheap, ML is getting cheaper, and we have language models that can generate a model of an individual or a composite of people.

If successful, what difference will it make?
You'd be able to have virtualizable people; enables high parallelization and delegation to a highly trusted individual or unblock groups who are operating in an "ask forgiveness, not permission" mode.

How much will it cost?
If we want to fine-tune a model of ~100 people, that'd cost on the order of $200; $40k for an org of 10k people +2 engineers for 4-6 months (@$150k/yr), ~ $500k for an initial prototype + testing with 4 organizations.

How long will it take?
It would be 4-6 months to a first prototype.

What are the mid-term and final exams to check for completeness?
Prototype in 4 months, test with companies in 6 months to validate success.

What we want/need
Funders would be great; Novel monetization strategies for open source products, and LLMs with open weights.
Contact: deger@objective.is

Norm Hardy Prize for usable security as part of civilization and AI security

Mark S. Miller, Agoric
Morgan Livingston, TechCongress Fellow
Allison Duettmann, Foresight Institute
Austin Liu, Cornell University
Dean Tribble, Agoric
Mikayla Maki, Zed

What are you trying to do?
We suggest building interaction design for usable security. Make it common to develop user interfaces in which the easy way is the secure way, i.e. "the actions that users want to take, they naturally take in a secure fashion".

How is it done today? What are the limitations of the current system?
Civilization's computer infrastructure is patchwork and insecurable. Problems include attacks on electric grid, major financial meltdowns, and other attack vectors propagated by advanced AI.

What is new in your approach and why do you think it will be successful?
Make all decisions an explicit part of the user interface, which will allow users to understand the implications of their actions. The entire problem is too large to tackle. What we can do is create a prize, i.e. the Norm Hardy Prize, to honor the late computer security pioneer and promote the secure use of computers.

If successful, what difference will it make?
Voluntary cooperation is only meaningful when based on informed consent. UIs are where the human world meets the crypto world. The UI security issue *is* the issue of enabling the human to understand what they are consenting to.

How much will it cost?
Ideally, $10K annually for 5 years. Plus $5K for admin, physical award, travel stipends to speaking engagement, press release and PR.

How long will it take?
The prize will be given annually.

What are the mid-term and final exams to check for completeness?
To check for prize awardees quality: use rank votes for submissions.
To check for success of prize purpose, determine if prize winners' systems are adopted in widely used systems afterward.

Reforming Academic Incentives

Robin Hanson, George Mason University
Philipp Koellinger, DeSci
Kydo, Stanford University
Alex Aleksyenko, Independent
Ryan Grant, Independent
Nick Matthew, Figment Capital
Eduardo da Veiga Beltrame, ImYoo

What are you trying to do?
We want to create reliable ex-post measures of prestige. A long-term prediction market could be set up today to measure a consensus of experts (to be chosen one hundred years in the future) about what current research most impacted valued outcomes.

How is it done today? What are the limitations of the current system?
Current prestige is measured by journal publication, job title, citations, and grants. Prestige is constructed in and informal/social way. Having prestige doesn't equate to having scientific insight.

What is new in your approach and why do you think it will be successful?
A long-term prediction market would separate the living human career concerns from the ultimate value of the science. It is more worth trying than known to work.

If successful, what difference will it make?
Academia/science would make more/better useful intellectual progress. Broader public may begin talking about this market, thus showing interest in what science research is effective.

Cost and Timeline?
Expert analysis may need 6 profs for 6 years: $5m; then use $100m to pick 10 people on the basis of these market.

What are the mid-term and final exams to check for completeness?
Mid-term exams:
Is there prediction market activity?
Can we find a way to rate past progress?
Final exams:
Is the prediction market still in operation?
Is academic world influenced by prices?

Modeling Theory of Change

Amanda Ngo, Ought
Shaun Conway, iXO
Isabella Duan, University of Chicago
Kevin Compher, IN-Q-TEL

What are you trying to do?
We want to help decision makers to make effective choices that are transparent and easy to understand.

How is it done today? What are the limitations of the current system?
Opaque (black box group-decisions) context around allocation of resources and program design to achieve measurable outcomes. Limited transparency create difficult to troubleshoot steps, thus hindering discovery of optimal paths.

What is new in your approach and why do you think it will be successful?
Our approach uses large language models semantically to recommend process to optimize outputs.

If successful, what difference will it make?
Find, evaluate, optimize best processes across nearly any outcome, for instance creating better jobs/increase income. The public or other stakeholders that can hold decision-makers accountable.

How much will it cost?
The estimated cost is roughly $10k.

How long will it take?
It could be very soon. We may need GPT4 or 5 to get it to the point of excellent proposals.

What are the mid-term and final exams to check for completeness?
Mid-term exams: Create a generalized/abstracted system that works across a diverse set of outcomes.
Final exams: Evaluation and feedback of the model against tested proposals.

Who is interested in exploring the working project further?
Amanda: I'm interested in helping people build a prototype (amanda@ought.org)
Kevin: Follow-up with team, gauge interest in collaborating on prototype. (kcompher@iqt.org)

Windfall Trust

Anna Yelizarova, Future of Life Institute
Silke Elrifai, Independent
Michael Freedman, Field's Medalist

What are you trying to do?
Set up an internationally robust legal set of entities to implement global UBS and be the target of a WC.

How is it done today? What are the limitations of the current system?
We need to learn more about the following entities:
Swiss Association (agility and ease of prototyping)
Guernsey Trust Structure
Gibraltar Company by guarantee lease share capital
Islamic Perpetual Trust WAQF
Singapore company limited by guarantee

What is new in your approach and why do you think it will be successful?
Success will depend on future Proofing some of these legal design challenges for a world that is hard to predict.

If successful, what difference will it make?
If successful, it could help all humanity build resilience for the economic effects of automation, job loss and growing inequality. Even if Windfall is never triggered, this provides a infrastructure for global Universal Basic Income.

Cost and Timeline?
Without seeding the Windfall Trust with funding, around 100k assuming pro bono work.
An initial prototype can be set up in a year, but something more sophisticated would take >5 years.

What are the mid-term and final exams to check for completeness?
The initial milestones involve completing the paper and setting up a legal entity that can prototype some of these concepts and allow philanthropic donations to be directed through partners working on distribution.