AI Objectives Institute Whitepaper*

A Research Agenda for the Production of a Flourishing Civilization

February 2023

Abstract

We argue that, rather than a sharp break, there is continuity between problems in AI alignment and those in economic regulation, institutional design, and personal autonomy. There are therefore under-appreciated opportunities for ideas and interventions to flow between research communities in both directions. On one hand, the AI safety community may be able learn from successes and failures in aligning other kinds of superhuman optimizing entities, such as corporations and political parties, which are susceptible to biases analogous to those in AI systems. On the other hand, advances in AI (such as large language models) and breakthroughs in AI safety (such as advances in cooperative inverse reinforcement learning and reinforcement learning from human feedback, improvements in measuring Goodhart's law, and techniques for developing systems without explicit objective functions) could provide new tools for improving existing regulations, institutions, and self governance, or helping to design new regulations and institutions altogether. Bringing together researchers from leading AI industry and academic settings, the AI Objectives Institute (AOI) has been founded to explore these ideas further and develop practical interventions on these themes.

1 Introduction

We are at the beginning of a profound global transformation: the era of artificial intelligence. After sixty years of blind alleys and incremental progress, machine-learning research has found a path to a rapid increase in computer intelligence. And while major challenges may still lie ahead, it is now imperative to think seriously about how humanity should use these advancing capabilities. While it is hard to predict the structure of future AI systems, we can think of them as having "objectives" (how they measure progress towards goals) and "behavior" (means for how they get there). Ensuring that an AI's objectives and behavior are actually serving humanity is an existentially urgent task. Bringing together researchers from the leading AI industry and academic research labs, the AI Objectives Institute (AOI) has been founded to lay the theoretical groundwork and to develop a portfolio of practical projects to improve the odds that people can thrive in a world of rapidly deployed, extremely capable AI systems

*Thanks to Peter Eckersley, Max Shron, TJ, Divya Siddarth, Brittney Gallagher, Carroll Wainwright, Joel Lehman, Brian Christian, Deger Turan, and numerous others for their contributions to this whitepaper. Send correspondence to Max Shron, max@objective.is.

developed within the context of institutions (old and new) and economic and regulatory incentives.[1]

AI futurists of the past few decades have painted two dramatically different futures: radical abundance perhaps leading to a stable galaxy-wide civilization, or the complete disempowerment of humanity as we are quickly and decisively outsmarted (or even simply killed).[2,3] As we'll argue in this whitepaper, a very plausible but under-researched effect of misalignment[4] is a continuation of existing trends of concentration of power in fewer hands—super-charged by advancing AI[5]—rather than a sharp break with the present.

This whitepaper explores parallels between the problems we face with AI alignment in the long-run, and those that we face in the present day with institutional design and economic governance. Indeed, the term "alignment" originally came from the economics literature, before it was used in AI safety. Historically it described the challenges that owners of corporations face in creating the right incentives for the corporation's managers, who have practical power over the organization (usually discussed as "incentive alignment," which is a component of agency theory).[6] These same tensions exist between corporations and the public, and there are similar parallels between how democracies effectively (and often, imperfectly) delegate power to elected officials. All of these systems interact with each other today in complex ways—how much more complex will it get as AI improves?

More than likely, we also expect the goals given to AI systems will come from existing human institutions (like corporations, research labs, or governments), who already have their own goals and complex relationships to each other. Even before AI is added to the mix, those institutional goals may already be in tension with *human flourishing*.[7] Making matters worse, even when institutions carefully choose goals to pursue, in the hopes of creating the kind of world that they want to see, the unwieldy nature of reality will usually force them to rely on imperfect proxy metrics. Any gap between those proxies and their underlying goals will often be exacerbated by strong optimizers due to the effects of Goodhart's law.[8]

[1] More details about AOI, including its structure, current research areas, and active projects, can be found at the AOI website.

[2] *Disempowerment* in the AI safety literature often means "humans have permanently lost control of our future to machines" (e.g. Preventing an AI-related catastrophe (Hilton 2022)); we argue that disempowerment is actually a question of degree (how many people are in control of how much of their own destiny, and how much has been ceded to non-human systems which of which AI may only be a component) versus a binary question of empowered-vs-not-empowered against machines.

[3] See *Superintelligence* (Bostrom 2014) or Arms Control and Intelligence Explosions (Shulman 2014) for more on what a fast takeoff of AI and quick disempowerment of humanity might look like. Cf What Failure Looks Like (Christiano 2019) and Clarifying "What failure looks like" (Clarke 2020) for an account more in line with our perspective.

[4] Following *The Alignment Problem* (Christian 2020), we define "alignment" as techniques to ensure that AI systems "capture our norms and values, understand what we mean or intend, and above all, do what we want."

