Foresight Institute

# Foresight Institute's Existential Hope Transformative AI Institution Design Hackathon

SALESFORCE TOWER, SAN FRANCISCO, CA, USA

February 5 - 6, 2024

In collaboration with Future of Life Institute

Report writer: Beatrice Erkers

# Table of Contents

# Table of Contents

# About: Foresight Institute and Future of Life Institute

### About Foresight Institute

Foresight Institute supports the beneficial development of high-impact technology to make great futures more likely. Foresight Institute focuses on science and technology that is too early-stage or interdisciplinary for legacy institutions to support, including biotechnology, nanotechnology, neurotechnology, computation, and space exploration. Foresight Institute awards prizes, offers grants, supports fellows, and hosts conferences to accelerate progress toward flourishing futures and mitigate associated risks.

### About Future of Life Institute

Future of Life Institute (FLI) is a non-profit organization focused on steering transformative technologies away from extreme, large-scale risks and towards benefiting life. FLI's work includes grantmaking, educational outreach, and advocacy within the United Nations, United States government, and European Union institutions.

# Executive Summary

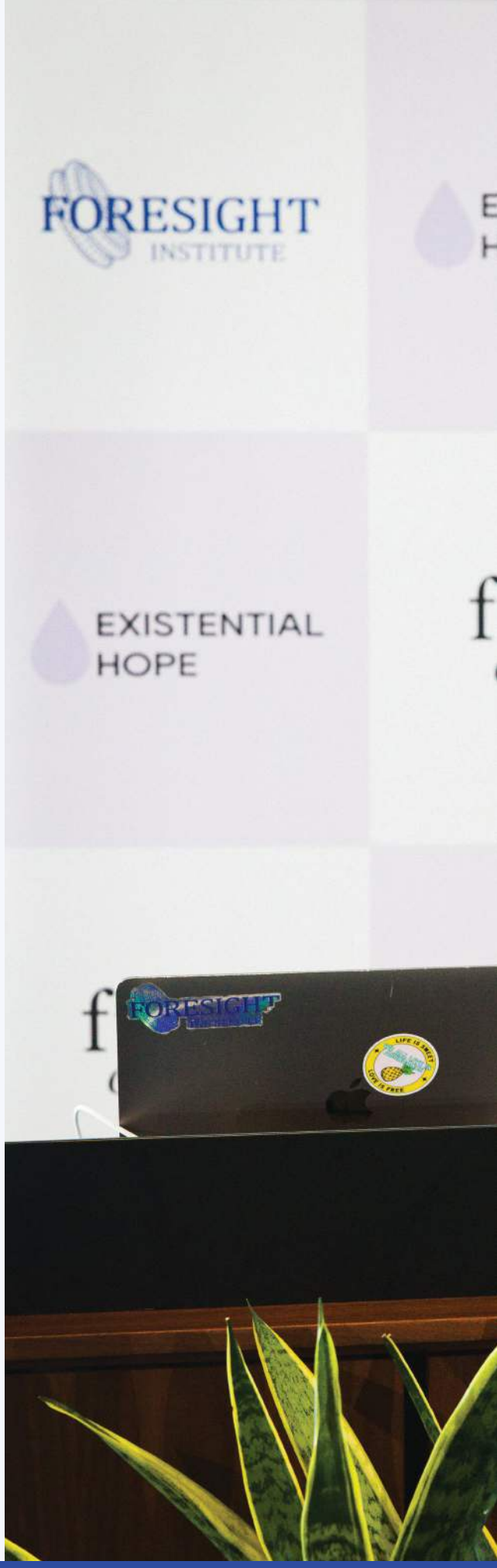
Foresight Institute, in partnership with FLI, hosted an Existential Hope Transformative AI Institution Design Hackathon on February 5-6, 2024, in San Francisco, to help catalyze the development of institutions that can steer the evolution of Transformative AI (TAI) towards outcomes that ensure a flourishing future for humanity.

Our working definition of TAI for this hackathon is potential future AI that precipitates a transition comparable to (or more significant than) the agricultural or industrial revolution.

This event brought together 61 participants (including mentors and judges) from a wide array of fields, including leading researchers, policymakers, and AI practitioners, to design possible institutions that could aid the beneficial development of TAI and its applications. This hackathon expanded upon discussions during our 2023 Existential Hope Day, which focused on reflecting and exploring positive future trajectories, including the potential of TAI.

Over the course of two days, hackathon participants developed potential solutions to bridge the gap between the potential of TAI and the existing institutional frameworks capable of guiding its development towards beneficial ends. Participants first identified the top 36 most important goals for TAI, before working to assess the adequacy of current institutions in addressing these aims.

Next, participants then leveraged the insights gathered and started conceptualizing new institutions that could effectively address the identified challenges and opportunities, or current institutions that could be re-designed to better address them. After listing these



# Executive Summary

proposals, the participants prioritized the top nine most tractable institution propositions to work on for the remainder of the hackathon.

Teams then refined their concepts with a special focus on the practical implementation of these institutions. They worked on creating practical institutional prototypes which could be created in the short time of the hackathon, such as for example a policy draft, a test case, or a survey. This final stage concluded with a series of presentations, where each team showcased their institutional prototype to a panel of judges.

Nine promising institutional sketches and their first prototypes were the key outcomes of this hackathon, designed to steer the next steps of ensuring successful TAI governance. These institutions were:

1. **The Flourishing Foundation (Hackathon Winner):** Aims to create well-being metrics and certifications for AI products to ensure they promote human well-being..
2. **The Global Deliberation Coordinator (Hackathon shared second place):** Focuses on establishing a platform for global discussions and decision-making on AI and other pressing issues.
3. **The Scenario Planning Institution (Hackathon shared second place):** Develops scenarios to explore potential, underrepresented futures of AI and its societal impacts.
4. **The Evals for Evals Institute:** Works on creating a standardized process for evaluating AI systems before launch, focusing on safety, fairness, and transparency.
5. **The World Convention on Transformative Artificial Intelligence:** Aims to prepare for a global conference on AI governance through diplomacy and research.
6. **The Common Knowledge Generator:** Working towards launching a platform with verified information on complex issues to improve understanding and decision-making.
7. **The Delphi Collaboration Protocol:** Seeks to build a platform for collaborative world modeling and scenario analysis to inform data-driven decisions in the future.
8. **The Open Source BCI Project:** Strives to create an open-source brain-computer interface (BCI) operating system to enhance human cognitive abilities and privacy in the AI era.
9. **The SECHI Institute:** Aims to design a software system for “Simulation-Enabled Cooperative Human Intelligence” (SECHI) to manage Earth using a model predictive control approach.

# Executive Summary

To incentivize and support the continuation of the work initiated during the hackathon, the event concluded with the distribution of developmental grants to the most promising projects. The winning team was awarded $10,000 – recognizing their outstanding contribution and the potential impact of their proposed institutional prototype. In acknowledgment of the high caliber of submissions, two teams were selected as runners-up, each receiving $5,000. These grants were intended to provide financial support for the further development of the institutions, facilitating the winners and runners-up in their efforts to bring their concepts into reality.

# Hackathon Goals

This hackathon focused on generating institutional sketches and prototypes that could help build positive futures from TAI.

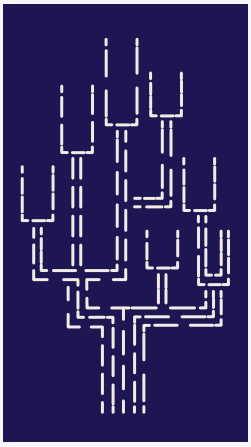
Hackathon goals were:

1. Identifying goals for a future defined by positive TAI and evaluating the role of existing institutions in guiding AI development toward these goals.
2. Designing sketches for institutions that are better positioned to reach the positive TAI goals and develop prototypes of such institutions.
3. Evaluating the new TAI institutional concepts and prototypes and exploring the next steps to initiate the creation of the leading institutions.

# Hackathon Format

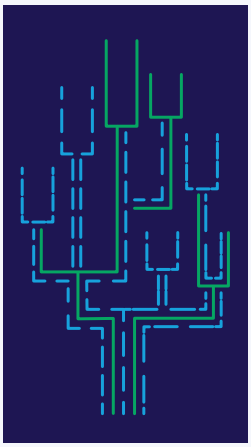
The Existential Hope TAI Institutions Hackathon unfolded over two days, segmented into three distinct phases:

## Phases



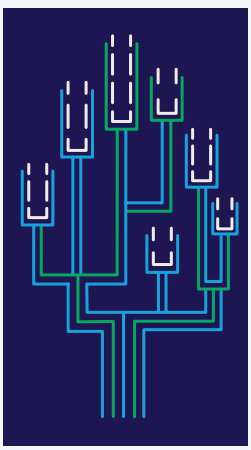
**1. SCAN:**

The SCAN phase set the stage by engaging participants in an exploration to identify and articulate TAI goals, aiming to create a shared understanding of the positive futures we might strive towards with TAI. This was coupled with an assessment of the current landscape of institutions, evaluating how existing frameworks align with, or fall short of, these aspirational goals. Through collaborative discussions, participants explored the gaps and limitations of current institutions in achieving the set TAI objectives.



**2. FOCUS:**

In the FOCUS phase, the hackathon harnessed the collective insights from the SCAN phase to form dedicated teams around the most compelling TAI goals. These teams worked together to conceptualize improved institutions that could effectively bridge the identified gaps. Efforts were concentrated on sketching out innovative institutional designs, refining these ideas into actionable prototypes, before subjecting the emerging concepts to rigorous feedback through a process of red-teaming. This critical evaluation aimed to challenge assumptions, test the resilience of the proposed solutions, and honing the concepts.



