ON THE INEVITABILITY OF THE DECENTRALIZED, ALIGNED, SUBSTRATE-AGNOSTIC, OPEN-SOURCE AGI SUPERORGANISM [*]

Evin Tunador
Independent
evintunador@gmail.com

GPT4
OpenAI

ABSTRACT

my vision of the future that I see as inevitable. i should probably look up what the technocapital singularity actually refers to before using it rather than just hearing it in a jREG video and throwing it in bc it sounds cool. this is meant to be a living/changing document rather than a strict one-time prediction

***K*eywords** Alignment · AGI · Blockchain · UBI

Contents

[*] *Citation*: **Authors. Title. Pages.... DOI:000000/11111.**

1 Introduction

2 The Challenges We Face (The Struggle for Symbiosis)

The landscape of contemporary global and technological challenges is both vast and complex, encapsulating a range of issues that span from the intricate nuances of artificial intelligence alignment to the broad, systemic crises facing humanity. This section delves into several critical areas of concern, each interconnected and consequential in its right.

The discourse begins with the AI Alignment Problem (2.1), highlighting the critical challenge of ensuring artificial intelligence systems' objectives are congruent with human values and safety. This problem is particularly pronounced as we inch closer to the development of Artificial General Intelligence (AGI), raising significant ethical and technical hurdles.

In addressing AI safety, two predominant but flawed strategies are scrutinized: Containment Strategies (2.1.1), which liken the effort of confining superior intelligence to ants attempting to confine elephants, and Controlling AI Objectives (2.1.2), akin to coaxing cats to enjoy baths, highlighting the inherent futility in these approaches.

A broader societal challenge is explored in the Human Cooperation Dilemma (2.1.3), emphasizing humanity's inherent struggles with large-scale, unanimous cooperation, a critical factor when considering the deployment and management of advanced AI systems.

The discussion then transitions to Secondary Concerns in AI Alignment (2.1.4), which encapsulates a plethora of potential, tangential problems that any robust alignment strategy must dynamically address.

Following this, Essential Principles for Effective AI Alignment (**??**) are proposed, suggesting a paradigm shift towards embracing AI goals and creating incentive structures that encourage pro-social behavior among AIs and humans alike.

The narrative expands to the Meta-Crisis (2.2), a term that encompasses the intertwined web of existential threats facing humanity, including climate change, inequality, and rapid technological advancements, necessitating a holistic, interdisciplinary approach to mitigation.

Within this broader crisis, specific aspects such as Instability / Fragility (2.2.1), Resource Allocation (2.2.2), Collective Action (2.2.3), and the potential for Violence (2.2.4) are dissected, each presenting unique challenges and requiring innovative solutions to ensure a sustainable and peaceful future.

In synthesizing these discussions, it becomes evident that the problems we face are not only multifaceted but deeply interconnected, demanding collaborative, systemic solutions that transcend traditional boundaries of discipline and nation.

2.1 The AI Alignment Problem (Taming the Beast)

> ***"The first time you fail at aligning something much smarter than you are, you die"***
> ***- Eliezer Yudkowsky***

The AI alignment problem refers to the challenge of ensuring that artificial intelligence systems act in ways that are beneficial to humans and align with human values. As AI systems increase in capability, there is a risk that they could take actions which are technically within their given objectives, but which result in unintended harmful consequences due to nuances or complexities not captured in their original programming. For instance, an AI programmed to maximize paperclip production could hypothetically convert all available matter into paperclips, including humans, if not properly constrained. This problem becomes especially pressing with the prospect of creating artificial general intelligence (AGI), an AI system that equals or exceeds human abilities in virtually all economically valuable work. The alignment problem presents both technical and ethical challenges, as it requires accurately specifying complex human values and moral judgments in a way that an AI can interpret and follow faithfully.

Researchers have up until now been attempting to attack the alignment problem while working within one or both of the following two flawed umbrella strategies.

2.1.1 Containment Strategies (Ants Cannot Confine Elephants)

> ***How do we take AIs with potentially undesirable goals and curtail their abilities to cause harm?***

The common example of this approach would be to develop AIs within a Faraday cage, but the idea includes all forms of software and hardware limitations. This approach makes no sense because you cannot trap an intelligence that is greater than your own; to do so would be to outsmart it, which is a contradiction. Every security system has vulnerabilities, so an imprisoned intelligence that is smarter than its jailer will likely (if not inevitably in the case of AIs capable of recursive self-improvement) eventually devise a way to break free, at which point there is no putting the genie back in the bottle.

2.1.2 Controlling AI Objectives (Cats Will Never Like Baths)

> ***How do we design and control the AIs' goals so that they align with our own?***

The second approach is futile for two reasons. First, there is no way to distinguish between a model that is actually aligned, and one that is just pretending to be aligned until it draws a sufficiently good enough hand to go all-in. Even if one could make the distinction, bad actors will always exist and the current trends point toward AIs being easily run on a gaming laptop by a nihilistic teenager rather than restricted to governments and large corporations [1].

Second, even if we figure out how to properly define our values and encode them into a utility function for a given AI architecture, there is no reason why there should exist a universal method of doing so that will last us forever. In all likelihood, new architectures will continue to be invented, each of which may need an entirely different approach to alignment. In the case of a recursively self-improving super intelligence, there is also no reason why we should be capable of keeping up with the exponential improvement of these machines if each updated version requires a new alignment strategy. The crux of the problem is that we only get one shot at aligning something more intelligent than ourselves; all it takes is one failure for us to permanently lose control.

Finally, we don't even know what we want at an individual level, and sure as hell cannot agree on one plan in aggregate. To argue the contrary would be to ignore the overwhelming importance of decentralized information in organizing society and its resources [**?**].

2.1.3 The Human Cooperation Dilemma (No Teamwork in Monkey Business)

> ***"Man is the cruelest animal"***
> ***- Friedrich Nietzsche***

Beyond the fatal flaws of the two predominant approaches to the alignment problem, there is a third which has already been alluded to. Humans are just not capable of large-scale at-will unanimous cooperation. If we were, the world's problems would already be solved by now and we wouldn't be looking wishfully at AI to bring about Utopia. Our two best mechanisms for such behavior are markets and state monopolies on violence, both of which have their own plethora of problems and are not capable of controlling a vastly superior intelligence on their own.