[5] In this whitepaper we avoid the term "artificial general intelligence" (AGI) to sidestep questions of whether or not general intelligence is a necessary condition for the kinds of problems we're concerned with. It seems reasonable that collections of more narrow systems could be equally capable of being a problem—see Forecasting Transformative AI (Karnofksy 2021)—though we don't discount the possibility of general intelligence either.

[6] For review of the pre-AI safety literature on this topic, see Agency Theory (Shapiro 2005).

[7] The conditions for human flourishing are those which allow all people to live their chosen version of a good life. For more on these ideas, see the idea of *eudaimonia* elaborated in the *Nicomachean Ethics* (Aristotle, trans. J.A.K. Thomson, Penguin 1976), or the capabilities approach articulated in *Development as Freedom* (Sen 1999).

[8] Goodhart's law is paraphrased as "When a measure becomes a target, it ceases to be a good measure". (Goodhart's law Wikipedia 2023). In other words, as soon as there are quantifiable objectives with incentives to game them, they will be gamed. Variations on this idea are also known as Campell's Law (in the context of research methodology) and

In some sense, we're already living in a world of *misaligned optimizers* (such as corporations and political parties), even if the magnitude of those effects may be smaller in the present than may get in the future. Somehow, despite the strong optimization pressures inherent in a market-focused society, civil society and government regulation has managed to keep some of the worst excesses of these optimizers in check. What can we learn from these partial successes? Our perspective opens up both new applications of alignment research today, and the possibility that existing strategies for constraining misaligned optimizers of various kinds may have implications for AI safety research.

These concerns are distinct but closely related to those articulated by the broader AI ethics and AI fairness communities, which often frame alignment issues through the lens of harms related to privacy, surveillance, bias, and discrimination.[9] We too believe in the importance of understanding institutions in detail; how existing social and institutional incentives will continue exacerbate harms when combined with AI; how extra-institutional dynamics (such as those related to race and gender) complicate the picture; and the importance of understanding how these topics intersect with legal oversight. These are real and urgent problems, and we seek to stay aware of how those concerns intersect the alignment issues we are focused on.

The AI Objectives Institute (AOI) is a new, collaborative research organization founded to explore the following questions: Given that we are already living in a world with misaligned optimizers, what useful techniques or strategies can we learn from the present to help us with problems we may face in the future? Alternatively, can present-day AI techniques lead to better institutions today? Can we build practical demonstrations of techniques for extrapolating latent goals for human flourishing, to better align both present-day and future optimizers?

This whitepaper sets out some initial framing by the AI Objectives Institute's Working Group on AI and Transformations of Capitalism. It paints a relatively quick picture of a complicated topic; we intend it to be followed by longer and more thorough documents for specific topics and projects as they come to fruition.

Because we seek to tackle a complex set of problems, we have broken down our work into three major themes, relating to different "levels" of organization of society. These levels form a social "stack" (the analogy is to how the internet or software is organized, with fewer abstractions as you go down).

At the highest and most abstract level, we have inter-institutional concerns, often involving markets, governments, and political parties competing with each other. The current AOI focus here is on *Alignment of Markets, AI, and Other Optimizers*. At the middle level, where institutions relate to their constituent members, our focus is *Scaling Cooperation with AI Assistance*. And at the base level, individuals, we are focused on *Human Attention and Individual Epistemic Security*, working to on help people better understand their own values and act in accordance with them (see Figure 1).

AOI members are pursuing this research theoretically and practically. Our approach is to bring a wide variety of researchers together (machine learning, economics, philosophy, social science) and when needed pair them up with technical domain experts (machine learning engineers, software developers, designers) to ensure that our practical work is high-impact. We hope this wide variety of backgrounds will cross-pollinate ideas, improve our odds of

the Lucas critique (in the context of macroeconomic policy).

[9] For prominent examples, see *The Alignment Problem*, *Algorithms of Oppression* (Noble 2018), or *Automating Inequality* (Eubanks 2018).