**3. ACT:**

Transitioning to the final phase, – ACT, the emphasis shifted towards the practical aspects of bringing the institutional prototypes closer to reality. Teams engaged in intensive sessions to detail the development process for their prototypes and strategized on scaling these ideas into fully functioning institutions. The final part of this phase was a series of presentations where each team showcased their prototypes to judges and other participants, for constructive feedback and collaborative refinement.

# Hackathon Format

## Prototyping Process

- **Drafting the institutional sketch:** Teams outlined the foundation of their proposed institution by identifying team members, confirming the TAI goal they aimed to address, and highlighting the unique attributes of the institution they envisioned. This included governance structure, technological reliance, and innovative approaches to overcome existing institutional shortcomings.
- **Designing the prototype:** With the institutional framework in place, teams began to create a meaningful prototype that could realistically demonstrate the institution's potential impact. This prototype could take various forms, such as an app, platform, policy draft, or simulation, and was designed to be a tangible representation of the institution's capability to achieve its stated goal.
- **Scaling from prototype to institution:** Participants outlined a roadmap for evolving their prototype into a fully operational institution. This involved detailing milestones, identifying key stakeholders, and planning initial actions; supported by a rough timeline and budget estimates, especially in light of the $10,000 development grant that the winning team would receive to financially aid in taking the first steps to realize their institution.

## Judging Process

Judges evaluated the institutions and their prototypes based on a comprehensive set of criteria, each scored from 1 (low) to 100 (high):

- **Relevance and impact:** The expected positive impact of the institution on society, focusing on how significantly the institution could contribute to addressing the targeted TAI goal.
- **Practical feasibility:** The practicality of implementing the proposed institution and realizing its intended impact, considering the detailed plan from prototype to full institution.
- **Prototype quality:** The effectiveness and quality, focusing on how well the prototype represented the proposed institution's capabilities and potential for success.

# Workshop Participants, Mentors, and Judges

## JUDGES

### Anthony Aguirre
EXECUTIVE DIRECTOR
**FUTURE OF LIFE INSTITUTE**

### Christine Peterson
CO-FOUNDER
**FORESIGHT INSTITUTE**

### Emilia Javorsky
DIRECTOR OF THE FUTURES PROGRAM
**FUTURE OF LIFE INSTITUTE**

### Hannu Rajaniemi
CO-FOUNDER AND CEO
**HELIX NANOTECHNOLOGIES**

### Tom Kalil
CHIEF INNOVATION OFFICER
**SCHMIDT FUTURES**

### Kipply Chen
TECHNICAL STAFF
**ANTHROPIC**

## MENTORS

### Andrew Maynard
PROFESSOR AT THE SCHOOL FOR THE FUTURE OF INNOVATION IN SOCIETY
**ARIZONA STATE UNIVERSITY**

### Anna Yelizarova
SPECIAL PROJECTS LEAD
**FUTURE OF LIFE INSTITUTE**

### Christian Tarsney
PROFESSOR OF PHILOSOPHY
**UT AUSTIN**

### Darren McKee
SENIOR POLICY ADVISOR
**ARTIFICIAL INTELLIGENCE GOVERNANCE AND SAFETY CANADA**

### Robert Trager
CO-DIRECTOR
**OXFORD MARTIN AI GOVERNANCE INSTITUTE**

# Workshop Participants, Mentors, and Judges

## PARTICIPANTS

### Amanda Ngo
PRODUCT MANAGER
**EX OUGHT**

### Aviv Ovadya
RESEARCH FELLOW
**NEWDEMOCRACY**

### Bear Haon
QUAD FELLOW
**SCHMIDT FUTURES**

### Bogdan-Ionut Cirstea
AI EXISTENTIAL SAFETY RESEARCHER
**INDEPENDANT**

### Brandon Goldman
PARTNER
**LIONHEART VENTURES**

### Bryce Hidysmith
RESEARCHER

### Chandler Smith
RESEARCH SCHOLAR
**ML ALIGNMENT & THEORY SCHOLARS**

### Colleen McKenzie
DIRECTOR OF STRATEGY
**AI OBJECTIVES INSTITUTE**

### Connor McCormick
RESEARCHER
**GAIA INSTITUTE**

### Deger Turan
PRESIDENT
**AI OBJECTIVES INSTITUTE**

### Diogo de Lucena
CHIEF SCIENTIST
**AE STUDIO**

### Dusan Desic
OPERATIONS LEAD
**PIBBSS**

### Elyse Lefebvre
INDEPENDENT

### Evan Miyazono
FOUNDER AND CEO
**ATLAS COMPUTING**

# Workshop Participants, Mentors, and Judges

### Fin Moorhouse
RESEARCH ANALYST
**LONGVIEW PHILANTHROPY**

### Jan Kulveit
RESEARCH FELLOW
**FUTURE OF HUMANITY INSTITUTE (UNIVERSITY OF OXFORD)**

### Jelena Luketina
DPHIL SCHOLAR
**FUTURE OF HUMANITY INSTITUTE**

### Joel Christoph
DIRECTOR | PHD RESEARCHER | FELLOW
**EFFECTIVE THESIS | EUROPEAN UNIVERSITY INSTITUTE | ATLANTIC COUNCIL**

### Joel Lehman
RESEARCH ADVISOR
**STABILITY AI**

### José-Jamie Villalobos
RESEARCH AFFILIATE
**LEGAL PRIORITIES PROJECT**

### Joshua Tan
FOUNDER AND LEAD
**METAGOV**

### Judd Rosenblatt
FOUNDER
**AE STUDIO**

### Justin Bullock
SENIOR RESEARCHER
**CONVERGENCE**

### Keenan Pepper
SOFTWARE ENGINEER
**SALESFORCE**

### Konrad Seifert
CO-FOUNDER
**SIMON INSTITUTE FOR LONGTERM GOVERNANCE**

### Kyle Killian
DEPUTY DIRECTOR
**TRANSFORMATIVE FUTURES INSTITUTE**

### Lewis Hammond
CO-DIRECTOR
**COOPERATIVE AI**

### Dr Li Zi
INDEPENDANT

### Ludwig Illies
FELLOW
**NON-TRIVIAL**

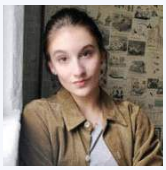
### Malcom Murray
RESEARCH AFFILIATE
**CENTER FOR THE GOVERNANCE OF AI**