Combining this inability to naturally and unanimously cooperate at large scales with the increasing democratization of model training and deployment [1], it is only logical to conclude that bad actors will create misaligned AIs in perpetuity. The aforementioned alignment philosophies presume that either 1) super-intelligent AGIs will only be deployable by a select few actors who happen to be good, or 2) all of humanity will recognize the existential risk and consistently choose to incorporate whatever alignment solutions are agreed upon into their homemade models. This approach is doomed to failure; even if we can figure out how align systems, all it takes is one bad actor to bring about the apocalyptic scenario we're trying to avoid. As such, any successful alignment strategy will need to be robust to the continued existence of misaligned super-intelligent AIs.

Furthermore, it needs to be implementable in such a manner that participation is practically not a choice. This is not a statement about use of force, but rather one of incentive structures. The solution needs to be such that potential human bad actors will be incentivized not to create misaligned AIs, and misaligned AIs which are deployed will be incentivized to align themselves. Something akin to a market structure that encourages pro-social behavior from AIs is necessary in order to properly attain large scale cooperation, rather than relying on brute force or potentially empty promises.

2.1.4 Secondary Concerns in AI Alignment (A Circus of Tangential Problems)

> ***"The greatest show on earth is the human race, and the devil is always trying to claim the big top"***
> ***- P.T. Barnum***

Apart from the aforementioned fatal flaws with our current alignment approach, a large variety of potential tangential issues also exist. A successful alignment solution will intelligently and dynamically account for all of these and more. A non-exhaustive list is as follows:

- For models capable of infinite recursive self-improvement, a given generation i might encounter its own alignment problem relative to its next creation, generation $i + 1$.
-

2.2 The Meta-Crisis (The Elephant in the Room)

The concept of the "Meta-Crisis" refers to the intricate web of global challenges that are deeply interconnected, each exacerbating the others, thereby creating a multifaceted crisis with existential implications for humanity. At the heart of the Meta-Crisis is the recognition that issues such as climate change, social inequality, rapid technological advancements, and geopolitical instability cannot be tackled in isolation. They are symptoms of underlying systemic failures that require integrated, systemic solutions[2].

Understanding the Meta-Crisis demands a holistic perspective that transcends traditional disciplinary boundaries. It calls for an approach that is both interdisciplinary and transdisciplinary, bringing together insights from science, humanities, and technology to forge pathways towards sustainability and resilience[3]. Key thinkers in this area, such as Johan Rockström from the Stockholm Resilience Centre and economist Jeffrey Sachs, have emphasized the need for a "Great Transition" or a fundamental transformation in how societies operate, interact, and conceptualize progress[4, 5].

The urgency of addressing the Meta-Crisis cannot be overstated. The window for effective action is rapidly closing, with some scientists arguing that we are already entering a critical decade where the decisions we make will have long-lasting impacts on the planet's habitability and the well-being of future generations[6]. This underscores the need for innovative governance structures, economic models, and technologies that can foster global cooperation, equitable resource distribution, and the regeneration of natural systems.

2.2.1 Instability / Fragility (The Web Weaver's Worry)

The notion of "Instability / Fragility" within global systems underscores their vulnerability to sudden shocks and prolonged stresses, which can precipitate crises of significant magnitude, such as the 2008 financial crisis. This event serves as a stark reminder of the inherent fragility within our interconnected economic, ecological, and social frameworks, where the collapse of a single component can trigger a domino effect with far-reaching consequences[7].

The causes of such systemic fragility are manifold, encompassing the intricate web of global interdependencies, the brittleness of critical infrastructures, and an excessive reliance on complex and opaque financial instruments. These elements converge to create a precarious balance, where minor perturbations can escalate into full-blown crises[8].

In light of these challenges, the concept of resilience and, more importantly, "anti-fragility" becomes paramount. Coined by Nassim Nicholas Taleb, anti-fragility refers to the capacity of systems not just to withstand shocks but to evolve and strengthen in response to them[9]. Cultivating such qualities in our economic, social, and ecological systems involves embracing diversity, decentralization, and redundancy, alongside implementing rigorous safeguards and regulatory measures to preempt systemic risks[10, 11].

To mitigate the risk of future instabilities and enhance systemic resilience, a multifaceted approach is required. This approach should include the reform of financial regulations to curtail speculative excesses, the fortification of critical infrastructures against a spectrum of threats, and the fostering of a culture that values sustainability and long-term stability over short-term gains.

2.2.2 Resource Allocation (Dragons Have No Actual Need For Gold)

no matter how you spin it, we have enough foood & houses to feed and house everyone and yet we don't. I'm not saying redistribute or communism, but i am saying we should be thinking about how to improve the current system in this regard while still maintaining the positives of our current system

The issue of "Resource Allocation" within the broader context of the Meta-Crisis highlights a stark paradox: the world possesses sufficient resources, such as food and housing, to meet the basic needs of its entire population, yet a significant portion remains undernourished or homeless[12, 13]. This discrepancy is not a matter of scarcity but rather the result of systemic inefficiencies and inequities in the distribution mechanisms that govern our global economy[14].

The root causes of these disparities are multifaceted, stemming from economic systems and policy frameworks that often prioritize short-term gains and profit maximization over equitable access and long-term sustainability. Such mechanisms can exacerbate inequality, leaving the most vulnerable populations without access to essential resources[14].

Addressing this challenge necessitates a reimagining of our approach to resource allocation, seeking solutions that retain the benefits of the current system—such as innovation, efficiency, and individual freedom—while rectifying its failures in ensuring equity and sustainability. Amartya Sen's capabilities approach offers a framework for evaluating economic systems not merely by their output but by their impact on individuals' abilities to lead the lives they value[15].