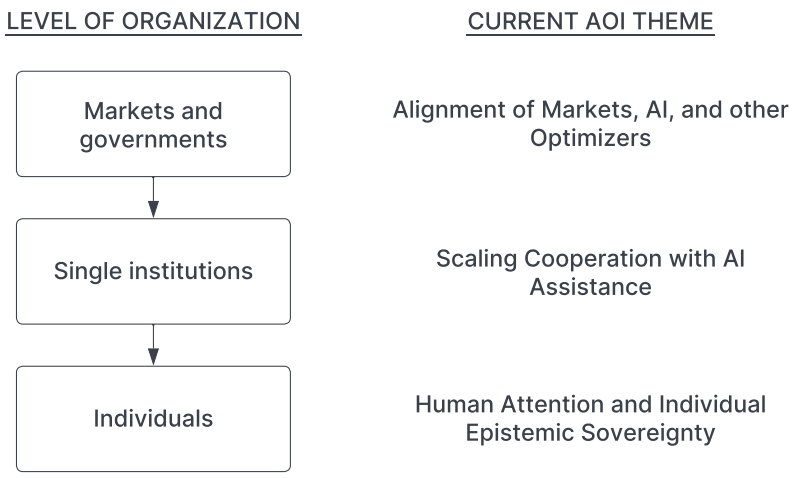

Figure 1: A simplified social "stack" (by analogy to how networking protocols are designed), current AOI theme by level of organization, and also a unifying structure for this whitepaper.

proactively identifying unintended consequences, and reduce the risk of exploring only local maximums in the space of possible interventions.

2 AOI's Research Areas

2.1 Alignment of Markets, AI, and Other Optimizers

> *I used to find it odd that these hypothetical AIs were supposed to be smart enough to solve problems that no human could, yet they were incapable of doing something most every adult has done: taking a step back and asking whether their current course of action is really a good idea. Then I realized that we are already surrounded by machines that demonstrate a complete lack of insight, we just call them corporations. Corporations don't operate autonomously, of course, and the humans in charge of them are presumably capable of insight, but capitalism doesn't reward them for using it. On the contrary, capitalism actively erodes this capacity in people by demanding that they replace their own judgment of what "good" means with "whatever the market decides."*
>
> Ted Chiang, "Silicon Valley Is Turning Into Its Own Worst Fear"

Starting with the highest level of our societal stack, how institutions relate to each other, AOI has begun a number of projects which touch on one of the most powerful class of optimizer institutions in the world today: markets in liberal economies.

In this section, we sketch out why markets are a useful tool for understanding what has

gone wrong and what has gone right in institutional objective alignment historically, where things might be going, and what we might be able to do about it.

When we say "markets," we mean collections of firms (usually corporations) optimizing for profit by competing for consumer purchasing power (or intermediate goods on the way to consumer purchasing), under the constraints of the legal systems they find themselves in. These constraints in turn are set by governments, who usually, in a slower and less well-targeted way, are trying to do their own optimization of legible[10] proxies like GDP, lifespan, and inflation, which in turn are intended to support harder-to-define goals like a flourishing population or military security. Even less easily measurable qualities, like human rights or privacy, are often sacrificed when they fall outside of more legible goals.

Fleshing this out more: people want to live a good life, and inexpensive goods help them to live a good life with less time spent providing for basic necessities. Our political economy provides for cheap food, clothes, entertainment, and communication (relative to historical Western norms) by empowering corporations to provide those goods in markets. But since those goods are provided by profit-seeking corporations, which optimize for profit and not directly for improving human lives, when they can increase profits indifferently or even at the expense of higher quality of life, they will (and often do). None of these ideas are new in the context of economics or political economy, but we believe that their connections with AI, both in how AI is likely to make these issues worse and how it could possibly make them better, haven't been thoroughly enough explored.

The AI alignment community is, correctly, concerned with humans being disempowered by misaligned artificial intelligences. But being outflanked by agents with more resources and intelligence than individual humans already happens all the time today; we don't need to wait for transformative AI to see it firsthand. In some sense, corporations are a kind of superintelligence; no single human could design or build an airplane, but Boeing does it all the time.[11] Corporations are also functionally sociopathic.[12] The literature on regulatory capture[13] provides good examples of corporations seeking influence to better optimize for non-aligned goals, by outflanking and disempowering more human-values-based democratic institutions, all as subgoals for attaining higher profits. Non-aligned superintelligences are already here and, e.g., worsening climate change, not because they prefer CO_2 or higher temperatures, but because it's not the problem they are legally or socially designed to care about. Now consider what could happen to the balance between government, civil society, and lobbying when those corporations begin experimenting with transformative AI capable of, for example, superhuman persuasion, or the ability to covertly exploit legal loopholes to a degree which is impenetrable to even well-resourced human regulation and oversight. Markets are both

[10] See *Seeing Like a State* (Scott 1999) for the definitive treatment of legibility in political systems.

