The Collective Intelligence Project

Table of Contents

Introducing the Collective Intelligence Project

Solving the Transformative Technology Trilemma through Governance R&D

The Collective Intelligence Project (CIP) is an incubator for new governance models for transformative technology. CIP will focus on the research and development of *collective intelligence capabilities:* decision-making technologies, processes, and institutions that expand a group's capacity to construct and cooperate towards shared goals. We will apply these capabilities to *transformative technology:* technological advances with a high likelihood of significantly altering society.

Collective intelligence (CI) is how we set and execute on collective priorities. Innovations in CI systems, like capitalist markets or nation-state democracy, have shaped the modern world. As collective problems have become more complex, our CI systems have too: global governance institutions and transnational corporations, standards-setting organizations and judicial courts, the decision structures of universities, startups, and nonprofits . These have allowed us to build incredible things. But they have also failed us. Rigid democratic institutions fail to serve their constituents or coordinate to solve global crises. Market mechanisms flatten complex values in favor of over-optimizing for cost, profit, or share price. **Our most pressing challenges are fundamentally collective intelligence challenges: pandemics, climate change, plutocracy, and catastrophic risks from technology all require better ways to set and execute on priorities.**

These failures are most evident when we apply existing CI systems to accelerating technological capacities. We have made little progress on regulating decades-old social media platforms, and we can barely talk about the dramatic resourcing shifts necessary to address growing climate risks. But new risks and opportunities continue to arise: we are faced with powerful AI models, blockchain-based financial and social technologies, expanded bioengineering capabilities, and large-scale labor automation. Directing technological development towards good outcomes requires working on the processes and institutions that drive effective decision-making around transformative technology. CIP is a response to the inevitable need for innovation brought about by the problems that existing CI systems could not solve.

At CIP, our core belief is this: Humans created our current CI systems to help achieve collective goals. We can remake them.

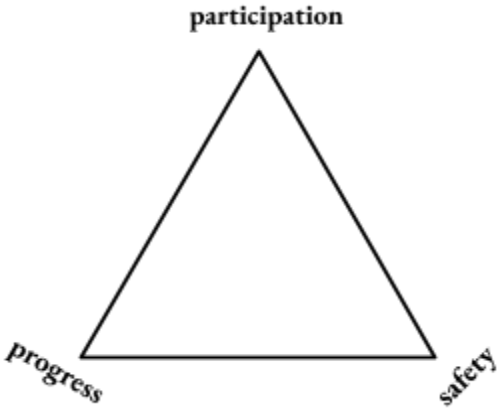
Here is what that means in practice. **First, we need new models of value elicitation: ways to develop scalable processes for surfacing and combining group beliefs, goals, values, and preferences.** Discussions of democratizing technological development abound, but they often leave aside the core question of actual collective input. Nation-state models of voting and representation are crude approximations of collective values and are ill-suited as inputs to technological development. We will accelerate promising alternatives. Currently, we are testing the use of liquid democracy for creating more values-aligned recommender systems; augmenting the emerging discussion platform Narwhal with language model capabilities, and supporting consortia-based efforts from complementary projects such as Pol.is, the AI Objectives Institute, the Consilience Project, RadicalXChange, and New Public.

Second, we must remake technology institutions. It is not enough merely to *understand* collective values. We must be able to *execute* on collective values. This requires developing hybrid philanthropic, public, and private funding models for technology development beyond the existing options of non-profit, VC-funded startup, or academic project. Our first pilot will be the 'CI Corporation': a scalable, capped-returns model for technology development and deployment. This builds from our past work on developing data intermediary institutions, frameworks for decentralized governance and metagovernance, internet standards-setting, and pandemic prevention consortia.

In the grand sweep of human history, it is highly unlikely that we've already somehow landed on the *best ways to make collective decisions for the collective good.* Transformative technologies give rise to new problems, and our collective intelligence must evolve to solve them. Our aim is to accelerate this necessary evolution by catalyzing an ecosystem of aligned governance research and development projects.

The Transformative Technology Trilemma

Transformative technologies (TTs) refer to technological advances with a high likelihood of significantly altering society, such as birth control, air travel, or the Internet. TTs will affect vast numbers of people and change long-term civilizational trajectories. The outcome of TT development, at least in the initial stages, are liable to be determined by a small proportion of those people and the institutions that house them, based on a fairly narrow set of priors and assumptions.