# Workshop Participants, Mentors, and Judges

## Mamun Miah
PROJECT SCIENTIST
**LAWRENCE BERKELEY NATIONAL LABORATORY**

## Matteo Pistillo
RESEARCH SCHOLAR
**LEGAL PRIORITIES PROJECT**

## Megan Cansfield
POLICY ANALYST
**U.S. DEPARTMENT OF HOMELAND SECURITY**



## Morgan Livingston
SCHWARZMAN SCHOLAR
**TSINGHUA UNIVERSITY**

## Ozzie Gooen
PRESIDENT
**QUANTIFIED UNCERTAINTY RESEARCH INSTITUTE**

## Saad Siddiqui
WINTER FELLOW
**GOVAI**

## Spencer Kaplan
PHD STUDENT
**HARVARD UNIVERSITY**

## Steve Caldwell
TECH LEAD IN INNOVATION
**AE STUDIO**

## Margarita Geleta
CS PHD STUDENT
**UC BERKELEY**

## Maximillian Negele
CO-FOUNDER AND AI GOVERNANCE LEAD
**CFACTUAL**

## Mingzhu He
CO-FOUNDER
**CONSCIOUS TECH COLLECTIVE**

## Nicolas Miailhe
FOUNDER AND CHAIR OF THE BOARD
**THE FUTURE SOCIETY**

## Ross Gruetzemacher
EXECUTIVE DIRECTOR
**TRANSFORMATIVE FUTURES INSTITUTE**

## Siméon Campos
FOUNDER AND PRESIDENT
**SAFERAI**

## Stephen Clare
RESEARCHER
**CENTER FOR THE GOVERNANCE OF AI**

## Steve Coy
RESEARCHER
**GAIA INSTITUTE**

# Workshop Participants, Mentors, and Judges

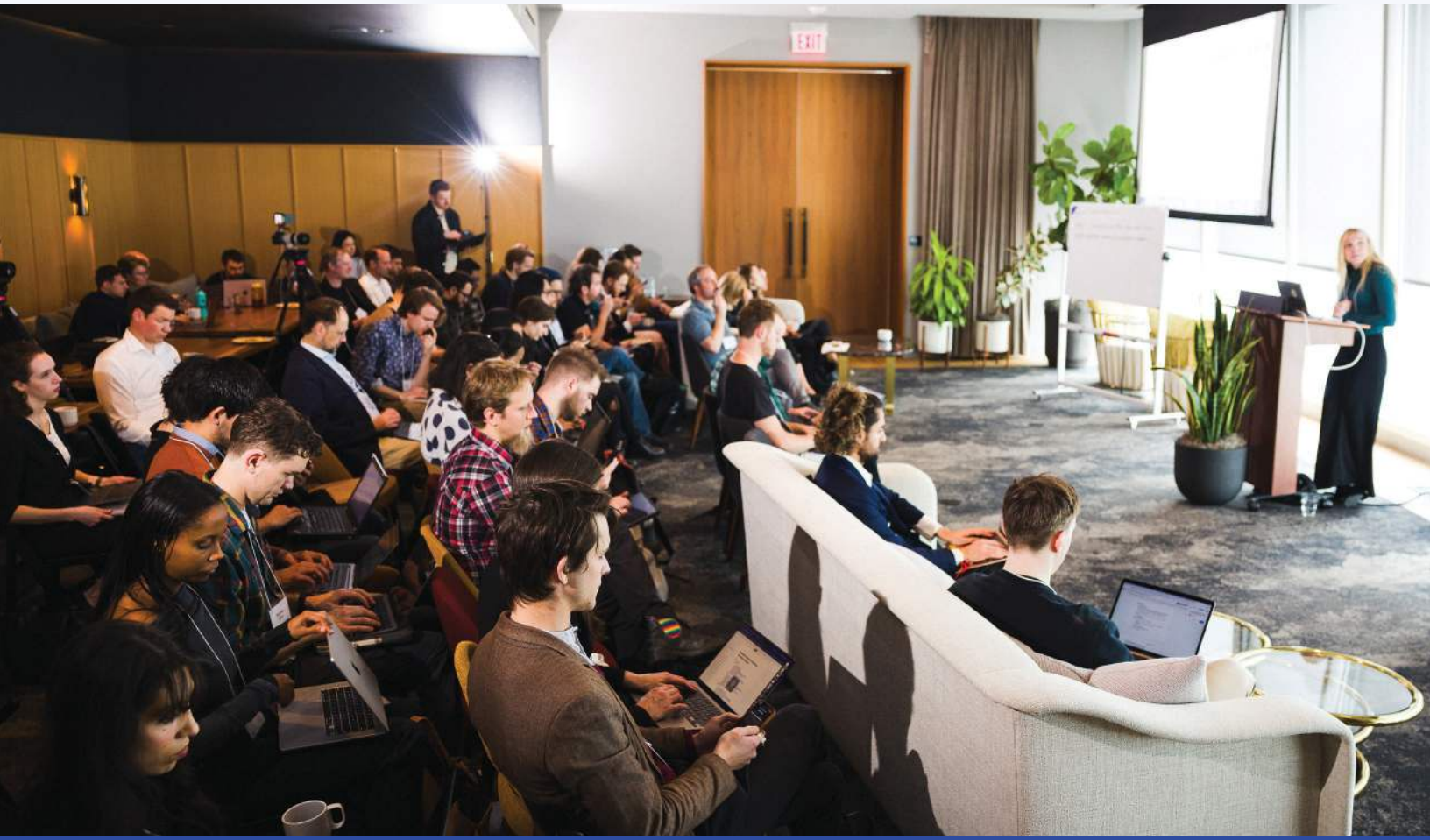
### Tilman Raüker
AI SAFETY RESEARCHER
**INDEPENDANT**

### Toby Pilditch
SENIOR RESEARCH SCIENTIST
**TRANSFORMATIVE FUTURES INSTITUTE**

### Tushant Jha (TJ)
RESEARCH SCHOLAR
**FUTURE OF HUMANITY INSTITUTE**

### Vilhelm Skoglund
CO-FOUNDER AND CEO
**IMPACT ACADEMY**

### Ziya Hunag
OPERATIONS MANAGER
**CONCORDIA AI**



PHASE I:

# SCAN

# Top TAI Goal Categories and Their Existing Institutions

The SCAN session of the workshop focused on identifying top TAI goals – see Appendix for a complete list of all individual goals identified. Breakout groups then formed to expand and deliberate on the leading goal clusters as determined by participant votes – each participant was given three votes. Finally, each breakout group divided the overarching goal cluster into the top goals within each category, and began mapping existing institutions to them.

Below is a list of the top TAI goals grouped into six different categories, along with their individual top goals and institutions. Please note that the listed institutions are not an exhaustive list.

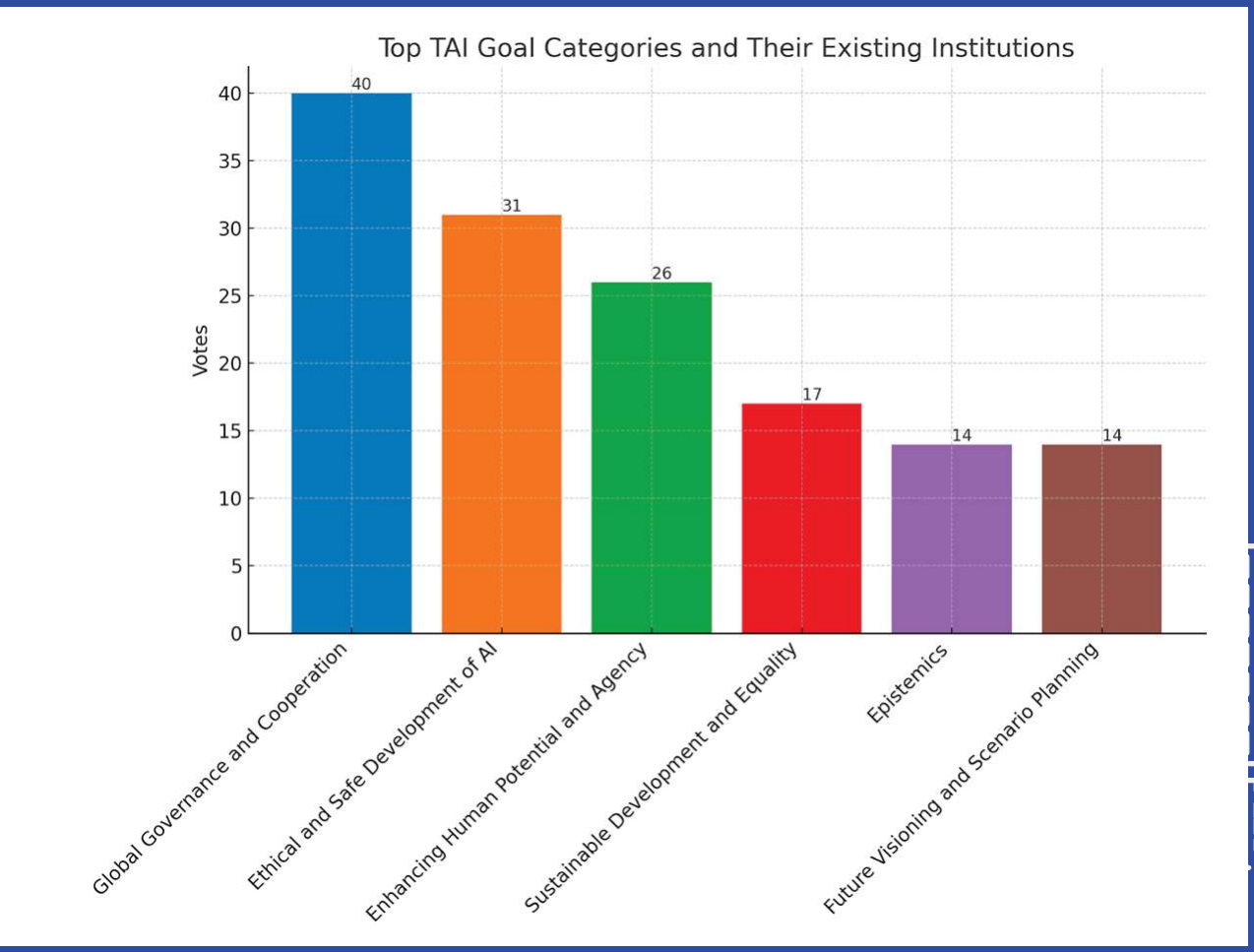



Image 1: Participant Votes of the Most Pressing TAI Goal Categories

# Category: Global Governance and Cooperation
## Participant votes: 40

This category aims to establish a comprehensive framework for international cooperation on AI, focusing on risk management, trusted agreement mechanisms, and enhancing the relationship with existing institutions to ensure global stability and equitable progress in AI development.

**Societal goals in this category include**

- Implement global auditing and risk management.
- Create trusted mechanisms for global agreements through deliberative processes.
- Develop infrastructure and operational capacity for execution of governance initiatives.
- Enhance connections with existing international institutions and map the adoption theory of change.
- Navigate US-China relations and manage the transition of legacy institutions without societal destabilization.
- Establish a Global AI Treaty with an enforcement mechanism.
- Designate a meta-organizer for cross-institutional collaboration.

**Existing institutions focused on addressing these goals include**

- Standards bodies (NIST, IETF, ITU, ISO/IEC), and national AI safety institutes.
- Consultancies (PWC, Deloitte), international forums (UNGA, UNESCO), and deliberative tools.
- NGOs, think tanks (CSIS), universities (Stanford, Harvard), and “neutral” governments.
- Regulations (Chips Act), rule of law initiatives, and global digital compacts.

# Category: Ethical and Safe Development of AI
## Participant votes: 31

This category prioritizes the safe development of AI within ethical guidelines and safety standards through risk assessments, safety coordination, and global monitoring, targeting responsible AI deployment worldwide.

**Societal goals in this category include**

- Conduct AI risk evaluations and develop science-based evaluation metrics.
- Coordinate safety capabilities and standardize AI safety measures.
- Promote global monitoring capabilities and certification/auditing of AI systems.

**Existing institutions focused on addressing these goals include**

- Standard bodies (NIST), red-teaming efforts, and global evaluation initiatives (IPCC for evals).
- Centers for AI ethics and safety (CHAI at UC Berkeley), and international metrology institutes.

## Category: Enhancing Human Potential and Agency

### Participant votes: 26

This category explores ways to boost human flourishing and agency, including through human augmentation and democratic involvement in AI governance, by experimenting with new social models and decentralized governance mechanisms.