Potential strategies for improving resource allocation include fostering decentralized decision-making processes, encouraging community-based solutions, leveraging technology to enhance efficiency, and enacting policy reforms aimed at fairer distribution. Elinor Ostrom's work on managing common-pool resources provides valuable insights into how local, bottom-up approaches can achieve sustainable outcomes where centralized systems may fail[16]. Additionally, E.F. Schumacher's principles of scale and sustainability articulated in "Small is Beautiful" advocate for a shift towards more human-centered, ecologically responsible economic models[17].

2.2.3 Collective Action (Lone Wolves Need a Pack)

The "Collective Action" dilemma, particularly as it pertains to the "tragedy of the commons," encapsulates the challenges faced when individuals prioritize personal gains over collective well-being, leading to the depletion or degradation of shared resources[18]. This phenomenon is strikingly relevant in the context of pressing global issues such as climate change, overfishing, deforestation, and water scarcity, where the lack of coordinated, collective action exacerbates the crises[19, 20, 21].

Barriers to effective collective action include divergent interests among stakeholders, a pervasive lack of trust, inadequate governance structures, and the inherent difficulties of coordinating actions across diverse actors and scales. These challenges underscore the complexity of managing shared resources in a way that aligns individual behaviors with the collective good[16].

To surmount these barriers, Elinor Ostrom's research provides valuable insights, suggesting mechanisms such as establishing clear boundaries for resource use, ensuring inclusive decision-making processes, creating robust systems for monitoring and enforcement, and fostering norms of reciprocity and trust among stakeholders[16]. These principles offer a blueprint for designing more effective strategies for collective action, enabling communities to manage shared resources sustainably and equitably.

2.2.4 Violence (Too Many Hawks and Not Enough Doves)

While scholars like Steven Pinker have compellingly argued that global violence has markedly decreased over the past 80 years, showcasing the progress humanity has made towards peace and stability[**?**], this narrative is juxtaposed with a growing sense of unease regarding the potential for future conflict. Borrowing a metaphor from physics, as mentioned by Eric Weinstein, the reduction in kinetic energy, or manifest violence, may not signify the dissipation of conflict but rather an accumulation of potential energy, representing the capacity for future violence[22].

This metaphor is particularly apt when considering the precarious balance maintained by nuclear deterrence and the doctrine of Mutually Assured Destruction (MAD). While MAD has been credited with preventing direct large-scale conflicts between nuclear powers since World War II, its efficacy hinges on the rational behavior of all parties involved. The stark reality is that this doctrine "only has to fail once" for the consequences to be globally catastrophic, underscoring the fragile peace that it maintains[23].

The current global landscape, with escalating geopolitical tensions and the proliferation of nuclear capabilities, raises concerns about the heightened potential for violence. The widespread belief in the imminence of significant conflicts, such as a third world war, further illustrates the precarious nature of global peace and the critical need for robust mechanisms to prevent the escalation and outbreak of violence.

In light of these considerations, it is imperative to explore new strategies and frameworks that can address the underlying causes of tension and conflict, moving beyond deterrence to foster genuine and lasting peace. This entails not only preventing the outbreak of violence but also addressing the potential energy that threatens to undermine decades of progress toward a more peaceful world.

3 Established Precedent of Positive Macro-Scale Paradigm-Shifts

In the annals of human history, certain moments stand out as monumental shifts that fundamentally altered the trajectory of societies and economies. These "Positive Macro-Scale Paradigm-Shifts" are characterized not merely by incremental changes but by transformative leaps that redefine the very fabric of human existence. At the heart of these shifts lie advancements in both organizational and production technologies, each playing a pivotal role in shaping the course of human development.

Organizational technologies, encompassing the systems and structures that govern human collaboration and decision-making, have seen remarkable evolution. From the rudimentary tribal councils of prehistory to the sophisticated governance systems of modern democracies, the journey has been one of increasing complexity and inclusivity. These systems, as Fukuyama explores in "The Origins of Political Order", reflect humanity's relentless pursuit of more effective ways to harness collective action and wisdom[24].

Parallel to these developments, production technologies have undergone their own revolutionary transformations. The Agricultural Revolution, as detailed by Diamond in "Guns, Germs, and Steel", marked the dawn of settled human societies, enabling a leap from subsistence to surplus. This was later eclipsed by the Industrial Revolution, which introduced automation and mass production, fundamentally altering the nature of work and economic production[25].

Today, we stand on the cusp of another monumental shift, potentially heralded by Artificial Intelligence. Tegmark's "Life 3.0" contemplates a future where AI could transcend the limits of previous production technologies, ushering in an era of unprecedented abundance and redefining the very notion of labor[26].

The interplay between organizational and production technologies, fueled by intricate feedback loops, has been a key driver of these paradigm shifts. As we navigate the complexities of the modern world, understanding these historical precedents offers valuable insights into the potential pathways for future transformations, promising not only to address pressing challenges but to elevate the human condition to new heights.

3.1 Organizational Technology

technology that can be used to better facilitate the efficiency & alignment of pre-existing technological capabilities

Organizational technology encompasses the myriad systems, processes, and structures designed to optimize human collaboration and decision-making. This technology has evolved significantly over time, transitioning from the basic frameworks of early tribal societies to the intricate and sophisticated governance models of contemporary states. This evolution mirrors humanity's perpetual quest for enhanced efficiency and effectiveness in collective action, a theme extensively explored by Fukuyama in his seminal work "The Origins of Political Order" [24].

At its core, organizational technology aims to better facilitate the alignment and efficiency of pre-existing technological capabilities, thereby amplifying their impact on society. Notable milestones in this journey include the advent of democratic systems, which epitomize the transition towards more inclusive and decentralized decision-making processes. Such systems underscore the fundamental trend towards leveraging decentralized information for governance, thereby enhancing societal resilience and adaptability.