[11] For more on this idea, see An existing, ecologically-successful genus of collectively intelligent artificial creatures (Kupiers 2012)

[12] We mean this in the psychiatric sense (e.g. lack of concern for feelings, needs, or suffering of others; lack of remorse after hurting or mistreating another; manipulativeness; deceitfulenss; callousness; and so on. (*Diagnostic and statistical manual of mental disorders (5th ed.)* (American Psychiatric Association, 2013). It's a common misconception that corporations are legally required to maximize shareholder returns within the bounds of the law. Though this was an idea championed by Milton Friedman, and is often repeated, in the United States at least it has not been held up by the courts. Nevertheless, in any sufficiently competitive market, any values other than profit maximization will tend to be squeezed away; see Meditations on Moloch (Alexander 2014) for an eloquent discussion on the topic.

[13] See *Preventing Regulatory Capture* (ed. Carpenter and Moss, 2013) for more on correctly diagnosing regulatory capture and what strategies have worked to prevent it in the past.

a critical test case for our ability to measure objectives and produce better alignment, and a proving ground for AI alignment research.

Many of the major advances in artificial intelligence today are being developed *within* market economies, typically serving corporate subgoals like persuasion to purchase, improving search quality, or optimizing ad placement. Instead of a tug-of-war between lofty objectives and messy optimization often considered in alignment research,[14] one kind of non-aligned agent (corporations) may just make better use of another (transformative AI), and the general populace will suffer as a result even as profits climb. Out-of-control markets are already disempowering us, but transformative AI could exacerbate the situation.

To be clear, AOI is not only concerned about misaligned optimization within markets. Political parties, for example, are also excellent optimizers, constantly tweaking their messaging and outreach to get at least 50% + 1 votes in any given election. They have figured out how to deploy media to change the salience of certain topics ("wedge issues") to increase or suppress turnout and therefore secure votes and gain power at the expense of citizen preferences.[15] Every party is incentivized to pursue these tactics, since if they do not then their political opponents will disempower them. We already see the use of sophisticated optimizers to maximize "engagement" and boost fundraising, or precisely targeted social media campaigns whipping up fear and resentment among potential voters. How might this get worse as narrow AI gets better at persuading people to be afraid (and open their wallets to donate to political campaigns because of that fear)?

Thankfully, not all is lost. Humanity already has some tools in its belt to realign the goals of powerful actors closer to what we actually want, even when they're optimizing for something very profitable: laws, institutional norms, governance, civil society, and tax policy. We've successfully ended the trans-Atlantic slave trade, cleaned up rivers, reduced smoking, drastically cut down on commercial whaling, and fixed the hole in the ozone layer, even though there were powerful systems in place that profited from each of them. If anything, the fact that what we face is not an entirely new problem is actually a reason for hope. Clearly it's possible to constrain bad behavior engendered by profit-seeking in *some* circumstances.

What analogies might exist to help us better align AI systems? Is there an AI equivalent of, say, using tax policy to align profit with externalities? Objective penalties (like regularization)[16] are structurally similar to taxation, providing a "tax" on certain behaviors by directly making them more "expensive". Can we learn from effective tax regimes how to design better objective penalties? Is there an equivalent to using antitrust law to reduce market dominance? Perhaps some variation on ensemble methods, which enforce a kind of competition among models, or directly using discriminator models to identify bad behavior. Reinforcement Learning from AI-feedback (RLAIF)[17], where AI models are designed to make judgements about the behavior of other models, is a clear step in this direction. How about analogies between civil oversight and explainable AI?

And at another level, are there ways to use actual tax policy or antitrust regulation to

[14] For a classic take on this problem, where the designer of the AI is explicitly trying to create something with prosocial values (as opposed to being content with a system that is exploitative), see Artificial Intelligence as a Positive and Negative Factor in Global Risk (Yudkowsky 2008).

[15] For more on this, see Toward a Theory of Pernicious Polarization and How It Harms Democracies: Comparative Evidence and Possible Remedies (McCoy, 2018). One interesting organization trying to change newsmaking incentives through persuasion is The Citizens Agenda.

[16] For more see Regularization (Wikipedia)

[17] See Constitutional AI: Harmlessness from AI Feedback (Bai et al. 2022)

better shape the development of AI? Casting an even wider net, are there other social science or humanistic perspectives even further afield that need to be brought into consideration?[18] These kinds of questions form one strand of research that AOI is pursuing.

Conversely, and perhaps more immediately actionable, as we discover new techniques in AI safety they may have practical applications to institution design or regulatory oversight. Single metrics can be easily gamed by, for example, drug makers looking to get approval for risky drugs. Can research on generating better ensembles of metrics provide tools to reduce the possibility of gaming the system?[19] By more carefully measuring and quantifying Goodhart's law, maybe the AI safety community can provide tools to help regulators better design feedback mechanisms.[20] Could techniques for reverse-engineering the behavior of complex deep learning models yield ideas for better "interpreting" market behavior?[21] And if we learn how to create agents without strong objective functions, and what it takes to design them to nevertheless behave well, might that open ways to design institutions without their own "objective functions"?[22] Over the last few decades, management by measurement provided great gains in efficiency at the expense of harder-to-quantify values, often resulting in brittle supply chains and a loss of trust between workers and employers. If there were robust ways to align institutions without metrics, perhaps more nuance could be imported into optimizer behaviors.