**Societal goals in this category include**

- Expand spaces for individual and collective experimentation with new social models.
- Measure and advance human flourishing and agency, including through human augmentation.
- Develop decentralized governance mechanisms and ensure democratic input in AI principles.

**Existing institutions focused on addressing these goals include**

- Meaning Alignment Institute, Collective Intelligence Project, and the OpenAI Democratic Input Grants.
- Educational innovations and initiatives promoting well-being and humane technology.

# Category: Sustainable Development and Equality
## Participant votes: 17

This category focuses on leveraging AI for sustainable development and equality, addressing the integration of AI in solving economic, environmental, and social challenges to promote inclusive growth and mitigate inequalities.

**Societal goals in this category include**

- Address economic system continuity, and incentivize inclusion of negative externalities in models.
- Incorporate AI to achieve sustainability goals and address distribution problems.
- Ensure fair access to AI and redistribute benefits to mitigate inequalities.

**Existing institutions focused on addressing these goals include**

- International frameworks and alliances (European AI Alliance, UN SDGs), financial instruments (Social Impact Bonds, IMF, World Bank).

# Category: Epistemics
## Participant votes: 14

This category concentrates on improving the collective understanding and forecasting abilities related to AI, aiming to safeguard the epistemic commons from misinformation risks and ensure a shared reality through consensus-building and trustworthy analysis.

**Societal goals in this category include**

- Improve shared reality conceptions and forecasting abilities.
- Prevent AI-driven degradation of epistemic commons and manage misinformation risks.

**Existing institutions focused on addressing these goals include**

- Initiatives for consensus-building and objective risk assessment in AI (global forecasting institutions).
- Tools for enhancing trustworthy analysis and facilitating collaboration (automated Wikipedia, expert-on-demand infrastructure).

# Category: Future visioning and Scenario Planning

## Participant votes: 14

This category focuses on anticipating future challenges and opportunities through sophisticated world modeling and creative storytelling, to navigate potential futures with improved foresight and preparedness.

**Societal goals in this category include**

- Build sophisticated world models for broad future scenario planning.
- Employ storytelling and horizon scanning to identify future challenges.

**Existing institutions focused on addressing these goals include**

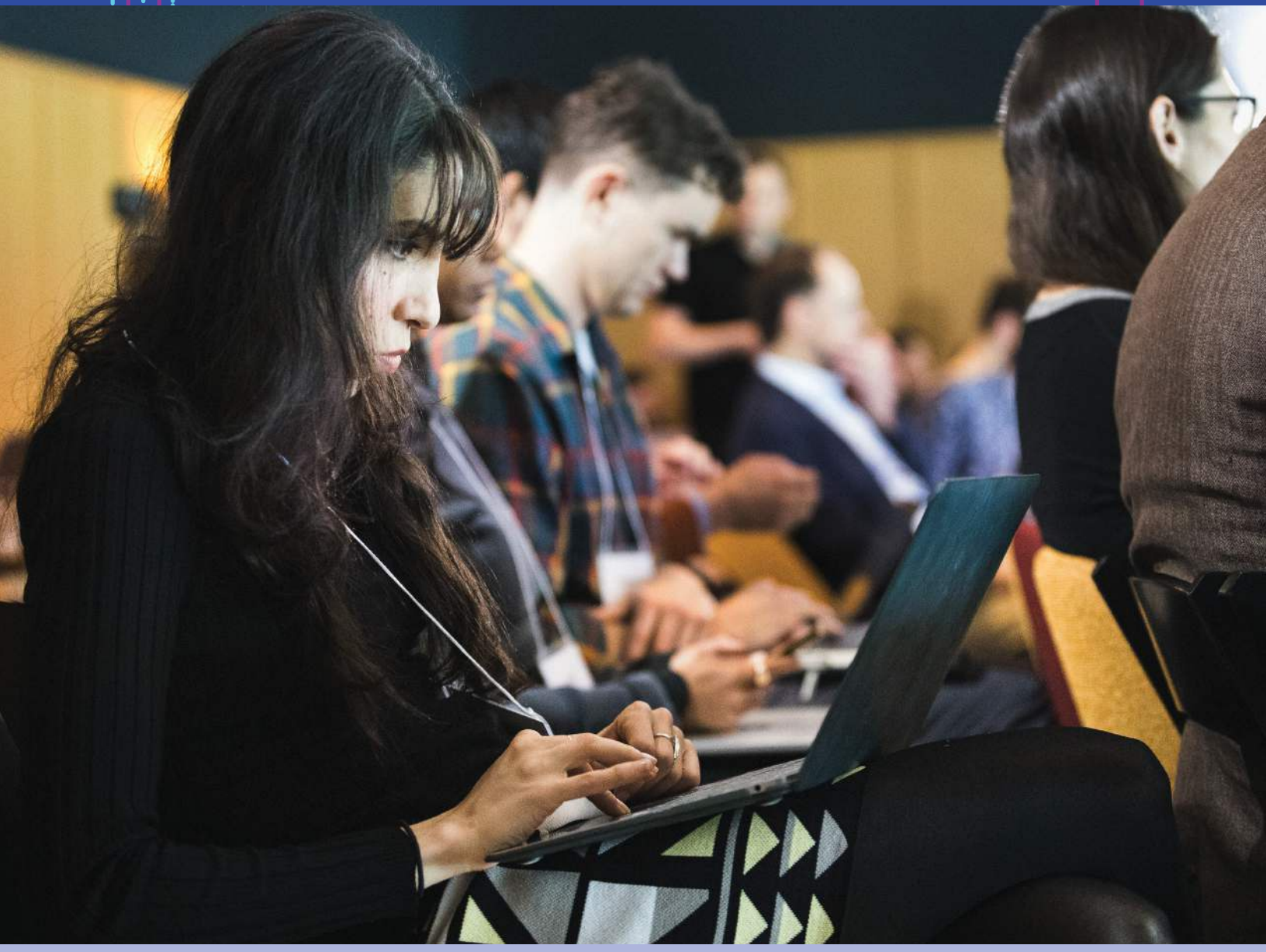
- Research and development centers focused on future studies (RAND, Singapore Centre for Future Studies).
- Collaborative platforms for speculative worldbuilding and forecasting (Metaculus, FLI, science fiction communities).

# Conclusion of the SCAN phase

The SCAN phase effectively set the direction for TAI governance development, identifying six main goal categories for TAI. These categories address global cooperation, safety, ethics, human potential, economic and social equity, misinformation management, and future scenario planning. Moving forward, the focus shifts to refining goals, crafting action plans, and either leveraging or establishing institutions to bridge identified gaps to ensure TAI's safe and beneficial progression.

PHASE II:

# FOCUS



In the FOCUS phase of the hackathon, participants leveraged their insights from the initial SCAN phase to narrow down the TAI goals to the most tractable ones, and form teams centered on addressing their current challenges. Thank you to our hackathon mentors Andrew Maynard (Arizona State University), Anna Yelizarova (Future of Life Institute), Christian Tarsney (UT Austin), Darren McKee (Artificial Intelligence Governance and Safety Canada), and Robert Trager (Oxford Martin AI Governance Institute) who advised the groups to help refine the proposals. Below is a list of the nine top TAI goals as prioritized by the participants, along with their individual challenges and proposed ways to address these challenges.

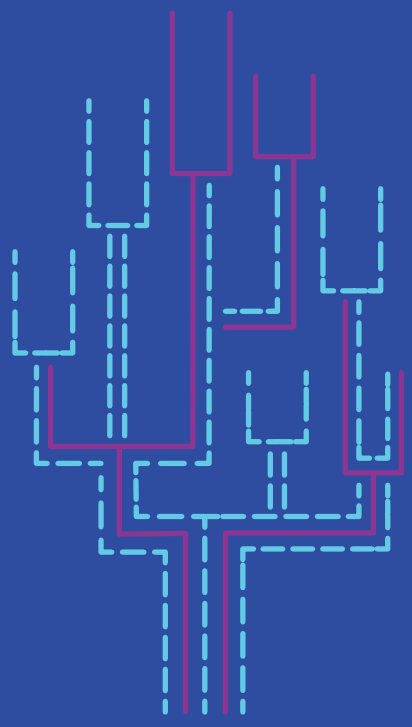
# Top Nine TAI Goals: Challenges and Design Improvements

## GOAL: ENHANCING SAFE AI THROUGH RIGOROUS EVALUATION SYSTEMS

As AI systems permeate various aspects of society, the need for robust evaluation frameworks, akin to the Consortium for AI Risk Evaluations, becomes critical. These frameworks must address the widening gap between the rapidly advancing functionalities of AI models and the existing evaluation methods employed by organizations like the Food and Drug Administration (FDA), National Institute of Standards and Technology (NIST), EU AI Office,Federal Aviation Administration (FAA), and RAND Corporation. This necessitates a collaborative effort towards developing a standardized, transparent, and effective system for evaluating AI risks.

### Challenges of Current Institutions

- Balancing the inherent complexity of biological systems with the "black-box" nature of AI, particularly Language Models (LMs). This can lead to relying on statistical averages instead of individual evaluations, as seen in the FDA's struggles.
- Inadequate infrastructure, expertise, and funding for comprehensive AI evaluations, exemplified by NIST's challenges.
- Difficulty attracting skilled personnel in AI safety and evaluation, as encountered by the EU AI Office, – highlighting a broader issue.
- Lack of understanding of advanced AI models and prevalent software cultures, as in the FAA,'s case, emphasizing the need for increased technical literacy and adaptability.
- Maintaining independence and mitigating bias in evaluation processes, a challenge faced, for example, by the Frontier Model Forum (FMF).
- The absence of shared definitions and terminology across institutions, hindering efforts to streamline and standardize AI evaluations.