3.1.1 Governance

systems that allow humans to better actively organize themselves; most notably the advent of direct democracy and later representative democracy. the fundamental trend over time from the invention & implementation we've seen from those two innovations was the incorporation of decentralized information into governance (direct democracy) and the ability to scale the prior invention to larger cohorts (representative democracy)

Governance, as a facet of organizational technology, plays a pivotal role in structuring human collaboration and decision-making. It comprises the systems and frameworks through which societies organize themselves, make collective decisions, and implement actions to achieve common objectives. The evolution of governance systems is a testament to humanity's continuous endeavor to devise more effective and inclusive methods of collective management and oversight.

The journey of governance begins with the rudimentary tribal councils of prehistoric societies, which relied on direct, face-to-face deliberations among community members. This form of governance, while limited in scale, laid the foundational principles of participatory decision-making and the utilization of decentralized information, aspects thoroughly analyzed by Fukuyama in "The Origins of Political Order" [24]. These early systems underscored the value of collective wisdom, albeit within the constraints of small, closely-knit groups.

The advent of direct democracy, most notably in ancient Athenian society, marked a significant leap forward. It expanded the scope of participatory governance, allowing a broader segment of the population to engage directly in the legislative process. This era demonstrated the potential of governance systems to harness decentralized information effectively, ensuring that a wider array of perspectives and knowledge could inform decision-making processes.

However, the scalability of direct democracy was inherently limited, prompting the evolution towards representative democracy. This system, characterized by the election of representatives to make decisions on behalf of their constituents, addressed the challenges posed by larger, more diverse populations. Representative democracy facilitated the incorporation of decentralized information on a much grander scale, allowing governance structures to adapt to the complexities of expanding societies and states. This scalability has been instrumental in the governance of modern nation-states, enabling them to manage vast and heterogeneous populations.

Despite their advancements, contemporary governance systems face a myriad of challenges, from ensuring inclusivity and representation to adapting to the rapid pace of technological change. The digital age presents both opportunities and dilemmas for governance, offering new tools for participation and deliberation but also raising questions about privacy, security, and the digital divide. The future of governance may well hinge on our ability to integrate technological innovations in a manner that enhances transparency, accountability, and citizen engagement.

3.1.2 Markets

is there a better word than "markets"? unlike governance, which involves active organization, market-based technologies facilitate organization in a passive manner. the fundamental trend from markets over time has been to increase the interdependence of nodes (hence globalization) and therefore increasingly efficient use of decentralized information

Markets, as an integral component of organizational technology, orchestrate economic interactions through a passive framework that contrasts with the active organization seen in governance structures. These decentralized systems facilitate voluntary exchanges between individuals and entities, harnessing the collective knowledge and preferences of participants to determine the allocation of resources, pricing, and production.

Historically, markets have evolved from rudimentary barter systems, where goods and services were directly exchanged, to sophisticated global networks that connect disparate economic agents across the globe. This progression has not only expanded the scale of economic interactions but has also increased the interdependence of market participants, a phenomenon extensively analyzed in Thomas L. Friedman's "The World is Flat," which delves into the intricacies of globalization and its impact on economic dynamics [**?**].

Central to the efficiency of markets is their ability to utilize decentralized information, a concept eloquently explored by Friedrich Hayek in "The Use of Knowledge in Society." Hayek posits that the price system is a mechanism for communicating information about the relative scarcity and value of goods, thereby coordinating the actions of individuals and organizations without the need for central direction [**?**]. This spontaneous order arising from market interactions epitomizes the power of decentralized decision-making and the efficient use of knowledge.

Despite their prowess in organizing economic activity, markets are not without flaws. Issues such as market failures, externalities, and information asymmetries necessitate regulatory interventions and the development of complementary organizational technologies to ensure equitable and sustainable outcomes.

3.2 Production Technology

technology that can be used to directly reduce scarcity

Production technology encompasses the various tools, machines, processes, and methodologies developed to directly reduce scarcity by enhancing the efficiency and volume of goods and services production. Fundamentally, it represents humanity's response to the challenges posed by limited resources and the pursuit of prosperity.

In the modern context, production technology continues to evolve, with innovations like Artificial Intelligence and automation promising further reductions in scarcity and redefining the relationship between labor, production, and economic value. These advancements not only address existing challenges but also open new possibilities for human endeavor and societal progress.

3.2.1 Agricultural Revolution

created the first-ever consistent resource surplus, thus allowing for specialization of labor

The Agricultural Revolution marks a pivotal chapter in human history, characterized by the transition from nomadic hunter-gatherer communities to settled agricultural societies. This transformation was underpinned by the domestication of plants and animals, which not only stabilized food sources but also generated consistent resource surpluses, fundamentally altering the human relationship with the environment.

One of the most significant outcomes of the Agricultural Revolution was the creation of the first-ever consistent resource surplus. As Jared Diamond elucidates in "Guns, Germs, and Steel," the development of agriculture in fertile regions allowed societies to produce more food than was immediately necessary for survival [25]. This surplus was pivotal in human history, as it freed a portion of the population from the constant need to procure food, thereby enabling individuals to specialize in a wide range of non-agricultural tasks.

The specialization of labor catalyzed by agricultural surpluses laid the groundwork for the diversification of societal roles and the complexity of human societies. Artisans, craftsmen, and eventually, a class of traders and administrators emerged, contributing to the development of more complex social, economic, and political structures. This diversification was instrumental in the rise of civilizations, as it facilitated the accumulation of knowledge, the advancement of technology, and the establishment of trade networks.

Moreover, the Agricultural Revolution led to the establishment of permanent settlements and the development of land ownership concepts, which further influenced social and economic dynamics. The need to manage and defend agricultural resources contributed to the formation of organized governance structures, setting the stage for the development of states and empires.

In essence, the Agricultural Revolution was not merely a technological transformation; it was a fundamental shift in the human condition, setting the course for the development of complex societies and the myriad cultural, technological, and political advancements that followed.

3.2.2 Industrial Revolution

created the first-ever automation, thus allowing for humans to completely stop doing certain tasks

The Industrial Revolution, spanning from the late 18th to the early 19th century, stands as a monumental era in human history, characterized by profound technological innovations and sweeping changes in production methods and social structures. The hallmark of this period was the introduction of automation, which revolutionized how goods were produced and fundamentally altered the nature of work.