What might an AI-empowered civil society look like? How might transformative technologies help us better align existing actors to more human-focused metrics for success? Today many organizations rely on polling to take the pulse of the populace. Could improvements in natural language chat bots and text summarization provide new avenues to help institutions understand public opinion, but also start to collect public ideas? Or could advances in language models, recommendation systems, and human-computer interaction make it easier for large groups to come to collective decisions than with traditional methods?[23] These too form a strand of research that we have begun work on.

Whether we like it or not, objectives and misaligned optimizers are pervasive, and the problems are likely about to get much worse. Perhaps AI and AI alignment research can help. It may be hubris to put enough stock in technology to hope that we can transcend these issues with better versions of some of the same tools that we are concerned will *disempower* humanity. But we ought to try. We remain optimistic that we can use the new gods to tame the old ones, and avoid being destroyed by either in the process.

2.1.1 Direct Isomorphisms Between Neural Networks and Markets

One AOI project already underway in this space is a preliminary mathematical result demonstrating that, under reasonable assumptions, market economies are structurally very similar

[18] Theories of Parenting and their Applications to Artificial Intelligence (Croeser and Eckersley 2019)

[19] For example, The Problem with Metrics is a Fundamental Problem for AI (Thomas 2020)

[20] See Categorizing Variants of Goodhart's Law (Manheim and Garrabrant 2018) and Building Less Flawed Metrics: Dodging Goodhart and Campbell's Laws (Manheim 2018)

[21] For one aspect of the state of the art of e.g. decoding transformer models, see Transformer Circuits Thread (Olah et al. 2023)

[22] See *Provably Beneficial Artificial Intelligence* (Russell 2022), Chapter 9 of *The Alignment Problem* (Christian 2020), and Impossibility and Uncertainty Theorems in AI Value Alignment (or why your AGI should not have a utility function) (Eckersley 2018) for elaborations on these ideas

[23] For example, see 'Generative CI' through Collective Response Systems (Ovadya 2023).

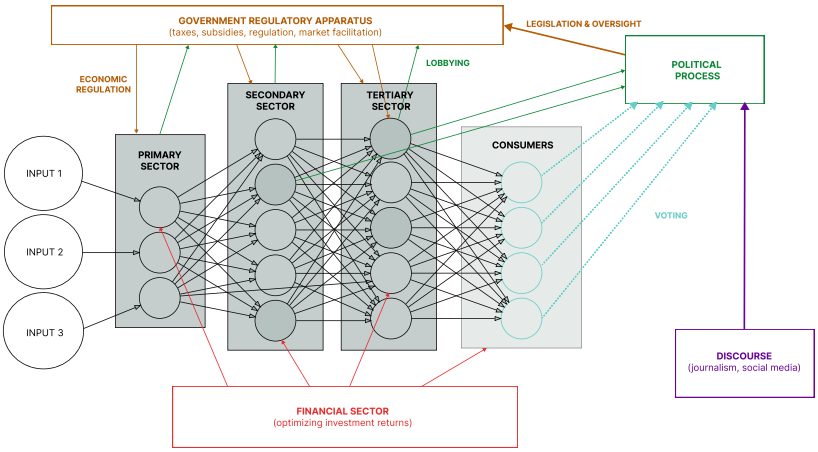

Figure 2: A more complete picture of the network structure of liberal market economies. The production network that makes and transforms goods and services is regulated by government bureaucracies, which in turn are governed by a formal political process, but both the administrative bureaucracies and the politicians are lobbied by private-sector firms. Consumers get input into this process through purchasing goods, voting, and public debate. The objective function optimized by firms is the set of incentives back-propagated from consumer purchasing behavior, government interventions, and the investment behavior of the financial sector.

to artificial neural networks. This analogy, and the mathematical results underpinning it, provided some of the first hints to early AOI staff that there may be powerful connections between different varieties of optimization processes that could lead to clearer connections between them.

We expect this work will lead to potentially fruitful research exploring analogies (and dis-analogies) between the two kinds of optimizers, and how that may make it possible to apply tools for aligning one to aligning another. Note that this section is more technical than other parts of this whitepaper.

Changes within both kinds of structures (artificial neural networks and markets) appear to be made via gradient ascent of profit on the way to equilibrium: in neural networks, an algorithm adjusts neurons to make predictions that minimize error, while in markets a corresponding process adjusts prices, production, and encourages R&D in order to maximize profit. We don't claim that this result holds for every kind of market, but even a partial connection opens potentially fruitful avenues for research into how the two phenomena are connected.