### Avenues for Addressing These Challenges

- Develop common definitions and share best practices for conducting evaluations.
- Encourage cooperation among stakeholders and manage conflicts of interest.
- Provide sufficient funding, infrastructure, and communication channels for effective evaluations.
- Recognize the need for a variety of evaluation methodologies to address different AI risks.
- Establish a network for knowledge exchange, harmonization of standards, and collective risk mitigation.

## GOAL: TOWARD TRUSTED DELIBERATIVE MECHANISMS

This objective seeks to establish new, reliable mechanisms for global agreement and decision-making, based on open discussions. This aims to overcome the limitations of existing institutions and frameworks, such as the Ethical Framework for Artificial Intelligence (EFT), UN General Assembly (UNGA), and UNESCO, which have traditionally led global cooperation and decision-making. Whilst valuable, these entities currently face restrictions hindering their effectiveness in addressing the complex challenges of trustworthy AI.

### Challenges of Current Institutions

- The lack of a random selection process for decision-makers undermines the fairness and democratic legitimacy of the governance process.
- Inadequate translation services impede meaningful participation and understanding from diverse global audiences.
- Uneven access to specialized knowledge across regions and sectors hinders well-informed decision-making.
- Challenges in achieving representation and buy-in within existing power structures undermine the potential for inclusive and effective governance.
- The process of gathering input, interpreting it, and translating it into actionable outcomes lacks efficiency and transparency.
- Ensuring decisions are understandable and verifiable by powerful institutions.
- Ambiguity regarding the scope of decisions, (e.g. addressing global risks) and whether they are recommendations or binding rules weakens the governance framework's impact.

# Phase II: FOCUS

### Avenues for Addressing These Challenges

- Leveraging existing frameworks like the UNGA, UNESCO, and citizen panel processes.
- Utilizing existing deliberative tools, such as Polis in Taiwan.
- Exploring alternative voting systems, like quadratic voting.
- Utilizing collective intelligence projects for decision-making.
- Utilizing practices from platform governance, such as traditional 'trust and safety' teams who function to ensure user protection.

## GOAL: ENHANCING US-CHINA DIALOGUE ON MILITARY AI

The complexities of US-China relations, particularly regarding military AI, demand innovative solutions in global governance. Existing Track II diplomatic dialogues facilitated by organizations, such as the Geneva Center for Security and various NGOs, have stimulated discussion, but many challenges continue to slow progress.

### Challenges of Current Institutions

- Many negotiation forums remain inactive, lacking proactive engagement with NGOs and focusing on immediate crises, rather than longer-term issues and solutions.
- Cohesive efforts among NGOs, think tanks, and universities are scarce, hindering sustained dialogue and solution development.
- Infrequent and unproductive closed-door meetings fail to address the complexity of these issues.
- Diplomats often lack the technical understanding for meaningful discussions on military AI, instead, focusing on traditional intelligence gathering.
- Financial limitations within ministries restrict the scope of specialized discourse.
- A lack of trust and insufficient incentives hinder constructive engagement.
- Domestic political agendas, such as regime stability and re-election, divert attention from and actively hinder cooperative AI governance.

### Avenues for Addressing These Challenges

- Existing forums require revitalization with a focus on long-term solutions and broader NGO participation.
- Equipping diplomats with relevant technical expertise to facilitate meaningful discussions on military AI.

- Increased funding for specialized discourse within ministries.
- Finding common ground beyond immediate political concerns.
- Establishing guidelines to determine when to prioritize bilateral, minilateral, or multilateral approaches for international cooperation. This framework should include considerations for escalation, de-escalation, and parallelization of efforts.

## GOAL: ACHIEVING EFFECTIVE AI REGULATION

Existing treaties fail to address the unique challenges posed by AI and often lack sufficient enforcement. This highlights the urgent need for a regulatory approach that balances the fast-paced innovation of the private sector with the structured oversight of government initiatives. To achieve this, we need a more adaptable, inclusive, and enforceable framework than what currently exists.

### Challenges of Current Institutions

- The tension, and resulting trade-offs, between government-led regulation and industry-driven approaches are hindering achieving effective regulation.
- Existing regulations struggle to address the specific challenges and opportunities of AI.
- Weak or absent enforcement mechanisms hinder the effectiveness of current frameworks.
- The current system doesn't effectively discourage countries from acting individually, leading to a "prisoner's dilemma" situation.
- Traditional treaty-making processes are too slow to adapt to the rapid development of AI.
- Low trust levels makes building consensus among nations difficult.
- Balancing the diverse interests of different stakeholders is a significant challenge.

### Avenues for Addressing These Challenges

- Combine industry-driven initiatives with government-led regulations to foster innovation and accountability.
- Tailor global treaties to take into account address the characteristics and implications of AI.
- Implement clear, enforceable mechanisms with incentives for compliance, and penalties for non-compliance.
- Design frameworks which  incentivize cooperation and discourage defection.
- Develop and adopt efficient and adaptable processes for negotiation and implementation that reflect the dynamic nature of AI.
- Create better channels of collaboration between nation-states.

# Phase II: FOCUS

## GOAL: CULTIVATING HUMAN FLOURISHING

Shaping a future that prioritizes human agency demands the creation of adaptable and forward-looking visions that respond to rapid technological and societal shifts. This necessitates a holistic approach that deepens our understanding of well-being, empowers future generations, and embraces agile learning and inclusive decision-making. By adopting this approach, we can cultivate a future where human agency and technological advancements flourish together.

### Challenges of Current Institutions

- Existing metrics, such as the Sustainable Development Goals, do not fully capture the impact of transformative technologies on well-being.
- Conventional thinking used by powerful institutions hinders agility and limits innovative approaches to human development.
- Despite smaller institutions being more adaptable, they lack the power to implement widespread change.
- Insights from philosophical discussions rarely translate into concrete action or policy.
- The emphasis on economic outcomes overshadows broader human well-being and flourishing.
- Current approaches to addressing AI Safety concerns is creating a closed-source AI development ecosystem, limiting any opportunities for open experimentation.

### Avenues for Addressing These Challenges

- Develop new mechanisms for identifying and measuring human values.
- Encourage powerful institutions to adopt more flexible approaches, learning from smaller and more agile entities.
- Create spaces for open experimentation with new technologies and societal models, enabling rapid learning based on real-world feedback.
- Ensure individual and community autonomy by allowing them to opt out of specific technologies or societal experiments.

## GOAL: PRESERVING AND ENHANCING HUMAN AGENCY IN HUMAN AUGMENTATION

The pursuit of human augmentation, particularly through Brain-Computer Interfaces (BCIs), offers potential for significant improvements in our capabilities and lives. However, it raises critical questions about human agency, including individual well-being and our power to shape the future. The key challenge is ensuring augmentation advances ethically: aligning with human goals, preserving our essential human qualities, and upholding the strictest safety and privacy standards. By addressing these complexities, we can empower individuals and society whilst safeguarding a future grounded in human values.

### Challenges of Current Institutions

- There is a lack of clear metrics to evaluate the impact of augmentation on agency, particularly in ensuring long-term alignment with human goals.
- There is no collective decision-making in deciding the current and future use and direction of augmentation technologies – decisions which may have profound and long-lasting effects on the future.
- Major biosecurity risks will develop with BCI and other augmentation technologies, unless addressed.
- It is currently impossible to ensure that technology developers prioritize human agency and self-regulation throughout the design and development processes.

### Avenues for Addressing These Challenges

- Develop robust metrics to measure the impact of augmentation on human agency, ensuring it aligns with empowering individuals.
- Implement stringent safety and biosecurity protocols that prioritize human health and ethical considerations throughout the development and use of these technologies.
- Create robust data protection protocols, such as homomorphic encryption, to be used across augmentation technologies, to safeguard user data privacy.
- Develop concrete agendas for the current and future plans for BCI, as well as determining whether human augmentation may be required for AI Alignment.

# Phase II: FOCUS

## GOAL: ADDRESSING SYSTEMIC FLAWS THROUGH AI AND ECONOMIC INCENTIVES

Organizations like the OECD, World Bank, Food and Agriculture Organization, and institutions such as Social Impact Bond and Investment Companies are actively working to address negative externalities – the unintended costs of economic activity. However, systemic flaws such as a lack of political support, siloed knowledge, outdated models, insufficient data, limited budgets, and a conservative approach to innovation, hinder their progress. The upcoming integration of TAI and new economic structures focused on incentivizing the mitigation of negative externalities offer a potential solution to overcome systemic obstacles.

### Challenges of Current Institutions

- A lack of stakeholder confidence alongside insufficient political support impedes integrating negative externalities into economic and development models.
- The compartmentalization of knowledge and methodologies obstructs a holistic understanding of complex global challenges.
- Reliance on outdated economics models limits adaptation to new insights and innovative approaches like TAI.
- Data narrowness leads to narrow models, perpetuating the data gap and undermining efforts to accurately account for negative externalities.
- Limited budgets restrict data collection and model development, hindering the understanding and mitigation of negative externalities.
- Overly conservative stances towards model recommendations and budget allocations stifle innovation and new methodologies.

### Avenues for Addressing These Challenges

- Increase political support by demonstrating the social and economic benefits of considering negative externalities.
- Incentivize continuous “integrated data” collection processes for paid communities, which would increase data fidelity.
- Advocate for increased resources and explore innovative financing solutions, such as Social Impact Bonds.
- Use AI to enhance model accuracy, adaptability, and learning from new data.
- Establish a collaborative system to expand data collection and metric coverage.