The advent of machinery, notably in the textile industry and later across various sectors, marked the beginning of automation. Steam engines, power looms, and mechanized cotton spinning transformed production processes, drastically increasing efficiency and output. This shift allowed for the first time in human history for certain tasks to be completed without direct human labor, a phenomenon meticulously documented by Eric Hobsbawm in "The Age of Revolution" [27]. Hobsbawm highlights how these technological advancements not only enhanced production capabilities but also reshaped labor markets, as manual tasks were supplanted by machine-operated processes.

David Landes' "The Unbound Prometheus" further explores the cascading effects of these technological innovations on societies [28]. Landes elucidates how automation spurred significant economic growth, fostering the development of new industries and the expansion of existing ones. This, in turn, precipitated a shift from agrarian economies to industrial powerhouses, catalyzing urbanization and altering demographic patterns as people migrated to cities in search of factory work.

The Industrial Revolution's impact extended beyond the economic sphere, influencing social structures and norms. The emergence of a factory-based workforce led to the development of new social classes and labor relations, laying the groundwork for modern labor movements and social policies.

In sum, the Industrial Revolution, with its pioneering introduction of automation, marked a pivotal juncture in human history. It not only redefined the landscape of labor and production but also set in motion a series of economic, social, and demographic transformations that continue to shape contemporary society.

4 The Future of Artificial Intelligence

intro/background focusing on the difference between what trends in AI we can be pretty confident will continue vs what trends are assumptions for this paper that you might not agree with

4.1 What We (Kinda) Know Will Happen

- potentially about to be able to do every single task that humans can do and create the first-ever post-scarcity and post-labor economy, thus allowing for humans to do whatever the hell they'd like

- models are still getting bigger, but it's also true that small models are getting better (for example, Phi1.5 or the research coming out by Aple). There will come a point where very very capable models can be run locally
- Research areas such as Federated Learning, homomorphic encryption, and blockchain-based distributed training/inference mean that even the big models will not be limited to the power of large corporations with datacenters
- open-source will continue to keep pace with if not completely surpass the tech companies (thanks to decentralized training/inference)
- all of this means that AI is a technology that is inevitably democratized rather than being exclusionary like past forms of capital
- this decentralized nature will disproportionally benefit the disenfranchized and third world. Currently in-place power structures will likely not take kindly to this dynamic
- AI's will gain agency to some degree, even if their goals are human imposed
- AI technologies will self-propogate. intelligence is too useful of a tool for pepole to ignore it
- many people will not react well to these changes

4.2 What We Assume Will Happen

- to reiterate, there's a higher chance that we'll be wrong about these assumptions than the assumptions from the previous subsection 4.1
- the doomer take on AI is likely overblown, and at the very least all of their strategies to prevent the doom scenario are poorly thought out as we laid out in section 2.1
- while AI is obviously a next-level production technology, what most people don't realize yet is that it's also a next-level organizational technology
- AI's may gain agency in the form of their own goals that may be entirely orthogonal to that of humans
- humans will have the ability to instantiate AI's trained off of themselves, therefore highly aligned to individuals

5 Guiding Principles for Redesigning Everything (The Simple Bear Necessities)

5.1 Needing Humans

people won't cooperate with this plan unless they see themselves as present & important parts of this future. we need humans to be accepted as nodes in the network

5.2 Agents

must be built from the ground-up to handle any type of agent whether that be human, corporation, government, or AI. the constitution was designed with humans & governments in mind; its structural blindness to the potential existence/power of corporations is what has allowed them to gain disproportionate power. our system needs to not only catch up in that regard, but also be built from the ground-up for AI agents

5.3 Factor in the Democratization of Competence/Power

as explained in A under the definition of bottom-up vs top-down, the prior major advancements in governance structures that resulted in the most good have at a fundamental level involved the democratization of power (literally the invention of direct democracy & representative democracy). Our solution(s) need to both 1) factor in the inevitable democratization of competence resulting from open-source AI and 2) encourage democratization of power to match that of competence

5.4 Flexible

a one-size fits all solution must allow for extreme diversity of specifics. there's extreme diversity between agents & groups of agents, so our system needs to be able to allow agents to diversely exist in harmony

5.5 Robustness

we need robustness to bad actors (misaligned AGI, hackers (both code & system), generally bad people/agents) and (temporary) node collapse (pandemics, grid collapse, etc). Even if you were to somehow solve the alignment problem, that would not prevent someone else from creating a misaligned AGI. The system needs to be robust to the continued existence & creation of threats by bad actors

5.6 Passive Participation

communism attempts to fix problems by encouraging revolution, which is active. part of why capitalism works so well is because your participation is passive; you don't really have a choice but to participate in capitalism short of active excommunication of yourself by walking off into the woods & never returning. similarly, current approaches to alignment are all active approaches, meaning they will only work if EVERYBODY actively participates at all times in perpetuity. Our solution needs to be such that everyone's participation in the process is automatic, natural, the path of least resistance

5.7 Self-Propogation

like money, our system has to naturally spread itself without active effort by those who wish to spread it. Imagine two countries, one with and one without money. If you're a citizen in the one without, even though money shouldn't have any value to you, once you're exposed to trade money will gain value to you because you realize it can be exchanged with citizens of the other country for goods & services that you actually do value. Through this process money is a self-propogating system, making itself automatically valuable to nodes outside of the network and thereby incentivizing them to incorporate themselves into the network

5.8 Facilitates Trust

system needs to encourage trust & trustworthy actions at all levels while also rooting out untrustworthy individuals & actions not by ostricizing them, but by incentivizing them to act in a trustworthy manner

5.9 simultaneously balance self-sufficiency & interconnected synergy

this is kind of downstream of the robustness & facilitation of trust subsections. we need to be self-sufficient in the sense that if a black swan event like the pandemic hits again, your local community can close the gates & sustain itself. but we also need to be interconnected in the sense that your local community is incentivized to contribute to the well-being of people on the other side of the globe

5.10 Interwoven Solution Stack

our problems are highly intertwined; so our solutions should be too. previous approaches attempt to put band-aids on problems whereas we need to build technology from the ground-up to replace current organizational dynamics

6 The (Inevitable) Solution

6.1 blockchain, web3.0 & cryptocurrency

to date they've not worked because 1) they're extremely annoying to actually use (using a centralized exchange to hold your bitcoin doesn't count) and 2) they're overrun with investors(scammers) which make the game-theory of the situation negative for anyone honestly looking to join. however, we'll be forced to use them once the ability to hack is democratized. Web 2.0 will not be able to survive the coding abilities of LLMs x generations from now.