Resulting governance models have fallen prey to the **transformative technology trilemma.** Coalescing camps implicitly or explicitly assume the need to accept significant trade-offs between **progress** (advancing technological capabilities), **participation** (enabling public input and self-determination), and **safety** (avoiding disproportionate risks). This reliably leads to a set of three failure modes.

I. **Capitalist Acceleration**—Sacrificing safety for progress while maintaining basic participation.

This path aims to incentivize and ensure technological progress, generally through a belief in free-market, profit-driven development. Participation comes in the form of consumer choice and investor agency, and risks

are taken by those who have the resources to take them. Perhaps this means a proliferation of VC-funded AGI and biotech startups optimizing for growth, or private geoengineering to offset climate risk, or the takeoff of unregulated decentralized finance. The upsides of this path include decentralized decision-making and necessary investments in tech advances (e.g. CRISPR). The downsides include proliferating risk (e.g. if startups use CRISPR to edit pathogens), and lack of public oversight (minimal regulation, auditing, or provision of public goods). Downsides may be especially significant when it comes to transformative AI—applying the structure of one value-flattening optimizer (profit-maximizing markets) to directing another (reward-maximizing artificial intelligence) could lead to catastrophic outcomes.

II. **Authoritarian Technocracy**—Sacrificing participation for safety while maintaining basic progress:

This path is built on the belief that ensuring safety requires entrusting only a few entities (individuals, companies, nation-states) with the ability to develop advanced technologies. This is coupled with the assumption that collective participation is too dangerous, too difficult to coordinate, too slow, or likely to lead to lower-quality decisions. Take the 'Vulnerable World Hypothesis', which advocates for total global surveillance in the face of catastrophic risks posed by emerging technologies. Or the CCP's response to Covid, which was defined by its strict monitoring, regulation, and calculation. The argument is simple: as the world becomes more dangerous, control structures must become more severe. Technological advances are harnessed for mass monitoring capabilities to ensure this control, eroding rights from privacy to free speech to due process. The upsides of this path include an understanding and avoidance of risk (e.g. it is possibly easier to coordinate a pandemic response). The downsides include the risks of illegitimacy (e.g. protests against China's zero-Covid policy and its dramatic reversal), the well-documented failures of central planning (e.g. the economic calculation problem and the challenges of gathering representative information for centralized decision-making), and the basic injustice of autocracy.

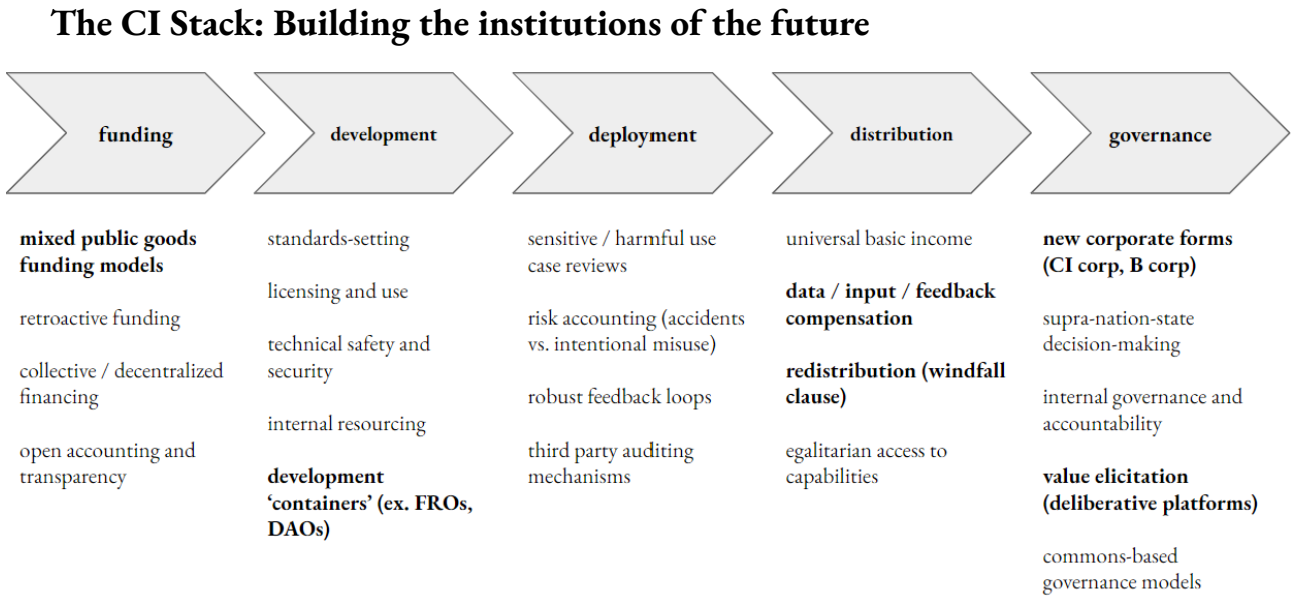
III. **Shared Stagnation**—Sacrificing progress for participation while maintaining basic safety:

This path combines anti-technology inclinations with concerns about worsening global conditions (such as climate change, inequality, bias and discrimination) due to current trajectories of progress. This is often paired with a desire for greater forms of direct democracy and local production and decision-making, and the explicit or implicit goal of forestalling technological advances. Tools for advancing this path can range from regulation and antitrust (recent EU policy points in this direction), to direct calls to halt technological investments and prioritize degrowth. The upsides of this path include an emphasis on shared power and decision-making (e.g. measures to protect privacy, distribute wealth, and regulate harm). The downsides include a lack of investment in necessary economic or technological development (e.g. nuclear power, public health, pharmaceutical advances), and undervaluing the need for large-scale coordination, e.g. via international bodies or large-scale production.

The Solution: Collective Intelligence R&D

Our goal is to find a fourth path, by developing a plurality of CI systems that encompass all three goals: participation, safety, *and* progress. When trade-offs must be made, they should be made in light of material outcomes and state-of-the-art information and preference gathering, not preconceived assumptions.

This requires CI R&D: the development of processes that reliably surface, choose, and execute actions that direct TT towards the collective good, including dynamic collective input on what 'good' means. We see possible vectors of CI R&D across the TT lifecycle, loosely represented below as the CI Stack, with areas of internal focus in **bold**.



Our initial R&D push will span two categories: value elicitation and remaking technology institutions.

I. **Value elicitation**: *Given a set of possible directions for technology deployment, how might we aggregate, understand, and incorporate the conflicting values of overlapping groups of people?*

We have few robust systems of collective information gathering and processing. Digital technology and innovations in governance design present affordances for increasingly granular preference elicitation, scaled deliberation, and collective alignment. These include **quadratic voting**, liquid democracy, and **deliberation tools** like Pol.is. In Taiwan, instead of optimizing for engagement (which often leads to polarization), civic technologists used Pol.is to surface useful, actionable statements that gathered broad agreement on various political questions, using them as a foundation for complex legislation. In addition, there are techniques for reshuffling resource allocation to prioritize goods that benefit more people, such as quadratic funding (QF) and

retroactive funding. Platforms like Gitcoin have implemented QF as a more democratic approach to funding public goods.

Innovative **polling methods** or **prediction markets** can help surface more accurate group forecasts to anticipate the future and understand people's preferences at scale. Sortition-based citizens' assemblies and other approaches to deliberative democracy can be conducted using digital platforms for better scalability, and potentially to decide on those platforms' policies. **Machine learning and natural language processing can be useful for managing this kind of large-scale deliberation**, to surface comments that bridge different perspectives, highlight areas of consensus between different people's statements, help to cite evidence, or summarize arguments. To systematically incorporate collective values, organizations could **democratically elect boards**, **replace a board member with a collective intelligence mechanism**, or create **platform assemblies** to make key decisions.

Our approach:
We will work on modularizing, combining, and experimenting with these systems across contexts. Our initial projects include:

- Running a citizen's assembly to understand how to gather and incorporate people's preferences, in order to align AI systems with a group's values — starting with the question of deployment decisions related to language models.
- Developing a set of strategies for LLM use in deliberative democracy and testing hypotheses in our partnership with the new discussion platform Narwhal (co-founded by *The Atlantic* and the Emerson Collective).
- Supporting consortia-based efforts from aligned projects such as pol.is, Talk to the City, the Consilience Project, RadicalXChange, and New Public.
- Experiments in taking a liquid democracy-based approach to designing recommender systems, with an initial focus on content moderation.