# Phase II: FOCUS

## GOAL: ENVISIONING AND NAVIGATING TAI FUTURES

Current efforts to envision and prepare for diverse AI futures fall short. To illuminate potential pathways and challenges, we require a more comprehensive approach. We need something like a TAI "Horizon Scanner" that utilizes advanced world models and simulations to identify blind spots in current understandings and projections, and to map future landscapes that explore a broad spectrum of potential futures.

### Challenges of Current Institutions

- Existing models tend to focus on either immediate issues or distant, theoretical scenarios, neglecting comprehensive near-future planning.
- The secrecy surrounding military scenario planning prevents international and private sector collaboration, which is crucial for addressing the future global impact of TAI.
- Institutions such as RAND, while historically relevant, struggle to adapt to the dynamic and multifaceted landscape of AI.
- Scenario generation often lacks diverse perspectives and fails to iterate or expand upon initial models.

### Avenues for Addressing These Challenges

- Engage a global community of scientists, ethicists, policymakers, and artists to co-create and refine diverse scenarios.
- Develop dynamic models that evolve with new information and perspectives, using AI for feedback and refinement.
- Seek consensus on the internal logic of scenarios (what happens within them) while remaining open about their likelihood.
- Identify the underlying assumptions and biases in each scenario through analysis and discussion.
- Utilize LLMs for feedback, synthesis, and simulating potential AI agent responses within scenarios.
- Design simulations and role-playing games based on scenarios to explore their implications.
- Prepare clear reports with diverse future possibilities and actionable recommendations.

## GOAL: BRIDGING THE REALITY GAP THROUGH COLLABORATIVE KNOWLEDGE PLATFORMS

A growing diversity of perspectives in society, and the widening of gaps between them, presents challenges in establishing common ground on factual information and making informed decisions. Platforms are needed that engender trust and understanding around information sources. However, existing platforms, like Wikipedia, face limitations in relation to TAI.

### Challenges of Current Institutions

- Existing institutions often struggle to keep pace with the latest information.
- The platforms in place are not sufficiently large or adaptable to encompass the full breadth of human knowledge. Although popular topics receive thorough coverage, niche areas often lack detail or immediacy.
- Crucial information is not always readily available, which can prevent timely decision-making and learning.
- Discussions and notes, despite their value, frequently lack the structure and verification needed for widespread acceptance.
- Mechanisms to bridge different perspectives or to translate between knowledge systems are missing.

### Avenues for Addressing These Challenges

- Utilize AI and LLMs for “automated Wikipedia” systems to generate and update content dynamically, ensuring comprehensiveness and timeliness.
- Implement robust verification through community expertise and automated fact-checking to maintain accuracy and trust.
- Develop inclusive processes for information curation, incorporating diverse viewpoints and methodologies.
- Adopt flexible formats to represent the full spectrum of human knowledge.
- Integrate tools that highlight and reconcile differences in perspectives, fostering mutual understanding.

# Conclusion of the FOCUS phase

In wrapping up the FOCUS phase, the hackathon effectively distilled broad TAI aspirations into nine critical areas for attention, covering everything from enhancing the safety of AI systems through stringent evaluations to promoting human prosperity in the age of automation. The exploration into these areas revealed key challenges, including the imperative to bolster human agency, tackle systemic issues through innovative AI applications and economic strategies, and the need to forecast the multifaceted futures shaped by AI advancements.

The collective response to these challenges was marked by a commitment to developing open, standardized evaluation frameworks, developing democratic governance in AI oversight, enhancing international cooperation on military AI use, and crafting flexible, responsive regulatory environments. Notably, these strategies underscore the importance of placing human interests at the core of AI development, exploiting economic models to mitigate unintended societal costs, and creating dynamic, inclusive platforms for knowledge sharing.

This phase's achievements lay in identifying actionable priorities and setting a collaborative course towards addressing the effective governance of TAI.

PHASE III:

# ACT

After having completed the narrowing down of top goals for TAI, and how to address their main challenges, the hackathon moved into the "ACT" phase, focusing on finalizing drafts for potential institutions and operationalizing prototypes. Teams presented development plans and strategies to scale their projects into viable institutions.

The phase ended with presentations showcasing prototypes and receiving feedback from peers. Judges then deliberated and assessed each project's potential based on its relevance, impact, feasibility, and prototype quality. Thank you so much to our judges Anthony Aguirre (Future of Life Institute), Christine Peterson (Foresight Institute), Emilia Javorsky (Future of Life Institute), Hannu Rajaniemi (Helix Nanotechnologies), Tom Kalil (Schmidt Futures), and Kipply Chen (Anthropic).

The event concluded with the distribution of developmental grants to the winning teams. The proposal that received the highest scores from the judges was awarded $10,000, recognizing their outstanding contribution and the potential impact of their proposed institution. In acknowledgment of the high caliber of submissions, two teams were selected as runners-up, each receiving $5,000. These grants were intended to provide financial support for the further development of the institutions, facilitating the winners and runners-up in their efforts to bring their prototypes into reality.

# Institution Prototypes

## 1. The Flourishing Foundation (Hackathon winner)

### TEAM MEMBERS

Amanda Ngo
EX OUGHT

Chiara Gerosa
IMPACT ACADEMY

Jelena Luketina
2024 FORESIGHT FELLOW

Mingzhu He
CONSCIOUS TECH COLLECTIVE

### ABOUT THE INSTITUTION

**Mission:**

Enable TAI-integrated consumer technologies to promote sustained human and planetary wellbeing by developing new norms and processes, and supporting a community-driven ecosystem.

**Goals**

- Support a foundational research agenda on human flourishing that is focused on developing well-being metrics.
- Create a public and transparent certification for how well products support well-being.
- Support people integrating well-being objectives into AI development and deployment via addressing current limitations in community engagement and knowledge sharing.
- Support the education and cultural shift of the wider venture ecosystem that scale and grow ventures.

# Phase III: ACT

## Success Metrics

- Develop and publish a well-being index based on the established research agenda.
- Launch and successfully implement the Flourishing Certification, a dynamic rating system for organizations and AI models that reflects community preferences for well-being.
- Support the technology for human flourishing community via a fellowship program, the early open exploration of project ideas via an incubator program for value-aligned products, and eventually a venture studio.

## Next Steps

- Conduct foundational research on human flourishing through an iterative and community-driven approach. This will generate practical insights and case studies to guide the broader use of the index and certification.
- Develop the Flourishing Certification, collaborating with a broad range of specific sectors and communities to ensure it meets diverse needs, ensuring that it dynamically updates based on people's real ratings. This includes analyzing similar certification initiatives to identify best practices and potential challenges.
- Launch a new fellowship program, and incubate value-aligned projects and organizations.
- Plan and host a conference on AI for Wellbeing.

# 2. The Global Deliberation Coordinator (Hackathon shared second place)

## TEAM MEMBERS

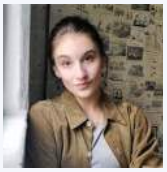
Aviv Ovadya
NEWDEMOCRACY

Bear Häon
SCHMIDT FUTURES

Evan Miyazono
ATLAS COMPUTING

Joël Christoph
EFFECTIVE THESIS

Joshua Tan
METAGOV



Morgan Livingston
TSINGHUA UNIVERSITY

Matteo Pistillo
LEGAL PRIORITIES PROJECT

## ABOUT THE INSTITUTION

**Mission:**

The Global Deliberation Coordinator (GDC) aims to be a pioneering institution for representative global deliberation on humanity's pressing challenges. It combines traditional deliberative democratic processes with AI-powered tools to address the need for rapid, cost-effective, and accessible global decision-making.

# Phase III: ACT

## Goals

- Pioneer Global Deliberation as a Service (GDaaS), demonstrating that it is possible for organizations and institutions worldwide to rapidly engage in global deliberative processes on major global challenges, such as AI development, climate change, etc.
- Implement a specific approach to GDaaS that is, and is seen as, trusted, fair, and accurate—and which leads to decisions which are implemented.

## Success Metrics

- Operationalizing a trusted and inclusive framework for global deliberations.
- Rapidly convening representative deliberative assemblies.
- Deliver respected and impactful decisions.
- Ensuring broad participation and representation in the process.
- Influence global governance and policy-making.
- Delivering accurate and rapid decisions, leading to tangible improvements in addressing the world's most pressing challenges.

## Next Steps

- Secure an Advanced Market Commitment by organizations (e.g., companies, and international organizations such as the UN) that would use a GDaaS, in order to spur rapid investment.
- Derisk key aspects of GDaaS through iterative pilots, and refine processes based on the results.
- Thoughtfully integrate AI and other digital tools to support scalable, secure, and effective deliberations on complex global challenges.

# 3. The Scenario Planning Institution (Hackathon shared second place)

## TEAM MEMBERS

Bryce Hidysmith
NEXAE SYSTEMS

Colleen McKenzie
AI OBJECTIVES INSTITUTE

Justin Bullock
CONVERGENCE

Keenan Pepper
INDEPENDENT RESEARCHER

Steve Caldwell
AE STUDIO

Li Zi
INDEPENDANT

## ABOUT THE INSTITUTION

**Mission:**

To enhance public awareness about the various futures TAI might create and its impact on society, focusing especially on outcomes that institutions such as corporations, large nations, and the military miss. By illuminating these overlooked futures and encouraging public discussion, the goal is to improve global comprehension and readiness for the intricate changes TAI may bring.

**Goals**

- Develop detailed and accessible analyses of TAI's potential impacts through a blend of formal models and engaging narratives.
- Ensure public transparency by freely sharing research insights and fostering global dialogue about TAI.