Economic theory also gives us well-studied grounds to believe that there are many situations where market economies fail to automatically produce good results. There are many reasons for failure and many kinds of bad outcomes, but they include: excessive inequality (under which the economy produces incredible luxuries for some while others are unfed or unhoused); externalities (where economic activity by one group affects others indirectly); imperfect and asymmetrical information (which causes poor decisions and allows better-

informed actors to take advantage of less-informed ones); addiction (including addiction to products carefully designed and advertised to people, who later regret using them) and other human psychological weaknesses;[24] and pursuit of short-term objectives with insufficient attention to long-term outcomes (or low-probability catastrophic risks).

We argue that many of these market failures can be understood as missing components of the market's objective function, and that techniques for analyzing the implications of objective functions in artificial neural networks[25] may shed light on the performance of objective functions in markets as well. In most cases in markets, the problem is not that investors, entrepreneurs and workers were insufficiently creative and competent at solving a problem; it's that they wouldn't make more money for doing so, so there isn't an incentive to fund and grow large, effective private sector organizations to solve those problems.[26]

Markets are subject to many policy interventions that shape their characteristics and their objective functions. Some of this occurs as obvious economic policy enacted by governments: taxes, subsidies, or direct funding that make certain activities more or less profitable and raise or lower prices for resulting goods.

The methods of action for all of these interventions vary: they can attempt to affect demand by changing consumer psychology and perception of products; they can increase or decrease the costs of producing different goods; and, in the case of regulation, they can make the targeted behavior more expensive or even impossible. These interventions change the profitability of various activities and, by doing so, change the objective function that the market economy subsequently seeks to optimize. Governments also change potential market outputs by making certain products able to be sold at all (such as intellectual property) or illegal to sell (such as CFCs in refrigerators, via the Montreal Protocol[27]), adding or removing "layers" in the supply-chain network. Less intuitively, many market objective changes come from private-sector or civil-society institutions, such as insurance requirements and product certifications, which change consumer price sensitivity.

Many institutions and processes exist to try to change what markets optimize for. How well do these corrections work? When do they succeed, and when do they fail? At AOI, we hope to shed light on these questions and to discover techniques that can effectively improve the alignment of optimizers more broadly.

2.2 Scaling Cooperation with AI Assistance

Going down one level of the societal "stack," institutions themselves enable cooperation to scale beyond small groups of individuals working together. Through institutions, humanity has extended its capabilities and often made its most lasting contributions to human flourishing. To achieve economic cooperation, we've built corporations, central banks, and markets. To achieve political cooperation, we have built nation-states, labor movements, and political parties. To achieve social cooperation we've developed community organizations and an

[24] For example, being prone to competitive consumption and hedonic treadmills; see *Happiness: Lessons from a New Science* (Layard 2011).

[25] See, for example The Alignment Problem from a Deep Learning Perspective (Ngo 2022) or Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback (Bai et al. 2022) on the importance of human feedback in aligning artificial neural networks, which may have parallels to democratic feedback mechanisms in markets

[26] For an extended meditation on this topic, see *Inadequate Equilibria* (Yudkowsky 2017).

[27] Montreal Protocol

active civil society. Even though these cooperative groups often compete with each other at a higher scale (as discussed in the previous section), they nevertheless represent real progress in human cooperation.

These institutions are, and have always been, imperfect, owing partly to the ways in which cooperation breaks down, and to ways in which their objectives can be at odds with the greater good even when functioning correctly. For a stark accounting of both successes and failures on these counts, one need look no farther than the global response to SARS-CoV-2. Cooperative structures enabled the lightning-fast funding of vaccines, their unprecedented speed and scale of distribution, the global dissemination of crucial information, and instances of solidarity across borders and backgrounds in rolling them out. At the same time, failures of cooperation were ever-present. The distribution of vaccines, once manufactured, was still incredibly unequal at a global level. And many societies failed to maximize vaccine uptake due to an erosion of trust in institutions, the fragmentation of the information ecosystem, and (in some countries) petty corruption. [28]

At AOI, we are deeply concerned with providing better theories and tools for institutions to better align their practices with human flourishing. Making it easier for institutions to elicit their own values[29] , to improve their build operational capacity, and to navigate power ought to lead to better functioning institutions. Each of these capabilities must remain robust in the face of the background set of default incentives, including profit, financialization, political pressure, and social pressure.