II. Remaking technology institutions: *Given competing incentives between progress, safety, and participation, how might we build a collectively-intelligent institution for developing and deploying transformative technology?*

Transformative technology is capital-intensive and high-risk. Its trajectory is thus informed by the interests of a limited number of private entities and well-resourced governments. How these organizations are financed and governed might greatly change how they develop and deploy technology, as well as how the benefits of the technology are distributed. **Questions around how to align incentives and distribute the returns to transformative technology are becoming critical.**

As it stands, the 'default container' for TT development remains the venture-capital funded startup. This model works well for asset-light, high-growth, low-accountability entities, but is dangerous when applied to societally consequential infrastructure projects, or as a default distribution model for exponential returns from AI advances. **We need a better container within which to build future technologies, from satellites to space travel to AI research.** The space for collectively-focused alternatives is growing: existing VC approaches have faced high-profile failures, tech-focused industrial policy and public funding are seeing a revival, and there is growing interest in standards-setting and auditing organizations. Couple this with initial forays into windfall redistribution and experimental approaches to philanthropy, and we can start charting a path that intentionally and effectively incorporates the public good.

Our approach:
We will extend and expand on existing experiments, from open source projects to benefit corporations to focused research organizations (FROs) to perpetual purpose trusts to cooperatives to decentralized autonomous organizations (DAOs). Our goal is to build a menu of alternate organizational structure, governance, and financing models available to TT founders and funders, and to enable multiple implementations of these new structures in the next 18 months. Current work includes:

- Experimental designs for a 'CI corporation' and the funding institutions that are necessary to sustain it, in partnership with the AI Objectives Institute. We are testing startup models that operate on a capped returns framework, including necessary legal and licensing innovations.
- Running an academic workshop in collaboration with the Cooperative AI Foundation on how AI could be used to improve and/or create designs for human institutions.
- Researching structures for governing generative models that account for the commons-based nature of the problems, in partnership with the office of Congresswoman Sara Jacobs.
- Expanding on previous work in positive-sum goods funding through actionable proposals for submodular (vouchers, auctions, bounties, tokens, etc.) and supermodular (digital commons, public matching funds, stakeholder oversight, auditing) processes for democratic financing.

Towards a Collectively-Intelligent Future

The ultimate goal of our work is concrete changes in real-world processes. Many promising mechanisms have already been proposed in theory, but these lack the empirical data to determine how well they work in different settings. To this end, we are working on piloting deliberative democratic tools with LLMs, building the infrastructure for data intermediaries, proposing new institutional forms for commons-based generative AI, and developing the CI corporation. Our areas of R&D focus will expand as CIP expands.

Nonetheless, there is far more work to be done across each layer of the CI stack than we can possibly imagine or implement on our own. Luckily, we are delighted to be joining a growing community of practice in this space, across technologists, policymakers, academics, scientists, activists, and citizens. Alongside our R&D projects, we

are working on developing a **Collective Intelligence Almanac**: a living map of the organizations, technologies, pilots, case studies, experiments, and platforms that make up this expanding ecosystem. This will also function as a guide for people to understand how to incorporate CI into their organizations.

As we embark on this journey, we welcome fellow travelers. A few ways to get started:

- **Join our community of practice:** Have an idea, a potential collaboration, or want to meet other folks in the space? Reach out to us here and we'll find ways to get you involved.
- **Contribute to the Almanac:** We are looking for people to work with on research, mapping, and sensemaking. Help us create easy-to-understand, implementable and well-researched modules for each stage of the CI stack.
- **Tell us about your CI project:** We are building a case library of CI work and experiments; we would love to add your project, and connect you to collaborators and supporters.
- **Collaborate with our allies:** CIP is proud to be part of a growing, multifaceted ecosystem building collective intelligence. The only thing better than working with us is working with our community.
- **Work on open questions (forthcoming):** We are compiling a list of wide-ranging open questions in CI, many of which we hope are tractable in the short- to medium-term. Work on one with your organization (and tell us about it).
- **Apply for a microgrant (forthcoming):** We are developing a microgrants program for scoped contributions to the CI stack. If you're interested in applying, reach out to us. If you have a project idea, reach out to us. If you're interested in funding microgrants, doubly reach out to us.

Building collective intelligence is both a human-scale and humanity-scale project. It will take ambitious experimentation through collective effort from diverse corners. But there is a path forward.

Edward O. Wilson once described the problem of humanity as having 'Paleolithic emotions, medieval institutions, and god-like technology'. This is not a sustainable trajectory. The time is ripe for new, collectively-intelligent institutions. We hope you will join us in building them.