# Phase III: ACT

- Generate nuanced and comprehensive models exploring underrepresented TAI futures.
- Foster a well-informed public capable of navigating AI uncertainties with preparedness.
- Influence policy, corporate strategy, and public opinion by highlighting potential futures demanding proactive action.

### Success Metrics

- Modeling a wide breadth of currently overlooked scenarios.
- Ensuring that the created models obtain stakeholder engagement and utility.
- Ensuring these scenarios reach the public eye, contributing to the public discourse of TAI development and governance.

### Next Steps

- Build a core foundation by forming a core team – including an Executive Director and modeling experts – as well as recruiting project advisors and advisory board.
- Identify a variety of diverse, underrepresented potential TAI futures for exploration, evaluating them based on their neglectedness, potential global impact and relevance to the public.
- Execute basic modeling process, including consulting domain experts to enrich models with depth and actionable insights.
- Publicly launch findings in engaging, publicly interactive formats; amplifying reach and public discourse via the current ecosystem and the wider media.
- Ensure institutional legal and financial stability to continue the Scenario Planning Project.

# 4. The Evals for Evals Institute

## TEAM MEMBERS

Bogdana Rakova
MOZILLA FOUNDATION

Maximilian Negele
CFACTUAL

Siméon Campos
SAFER AI

Kyle Kilian
TRANSFORMATIVE FUTURES INSTITUTE

Nico Miailhe
THE FUTURE SOCIETY

Tilman Räuker
SWISS EXISTENTIAL RISK INITIATIVE (CHERI)

## ABOUT THE INSTITUTION

**Mission:**

Establish uniform criteria for evaluating AI systems before they are launched, focusing on safety, fairness, and transparency.

Develop a protocol to certify evals used for frontier AI pre-deployment, focusing on safety, fairness, and transparency. This protocol will align with standards set by the EU AI Office and similar organizations dedicated to AI safety.

**Goals**

- Develop a certification protocol for AI evaluation organizations that the EU AI Office can use.
- Structure the field of AI evaluations by harmonizing diverse assessment methods.

# Phase III: ACT

- Prevent “safety washing” while encouraging evaluation innovation.
- Establish a dynamic evaluation ecosystem for continuous improvement.

**Success Metrics**

- A robust evaluation ecosystem with a diverse set of risk assessment and classification methods.
- These methods should be well-defined and rigorous, yet allow for innovation to guide the development of the entire evaluation field (value chain).
- This approach helps mitigate the risk of “safety washing” from poorly designed or conducted evaluations, while still fostering a thriving and diverse evaluation landscape.
- Ensure AI systems are assessed against these high ethical and technical standards.

**Next Steps**

- Draft Request for Evaluation (RfE) protocol: Outline the requirements for evaluation organizations seeking certification.
- Solicit input from external stakeholders, including industry and academia.
- Consolidate and develop a final white paper
- Engage the European Commission and other relevant agencies to push for implementation.

# 5. The World Convention on Transformative Artificial Intelligence

## TEAM MEMBERS

Brandon Goldman
LIONHEART VENTURES

Fin Moorhouse
LONGVIEW PHILANTHROPY

José Villalobos
LEGAL PRIORITIES PROJECT

Ludwig Illies
NON-TRIVIAL

Saad Siddiqui
CENTER FOR THE GOVERNANCE OF AI

Stephen Clare
CENTER FOR THE GOVERNANCE OF AI

## ABOUT THE INSTITUTION

**Mission:**

Create a dual strategy to prepare for a World Convention on Transformative Artificial Intelligence (WCTAI). The first part is focused on engaging diplomatically with government officials and policymakers to gain worldwide backing for the conference. The second part involves conducting research to set the stage for the event, covering the establishment of TAI development milestones, crafting inclusive and impactful discussion formats for the conference, and examining different governance models to guide conference debates.

**Goals**

- Develop a global consensus on the need for TAI governance through diplomacy.
- Conduct research on key areas to prepare for and to discuss during the conference, including, a) TAI development milestones, b) Deliberative processes for the conference, and c) Comprehensive examination of governance options.

# Phase III: ACT

- Establish a comprehensive preparatory framework for the conference.

**Success Metrics**

- Global agreement to host a WCTAI.
- Development of a roadmap to the WCTAI, with stakeholder precommitments of engagement when certain technological trigger conditions recognizing TAI development are met.
- Engagement of international actors in the preparatory process.
- Involvement of key international actors in the WCTAI when the aforementioned trigger conditions are met.

**Next Steps**

- Refine the WCTAI's structure and draft funding applications.
- Circulate the WCTAI's proposal with governments and other stakeholders to work on the consensus on the need for a global conference on TAI.
- Conduct preparatory research, including research to define the milestones signaling TAI development, to ensure inclusive representation and deliberation processes, and to evaluate potential global governance options.
- Draft preparatory documents for the WCTAI, including potential agreements for state signatories.

# 6. The Common Knowledge Generator

## TEAM MEMBERS

Dusan D Nesic
PIBBSS

Jan Kulveit
FUTURE OF HUMANITY INSTITUTE

Lewis Hammond
COOPERATIVE AI

Tushant Jha
FUTURE OF HUMANITY INSTITUTE

Ziya Huang
CONCORDIA AI

## ABOUT THE INSTITUTION

**Mission:**

Create a universally accessible, community-verified source of information that enhances understanding and informs decision-making of complex issues by generating shared models of reality.

**Goals**

- Develop a platform (“epistemic commons”) with verified information presented in various formats and perspectives.
- Address limitations of existing platforms by offering real-time updates, comprehensive overviews of diverse viewpoints, and resources at different complexity levels.
- Enhance user access to trustworthy information across disciplines, cultures, and languages.

# Phase III: ACT

### Success Metrics

- A qualitative improvement in global epistemic alignment (shared understanding).
- A quantitative increase of access to multi-perspective, multi-language information resources.

### Next Steps

- Create a functional platform demonstrating the initiative's potential to various stakeholders, continually iterating this pilot.
- Engage the scientific community as initial users and partners of the platform.
- Improve the platform capabilities in information distillation, translation, and presentation using technology, such as LLMs, as well as community input.
- Expand the platform scope beyond scientific knowledge, to include diverse fields and languages.
- Secure funding for platform development and expansion via engagement of potential stakeholders, partners, and the broader community.

# 7. The Delphi Collaboration Protocol

## TEAM MEMBERS

Connor McCormick
GAIA INSTITUTE

Deger Turan
AI OBJECTIVES INSTITUTE

Elyse Lefebvre
INDEPENDENT

Konrad Seifert
SIMON INSTITUTE FOR LONGTERM GOVERNANCE

Malcolm Murray
CENTER FOR THE GOVERNANCE OF AI

Rishi Patel
N/A

Toby Pilditch
TRANSFORMATIVE FUTURES INSTITUTE

## ABOUT THE INSTITUTION

**Mission:**

To become the go-to source for comprehensive, always up-to-date world models where decision-makers can make data-informed decisions in a post-TAI world.

**Goals**

- Build an institution with a high adaptive capacity, including to increasing automation, which will facilitate the solving of urgent world issues by enabling diverse actors to come to their own, contextualized conclusions.
- To help global decision-makers focus on the most important global questions.
- Increase the efficiency of responses to international crises, from actors such as the WHO and FAO.

# Phase III: ACT

### Success Metrics

- Decision-makers gain access to world models, including informed policy recommendations, that have been tested and refined by a large varied and pool of contributors.
- Decision-makers gain estimates of political legitimacy and uncertainty for diverse scenarios, aiding risk-level calibration in decision-making.
- Encourage collaboration between citizens across the globe, increasing data quality and political legitimacy through mutual quality checks.
- Increasing the automation of key processes in international organizations, such as the WHO and FAO, which enhances efficiency during crises.

### Next Steps

- Refine the prototype into a fully functional platform for collaborative modeling and scenario analysis, integrating diverse data sources and modeling approaches.
- Validate user demand for an ecosystem for model creation via holding a pilot modeling project on pre-existing knowledge challenges with predefined answers.
- Expand the model to other test-cases which are malleable and have a high temporal sensitivity. Use this to court funding and gain partners with government bodies, AI labs, researchers, and expert forecasters
- Create governance mechanisms and guidelines for platform use and contribution, to ensure information credibility and reliability.
- Launch the platform to a wider audience, iterating on functionality and model quality.

# 8. The Open Source BCI project

## TEAM MEMBERS

Diogo de Lucena
AE STUDIO

Judd Rosenblatt
AE STUDIO

Mamun Miah
LAWRENCE BERKELEY NATIONAL LABORATORY

## ABOUT THE INSTITUTION

**Mission:**

Develop a privacy-preserving, open-source BCI operating system (BCI-OS) that enhances human cognitive abilities and safeguards human-agency in the TAI era. This would contribute to a future where TAI serves as a tool for human flourishing, addressing challenges such as mental health and AI alignment.

**Goals**

- Integrated agency evaluations, model compatibility protocols, and robust data privacy standards in the BCI-OS.
- Secure widespread adoption within the neuroscience and technology sectors.
- Enhance human cognitive abilities and user agency in the midst of AI advancements.
- Preserve and strengthen data privacy as AI technologies evolve.
- Contribute to the curing of depression, mood disorders, as well as neurological diseases and paralysis.
- Leverage BCIs to address AI alignment challenges and foster a positive AI future.
- Improves all humans' baseline happiness (and other desired states) by three orders of magnitude.