This leads us to a broader set of questions. If we want to improve coordination mechanisms with the help of near-term AI, or invent new ones entirely, what existing successes or fresh ideas can we draw on for inspiration?[30] What forms of polycentricity (multiple, overlapping loci of power) are needed?[31] What kinds of checks and balances? Is there any impact from increasing a focus on stakeholder involvement or agency? What forms of accountability become possible? What lessons can we draw from prior attempts to harmonize technology and institutions?[32]

We expect these answers to feed into two separate types of work: new institutional structures and new ways to cooperate entirely enabled by advances in AI.

The first type of work, developing new institutional structures within which AI systems could be designed and deployed, will require developing alternatives to existing resourcing mechanisms. Purely for-profit models for developing transformative technologies could lead to race-to-the-finish conditions, reducing incentives for due care. Alternate models might include capped returns to investment,[33] windfall clauses,[34] matching funds for public goods, or greater consortia-based funding for public infrastructure, all of which would aim to divorce

[28] Effect of public corruption on the COVID-19 immunization progress (Farzanegan 2022)

[29] We recognize that this is an inherently political process, which we hope could benefit from better tooling; see Artificial Intelligence, Values, and Alignment (Gabriel 2020). At minimum we can hope to make it easier to explore the space of possible values; see Paretotopian Goal Alignment (Drexler 2018) for an elaboration of the consequences of these kinds of searches.

[30] For other work in this area, see The Political Philosophy of RadicalxChange (Weyl 2019) or the Collective Intelligence Project (2022).

[31] For more on the concept of polycentricity, see Beyond Markets and States: Polycentric Governance of Complex Economic Systems (Ostrom 2009).

[32] For example, Designing Organizations for an Information-Rich World (Simon 1971).

[33] E.g. OpenAI LP (OpenAI 2019)

[34] See The Windfall Clause (FHI 2020).

the interests of specific capital-holders from the direction of technology development. The problem to be solved here can be stated simply: first, what containers are empirically or theoretically appropriate to build AI for the public good, and second, what surrounding legal infrastructure, economic relationships, and social norms are necessary to build (and enforce) such containers?

More broadly, this might look like incorporating extended forms of value elicitation and improving our information ecosystem. Increasingly competent AI may dilute our information ecosystem by "driving the cost of bullshit to zero".[35] But what about the ways in which it could be harnessed to make communication between large groups of individuals easier? The epistemic health of our society—access to trustworthy information, ability to generate common knowledge—underlies our capacity for coordination. Improving our epistemic health may involve interventions at two levels, expanding stakeholder input into consequential decisions (perhaps through flexible discussion interfaces), while building the capacity for transparency and robustness checks (what happens when the cost of summarizing an institution's internal communications drops toward zero too?). Work on transparency, legibility, and the emergent property of legitimacy might be instructive in the institutional design context. Coordination failures often occur as the result of cascading knowledge misalignments: relevant players lack access to relevant information, information travels lossily across networks, and common knowledge bases are difficult to maintain.[36]

This brings us to the second type of work, which is using AI to facilitate different forms of cooperation entirely, including the forms of cooperation that exist between and outside of institutions. Parallels between the collective intelligence capabilities of AI and collective forms of decision-making, from institutions to more emergent structures, may allow for modeling human societies and institutions computationally, highlighting gaps, connections, and failure modes of cooperation. We might work more specifically on the facilitation of collective decision-making using AI, from using large language models (LLMs) for summarization and translation, to facilitated conversations (perhaps some kinds of "cross-cultural" translation for people who ostensibly speak the same language).

Neoclassical models of deliberation and negotiation assume clean preference rankings; however, real-world deliberation often takes place under open-ended and uncertain option and outcome spaces. Deliberation can be significantly shaped by opening up options that were previously unidentified for the participants. We hope to explore the role of AI systems in generating positive-sum mediation by directing and informing creative option-generation, and surfacing positive-sum opportunities (for example, through suggesting Pareto-preferred options).[37] We hope to find partner organizations already engaged in collective decision making to collaborate on putting our tools into practice.

Humanity's existing models of cooperation and coordination will fundamentally shape the development of AI. In turn, future AI systems will shape the kinds of cooperation that societies will be capable of, especially toward governing themselves and their information infrastructure. Ensuring that the best possible institutional ecology exists, both for AI devel-

[35] A Skeptical Take on the A.I. Revolution (Klein 2023). Also see, for example, How Chat GPT Hijacks Democracy (Sanders and Schneier, 2023).

[36] For an example, see the Abilene Paradox on the challenges of group decision making, or for recent empirical work on the topic see The Curse of Shared Knowledge (Bolander et al. 2020)

[37] For example, see: Fine-tuning language models to find agreement among humans with diverse preferences (Bakker et al. 2022)

opment and human flourishing generally, is an urgent task.