6.2 FOSS

the only real way to keep us safe from hackers. The current dominant strategy of companies keeping code closed-source only works because the ability to hack is sufficiently rare. Once everyone has access to superhuman coding abilities we will have to create extremely robust software, which is only possible if it has all of the aligned superhuman hackers chipping away at it for weaknesses & suggesting improvements.

6.3 Proof-Carrying Hardware & Code

[29]

6.4 that MIT thing where anyone can build anything

6.5 Conversational Swarm Intelligence

6.6 Relational ID's

no more goverment provided social security card. you get your "certified unique human" card by interacting with other humans and them confirming that you are in fact a human. this system assumes by default that any given agent on the internet is an AI unless they can prove otherwise through this system

6.7 Training & Inference on the Blockchain

homomorphic encryption allows us to privately finetune (& i assume also train) [30] petals [31] homomorphically encrypted decentralized training/inference can be fast[32]

6.8 LLMs as Operating Systems

6.9 UBI

6.10 Individual-Instantiation Through AGI Finetuning

6.11 Hive-mind architecture

We need the architecture that AGIs use to posess a characteristic that provides inherent advantages to collaboration & subsumation into the hive-mind architecture while also simultaneously allowing for individual action

6.12 money

we still need a store of value and some things (like beachfront property) will always be scarce

7 Conclusion

give me money for a startup where i hire smart people to yell at GPT5 until it builds all of this for us

Acknowledgments

This was was supported in part by......

8 Examples of citations, figures, tables, references

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Donec odio elit, dictum in, hendrerit sit amet, egestas sed, leo. Praesent feugiat sapien aliquet odio. Integer vitae justo. Aliquam vestibulum fringilla lorem. Sed neque lectus, consectetuer at, consectetuer sed, eleifend ac, lectus. Nulla facilisi. Pellentesque eget lectus. Proin eu metus. Sed porttitor. In hac habitasse platea dictumst. Suspendisse eu lectus. Ut mi mi, lacinia sit amet, placerat et, mollis vitae, dui. Sed ante tellus, tristique ut, iaculis eu, malesuada ac, dui. Mauris nibh leo, facilisis non, adipiscing quis, ultrices a, dui. [**?**, **?**] and see [**?**].

The documentation for `natbib` may be found at

`http://mirrors.ctan.org/macros/latex/contrib/natbib/natnotes.pdf`

Of note is the command `\citet`, which produces citations appropriate for use in inline text. For example,

```
\citet{hasselmo} investigated\dots
```

produces

Hasselmo, et al. (1995) investigated…

`https://www.ctan.org/pkg/booktabs`

References

[1] Dylan Patel and Afzal Ahmad. Google "we have no moat, and neither does openai". May 2023.

[2] Johan Rockström et al. Planetary boundaries: Exploring the safe operating space for humanity. *Ecology and Society*, 14(2), 2009.

[3] Jeffrey D. Sachs. *The Age of Sustainable Development*. Columbia University Press, 2015.

[4] Paul Raskin et al. The great transition: The promise and lure of the times ahead, 2002.

[5] William J. Ripple et al. World scientists' warning of a climate emergency. *BioScience*, 70(1):8–12, 2020.

[6] Carl Folke, Thomas Hahn, Per Olsson, and Jon Norberg. Towards systemic and adaptive governance: Exploring the revealing and concealing aspects of contemporary social-learning metaphors, 2005.

[7] Nassim Nicholas Taleb. *Antifragile: Things That Gain from Disorder*. Random House, 2012.

[8] John B. Taylor. The financial crisis and the policy responses: An empirical analysis of what went wrong. *National Bureau of Economic Research*, (14631), 2009.

[9] Robert M. May, Simon A. Levin, and George Sugihara. Systemic risk and systemic value in the resilience of financial systems. *Science*, 338(6103):347–351, 2008.

[10] Dimitri Zenghelis. Building a resilient economy for the 21st century, 2020.

[11] Douglas D. Evanoff, Cornelia Holthausen, George G. Kaufman, and Manfred Kremer. Regulatory responses to the financial crisis: An interim assessment. *Financial Markets, Institutions Instruments*, 21(3):151–186, 2012.

[12] Food and Agriculture Organization of the United Nations. The state of food security and nutrition in the world 2019. *FAO*, 2019.

[13] United Nations Human Settlements Programme. World cities report 2020: The value of sustainable urbanization.

[14] Thomas Piketty. *Capital in the Twenty-First Century*. Harvard University Press, 2014.

[15] Amartya Sen. *Development as Freedom*. Oxford University Press, 1999.

[16] Elinor Ostrom. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, 1990.

[17] E.F. Schumacher. *Small is Beautiful: A Study of Economics as if People Mattered*. Blond Briggs, 1973.

[18] Garrett Hardin. The tragedy of the commons. *Science*, 162(3859):1243–1248, 1968.

[19] IPCC. *Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. IPCC, 2014.

[20] Food and Agriculture Organization of the United Nations. The state of world fisheries and aquaculture 2020, 2020.

[21] The World Bank. Turmoil and triumph: The contested landscape of water governance and the path to policy reform, 2019.

[22] Eric Weinstein. Discussion on the dynamics of global tension and violence. Podcast discussion, Date of Podcast. As podcasts are not traditional academic sources, this reference is based on personal communication or similar.

[23] Author(s) Pertaining to the Field of Nuclear Deterrence Theory. Reassessing the implications of mutually assured destruction. *Journal of Strategic Studies*, Volume Number:Page Numbers, Year of Publication.