# Phase III: ACT

### Success Metrics

- Widespread adoption by neuroscience and technology sectors.
- Demonstrated improvements in human cognitive abilities and agency.
- Preservation and enhancement of data privacy.
- Curing depression, mood disorders, and neurological diseases.
- Progress towards addressing TAI alignment challenges.

### Next Steps

- Implement essential agency evaluations, privacy protocols, and compatibility standards.
- Collaborate with and engage BCI hardware manufacturers, software developers, neuroscientists, and AI researchers.
- Develop a governance structure that upholds human-agency and privacy principles.
- Open the platform to the open-source community for increased innovation and feedback.
- Conduct pilot projects with academic and industry partners to test, gather feedback, and refine the operating system.
- Creation of a set of standards and a standards body.

# 9. The SECHI Institute

## TEAM MEMBERS

Steve Coy
THE GAIA INSTITUTE

## ABOUT THE INSTITUTION

**Mission:**

Solve the "metacrisis" by developing a complete software system for "Simulation-Enabled Cooperative Human Intelligence" (SECHI) designed to oversee and improve the management of Earth and its interactions with external factors using a model predictive control (MPC) approach.

**Goals**

- Develop TimeLike, a first-in-class component-based software platform specifically designed to provide all the necessary software infrastructure required to support simulation-enabled cooperative human intelligence.
- Develop and deploy a global MPC loop encompassing the Earth and human activities.
- Integrate diverse models covering aspects like geology, space weather, and human systems on the TimeLike platform.
- Facilitate collaborative human intelligence and cutting-edge simulation technology to address global challenges.
- Enable humanity to navigate current and future crises for a flourishing future.

**Success Metrics**

- Successful implementation of the global MPC loop running on TimeLike.
- Improved ability to address global challenges and navigate crises.
- Successfully dealing with all of the major global crises that presently confront us, creating a flourishing future for humanity and the biosphere.

# Phase III: ACT

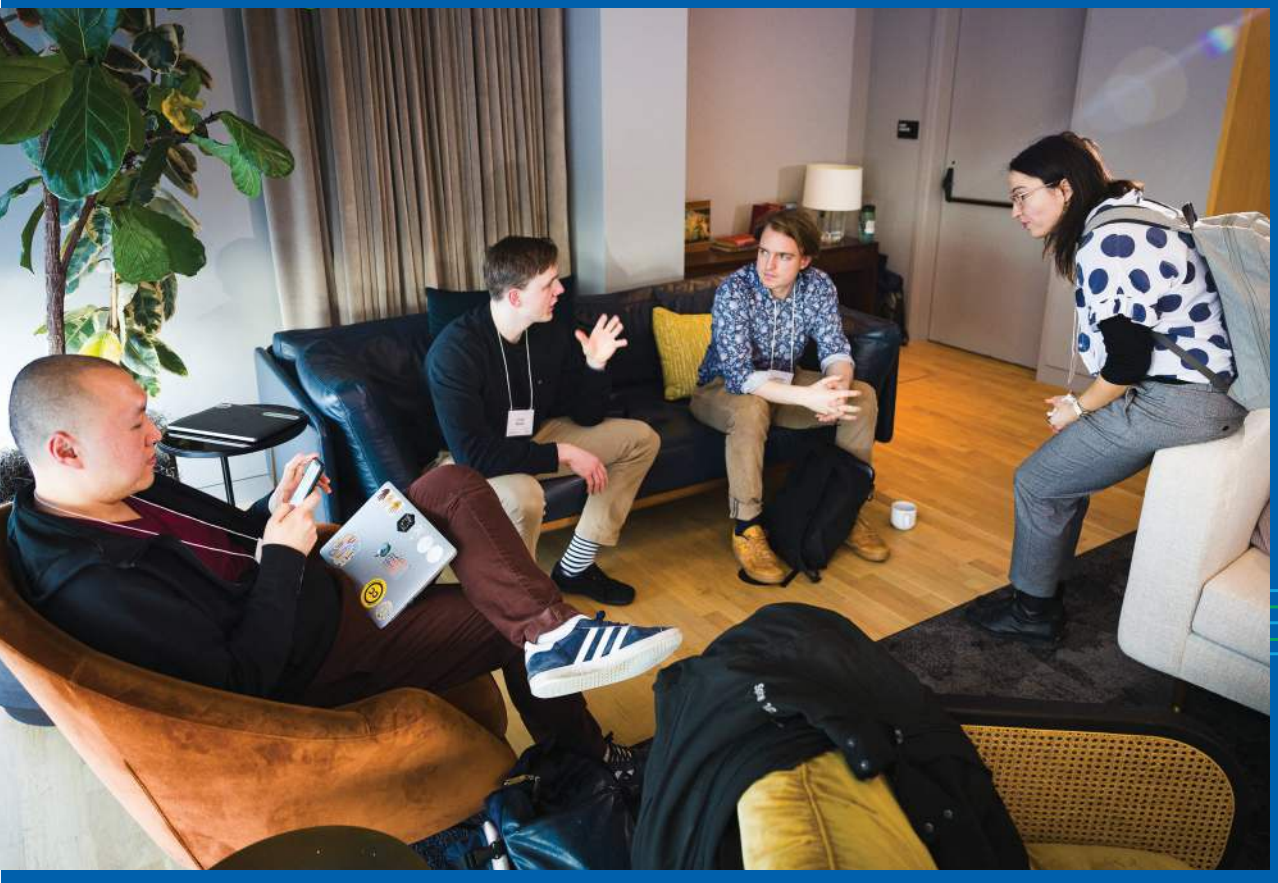
### Next Steps

- Secure key stakeholders and seed funding to support the core team in early development, before securing longer-term FRO-style funding.
- Create a real-world based, proof-of-concept demonstration and beta design.
- Promote widespread adoption of SECHI in global decision-making.
- Scale and implement the SECHI framework to address global challenges strategically.

# Conclusion of the ACT phase

In the "ACT" phase, the hackathon transitioned into finalizing drafts and operationalizing prototypes, with teams presenting their development plans and strategies to evolve their projects into viable institutions. This phase concluded with evaluations by judges, who assessed each project's potential based on its relevance, impact, feasibility, and prototype quality.

The culmination of this phase was the distribution of developmental grants, rewarding the most promising proposals with financial support to aid further development. These steps marked significant progress towards the hackathon's goal of fostering the development of institutions capable of guiding TAI towards safe and beneficial outcomes. This phase aimed to set a foundation for ongoing collaboration and innovation, even after the conclusion of the hackathon.



# Closing Remarks and Key Outcomes

The hackathon laid a strong foundation for future collaboration and innovation in TAI governance, especially highlighted by the creation of nine promising institutional concepts. We would like to extend our congratulations to the winning teams again for the high quality of their proposals.

Building on the hackathon's work, Foresight Institute is launching an Existential Hope Worldbuilding Course. This course focuses on imagining AI in various future scenarios, promoting optimistic visions of AI solving global challenges. Additionally, Future of Life Institute plans to continue supporting these efforts through its Futures program. This program aims to steer humanity toward the beneficial uses of transformative technologies, including offering new funding opportunities for research on safe AI applications to improve the world. These initiatives are crucial for refining and implementing the innovative ideas and organizations envisioned during the hackathon.

In closing, we extend our gratitude to all who contributed - from those who attended to the judges, mentors, and the Future of Life Institute team. While we recognize areas for improvement, we are grateful for the insights and enthusiasm brought by all participants, which enrich our shared mission to positively shape the future of TAI.

# Appendix

Data is available at: APPENDIX: Foresight Institute TAI Institutions Hackathon 2024. The document contains the complete list of TAI goals, and additional “wildcard” institutional proposals.

# Workshop Photos

The Future of Life team in attendance – Anthony Aguirre, Emilia Javorsky, Isabella Hampton, and Anna Yelizarova – discussing the prototypes.


The Hackathon winners, the Conscious Collective – Chiara Gerosa (Impact Academy), Amanda Ngo (Ex-Ought), Mingzhu He (Collective Intelligence Project), and Jelena Luketina (Future of Humanity Institute) – collecting their award.

Participants Dusan D Nesic, Ziya Huang, Vilhelm Skoglund, Stephen Clare, and Konrad Seifert in deep work mode.

Participants coming together to share ideas together during the beginning of the Hackathon.

The Global Deliberation Coordinator team – Joshua Tan (MetaGov), Aviv Ovadya (newDemocracy), Matteo Pistillo (Legal Priorities Project), Evan Miyazono (Atlas Computing), Joël Christoph (Effective Thesis), Bear Häon (Schmidt Futures), and Morgan Livingston (Tsinghua University) – celebrating after winning shared 2nd place.

Emilia Javorsky (Future of Life Institute) giving feedback on the prototypes, alongside fellow judges Anthony Agguire (Future of Life Institute) and Christine Peterson (Foresight Institute).

Steve Caldwell (AE Studio) and Diego de Lucena (AE Studio) brainstorming.

Different teams finalizing their prototypes before the judging.

Andrew Maynard (Arizona State University) giving additional insight to the teams at the end of Day one.

The Deep Green Scenario Planning team, including Bryce Hidysmith (Nexae Systems), Colleen McKenzie (AI Objectives Institute), Steve Caldwell (AE Studio), Justin Bullock (Convergence), and Keenan Pepper (Salesforce), discussing their prototype, which won shared second place.

Kipply Chen (Anthropic) delivering her reasoning for her rankings, alongside fellow judges Hannu Rajaneimi (Helix Nanotechnologies) and Tom Kalil (Schmidt Futures).

Ela Madej (50 Years) and Allison Duettmann (Foresight Institute) sharing ideas.