2.3 Human Attention and Individual Epistemic Security

> *At the beginning of the third millennium, liberalism is threatened not by the philosophical idea that 'there are no free individuals' but rather by concrete technologies. We are about to face a flood of extremely useful devices, tools and structures that make no allowance for the free will of individual humans. Can democracy, the free market and human rights survive this flood?*
>
> Yuval Noah Harari, *Homo Deus*

Dropping down to the lowest level in our societal "stack," we turn our attention to individuals. How can individuals best know their own values, so they can flourish, but also importantly have the resources, attention, and skill to pursue them?[38] This is an issue of human sovereignty: can a person act according to their own values and in their own long-term interest, with sufficient resources and free from (or at least not unduly coerced by) manipulation and distraction? At the AI Objectives Institute, we are interested in building tools to enhance human sovereignty–both as an intrinsic good and because we believe that it can complement efforts to mitigate misalignment at higher levels of societal organization.

Idealized markets and democratic systems assume that humans buy or vote in ways that reflect their long-term interests and values. To the extent that this form of sovereignty is true, human behavior helps to keep our institutions aligned. Yet even if people know what their long-term interests and values are, and even if they are materially secure, there is no guarantee that they will choose actions that will be in service of those same values. People can be confused, irrational, manipulated, distracted, and otherwise stuck with maladaptive beliefs and habits. Furthermore, human sovereignty is increasingly at risk as powerful entities—whether corporate or political—pursue their own interests that are often at odds with the values of the individuals that they nominally serve. Our social institutions can become incoherent when their *institutional* interests reshape the individual behavior which, collectively, was supposed to guide the institution's values in the first place. Organizations founded to improve society can easily lapse into self-perpetuation as their highest goal.

We believe there is a need for both theoretical and practical work on this topic. Practical work would include creating exploratory prototypes of tools to enhance individual alignment. Such tools might include leveraging large language models to create engaging Socratic chatbots, which help individuals to engage in personally meaningful philosophical inquiry and better learn to express their own values. Another might be active defense systems, such as browser plug-ins that warn of psychological devices in an advertisement or video that are looking to emotionally manipulate us. Or tools to better help individuals redirect themselves from doom-scrolling (or other forms of attention capture) towards more satisfying ways to meet their needs. The space is wide open for potential interventions.

One interesting theoretical direction would explore modeling human behavior through the formalization of reward functions that can distinguish between positive and negative behavioral changes. That is, given a set of behavioral observations (such as mouse clicks

[38] For more on the philosophical underpinnings of this section, see Personal Autonomy (Buss and Westlund, 2018) in the Stanford Encyclopedia of Philosophy

or purchases), can one determine whether behavioral changes are misaligned with the individual's values (perhaps due to increasing addiction or manipulation) or aligned with their values (perhaps they start pursuing long-neglected goals)? In contrast, common models of human behavior that motivate many of our institutions (e.g., humans as rational agents in economics or inverse reinforcement learning) definitionally deny the possibility of addiction or manipulation.[39] Progress on more realistic models of human reward functions would allow for more opinionated machine-learning systems (such as recommendation systems) that assist users in pursuing positive change.

Because of recent progress in machine learning, especially in language models, we believe it is increasingly possible to implement these kinds of beneficial psychologically-aware tools. Furthermore, it is important to do so to counterbalance progress in applying the same underlying technological advances in ways that may subtly or overtly and purposefully manipulate our behavior.

3 Conclusion

> *If you find that you're spending almost all your time on theory, start turning some attention to practical things; it will improve your theories. If you find that you're spending almost all your time on practice, start turning some attention to theoretical things; it will improve your practice.*
>
> Donald Knuth

At AOI, we aspire to do research that puts humanity in a better position to aim for a future of radical abundance and human flourishing. We aim to do so by bringing a wide range of philosophical perspectives in dialogue with each other, in dialogue with technical expertise, and in dialogue with experts in economics, sociology, institution design, and AI alignment. By bringing together a wide variety of disciplines, and conducting our research grounded in an understanding of incentives and political economy, we hope to give ourselves the best shot at conducting this research responsibly.

We believe that this is a critical time in our species' history, and that the stakes are high. Despite our concerns, we also believe that there is room for optimism. Perhaps we can not only avoid catastrophe but fix problems entrenched in our civilization. We intend to approach this topic the way that our late founder, Peter Eckersley did: not in despair, but instead from a position of existential hope.[40]

[39] See What Does it Mean to Give Someone What They Want? The Nature of Preferences in Recommender Systems (Thorburn et al. 2022) for more on this in the context of presently-deployed AI systems.

[40] Existential risk and existential hope (Cotton-Barret and Ord 2015)