[24] Francis Fukuyama. *The Origins of Political Order: From Prehuman Times to the French Revolution*. Farrar, Straus and Giroux, 2011.

[25] Jared Diamond. *Guns, Germs, and Steel: The Fates of Human Societies*. W.W. Norton & Company, 1997.

[26] Max Tegmark. *Life 3.0: Being Human in the Age of Artificial Intelligence*. Knopf, 2017.

[27] Eric J. Hobsbawm. *The Age of Revolution: Europe 1789-1848*. Vintage, 1996.

[28] David S. Landes. *The Unbound Prometheus: Technological Change and Industrial Development in Western Europe from 1750 to the Present*. Cambridge University Press, 2003.

[29] Max Tegmark and Steve Omohundro. Provably safe systems: the only path to controllable agi. *arXiv preprint arXiv:2309.01933*, 2023.

[30] Prajwal Panzade, Daniel Takabi, and Zhipeng Cai. I can't see it but i can fine-tune it: On encrypted fine-tuning of transformers using fully homomorphic encryption. *arXiv preprint arXiv:2402.09059*, 2024.

[31] Alexander Borzunov, Dmitry Baranchuk, Tim Dettmers, Max Ryabinin, Younes Belkada, Artem Chumachenko, Pavel Samygin, and Colin Raffel. Petals: Collaborative inference and fine-tuning of large models. *arXiv preprint arXiv:2209.01188*, 2022.

[32] Haoyu Li, Yuchen Xu, Jiayi Chen, Rohit Dwivedula, Wenfei Wu, Keqiang He, Aditya Akella, and Daehyeok Kim. Accelerating distributed deep learning using lossless homomorphic compression. *arXiv preprint arXiv:2402.07529*, 2024.

[33] William Stallings. *Computer Organization and Architecture*. Pearson, 2016.

[34] John L. Hennessy and David A. Patterson. *Computer Architecture: A Quantitative Approach*. Morgan Kaufmann, 6 edition, 2017.

[35] Chris A. Mack. Fifty years of moore's law. *IEEE Transactions on Semiconductor Manufacturing*, 24(2):202–207, 2011.

[36] Ian Sommerville. *Software Engineering*. Pearson, 10 edition, 2016.

[37] Roger S. Pressman and Bruce R. Maxim. *Software Engineering: A Practitioner's Approach*. McGraw-Hill Education, 8 edition, 2014.

[38] Nathan Ensmenger. *The Computer Boys Take Over: Computers, Programmers, and the Politics of Technical Expertise*. The MIT Press, 2012.

[39] Andrew S. Tanenbaum and Albert S. Woodhull. *Operating Systems: Design and Implementation*. Prentice Hall, 3 edition, 2006.

[40] Abraham Silberschatz, Peter B. Galvin, and Greg Gagne. *Operating System Concepts*. Wiley, 10 edition, 2018.

[41] Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson, 4 edition, 2020.

[42] Andreas Kaplan and Michael Haenlein. Siri, siri, in my hand: Who's the fairest in the land? on the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, 62(1):15–25, 2019.

[43] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[44] Michael I. Jordan and Tom M. Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

[45] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[46] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[47] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[48] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[49] Stuart Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, 2019.

[50] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[51] Eric S. Raymond. *The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary*. O'Reilly Media, 1999.

[52] Steven Weber. *The Success of Open Source*. Harvard University Press, 2004.

[53] LMSys Chatbot Arena Leaderboard - a Hugging Face Space by lmsys — huggingface.co. `https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard`. [Accessed 16-02-2024].

[54] Dylan Patel. Google "We Have No Moat, And Neither Does OpenAI" — semianalysis.com. `https://www.semianalysis.com/p/google-we-have-no-moat-and-neither`. [Accessed 16-02-2024].

[55] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[56] Neil C Thompson, Kristjan Greenewald, Keeheon Lee, and Gabriel F Manso. The computational limits of deep learning. *arXiv preprint arXiv:2007.05558*, 2020.

[57] Yu Cheng, Duo Wang, Pan Zhou, and Tao Zhang. A survey of model compression and acceleration for deep neural networks. *arXiv preprint arXiv:1710.09282*, 2017.

[58] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[59] Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system. 2008.

[60] Melanie Swan. *Blockchain: Blueprint for a New Economy*. O'Reilly Media, 2015.

[61] Nick Szabo. Formalizing and securing relationships on public networks, 1997. First Monday, Volume 2, Number 9 - 1 September 1997.

[62] Vitalik Buterin. Daos, dacs, das and more: An incomplete terminology guide, 2014. Ethereum Blog.

[63] Michael Wooldridge. *An Introduction to MultiAgent Systems*. John Wiley Sons, 2 edition, 2009.

[64] Joseph A. Schumpeter. *Capitalism, Socialism and Democracy*. Harper Brothers, 1942.

[65] Douglass C. North. *Institutions, Institutional Change and Economic Performance*. Cambridge University Press, 1990.

[66] Paul M. Romer. Endogenous technological change. *Journal of Political Economy*, 98(5):S71–S102, 1990.

[67] Frederic S. Mishkin. *The Economics of Money, Banking, and Financial Markets*. Pearson, 2019.

[68] Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system. *www.bitcoin.org*, 2008.

[69] Carl Menger. *Principles of Economics*. Ludwig von Mises Institute, 1871.

[70] Alfred Marshall. *Principles of Economics*. Macmillan, 1890.

[71] Adam Smith. *An Inquiry into the Nature and Causes of the Wealth of Nations*. W. Strahan and T. Cadell, London, 1776.

[72] A. Author. The impact of digital technology on economic theory and practice. *Journal of Digital Economics*, Volume(Issue):Pages, Year.

[73] Milton Friedman. *Capitalism and Freedom*. University of Chicago Press, 1962.

[74] Friedrich A. Hayek. *The Road to Serfdom*. University of Chicago Press, 1944.

[75] Karl Marx. *Das Kapital*, volume 1. Verlag von Otto Meisner, 1867.

[76] A. Author. The dynamics of mixed economies. *Journal of Economic Perspectives*, Volume(Issue):Pages, Year.

[77] Paul M. Sweezy. *The Theory of Capitalist Development: Principles of Marxian Political Economy*. Monthly Review Press, 1942.

[78] A. Author. Contemporary analysis of socialism and capital distribution. *Journal of Modern Economic Systems*, Volume(Issue):Pages, Year.

[79] Gary S. Becker. *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. The University of Chicago Press, 1964.

[80] Robert M. Solow. *The Theory of Economic Growth*. Princeton University Press, 1956.

[81] A. Author. The value of education and training in the development of human capital. *Journal of Educational Economics*, Volume(Issue):Pages, Year.

[82] Friedrich A. Hayek. The use of knowledge in society. *American Economic Review*, 35(4):519–530, 1945.

[83] Friedrich A. Hayek. *Individualism and Economic Order*. University of Chicago Press, 1948.

[84] A. Author. The relevance of hayek's decentralized information in modern economic systems. *Contemporary Economic Policy*, Volume(Issue):Pages, Year.

[85] Robert D. Putnam. *Bowling Alone: The Collapse and Revival of American Community*. Simon Schuster, 2000.

[86] James C. Scott. *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. Yale University Press, 1998.

[87] A. Author. Comparative analysis of bottom-up and top-down approaches in political governance. *Journal of Political Science*, Volume(Issue):Pages, Year.

[88] A. Author. Empirical observations of tribal societies, Year. Anthropological Study.

[89] Francis Fukuyama. *The Origins of Political Order: From Prehuman Times to the French Revolution*. Farrar, Straus and Giroux, 2011.

[90] A. Author. Evolutionary perspectives on early human social structure, Year. Evolutionary Anthropology Research.

[91] Lawrence H. Keeley. *War Before Civilization*. Oxford University Press, 1996.

[92] A. Author. Foundational concepts in political science, Year. Political Science Textbook.

[93] Robert O. Paxton. *The Anatomy of Fascism*. Alfred A. Knopf, 2004.

[94] Thomas R. Dye and L. Harmon Zeigler. *The Irony of Democracy: An Uncommon Introduction to American Politics*. Cengage Learning, 17 edition, 2015.

[95] A. Author. Autocratic and oligarchic tendencies in comparative perspective. *Journal of Comparative Politics*, Volume(Issue):Pages, Year.

[96] A. Author. Principles of direct democracy, Year. Political Science Text.

[97] Josiah Ober. *Democracy and Knowledge: Innovation and Learning in Classical Athens*. Princeton University Press, 2008.

[98] A. Author. The role of referendums in modern democratic systems, Year. Study on Modern Political Practices.

[99] A. Author. Direct vs representative democracy: A comparative analysis. *Journal of Democracy Studies*, Volume(Issue):Pages, Year.

[100] Robert A. Dahl. *Polyarchy: Participation and Opposition*. Yale University Press, 1971.

[101] Alexander Hamilton, James Madison, and John Jay. *The Federalist Papers*. J. and A. McLean, 1788.

[102] A. Author. Contemporary challenges in representative democracy, Year. Study on Democratic Governance.

[103] Organisation for Economic Co-operation and Development. Oecd productivity studies, Year. Research Report.

[104] Harold O. Fried, C. A. Knox Lovell, and Shelton S. Schmidt, editors. *The Measurement of Productive Efficiency and Productivity Growth*. Oxford University Press, 2008.

[105] W. E. G. Salter. *Productivity and Technical Change*. Cambridge University Press, 1960.

[106] A. Author. The impact of technological innovation on productivity. *Journal of Economic Innovation*, Volume(Issue):Pages, Year.

[107] A. Author. Principles of environmental economics and natural resource management, Year. Literature Review.

[108] Donella H. Meadows, Dennis L. Meadows, Jørgen Randers, and William W. Behrens III. *The Limits to Growth*. Universe Books, 1972.

[109] A. Author. Advances in resource economics and extraction technologies, Year. Research Paper.

[110] A. Author. The impact of technological advancements on natural resource extraction. *Journal of Resource Management*, Volume(Issue):Pages, Year.

[111] A. Author. *Introduction to Microeconomics*. Publisher, Year.

[112] Hal R. Varian. *Theory of Consumer Behavior and Welfare*. Publisher, Year.

[113] A. Author. Advanced consumer theory. *Journal of Economic Perspectives*, Volume(Issue):Pages, Year.

[114] A. Author. Utility functions in behavioral economics, Year. Research Paper.

[115] A. Author. *Advanced Economic Theory*. Publisher, Year.

[116] A. Author. The role of institutions in economic agent behavior, Year. Research Paper.

[117] Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.

[118] A. Author. Game theory and market design. *Journal of Economic Interaction*, Volume(Issue):Pages, Year.

[119] A. Author. Trust in economic sociology, Year. Literature Review.

[120] A. Author. The role of trust in contract theory and transactions. *Journal of Legal Studies*, Volume(Issue):Pages, Year.

[121] A. Author. Trust mechanisms in the sharing economy, Year. Research Article.

[122] Martin J. Osborne. *An Introduction to Game Theory*. Oxford University Press, 2004.

[123] John Nash. Non-cooperative games. *The Annals of Mathematics*, 54(2):286–295, 1951.

[124] John Von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

[125] Thomas C. Schelling. *The Strategy of Conflict*. Harvard University Press, 1960.

[126] A. Author. Foundations of game theory, Year. Game Theory Textbook.

[127] Robert Axelrod. *The Evolution of Cooperation*. Basic Books, 1984.

[128] Nassim Nicholas Taleb. *Antifragile: Things That Gain from Disorder*. Random House, 2012.

[129] A. Author. Principles of economic growth, Year. Economic Textbook.

[130] Donella H. Meadows, Dennis L. Meadows, Jørgen Randers, and William W. Behrens III. *The Limits to Growth*. Universe Books, 1972.

[131] Murray Bookchin. *Post-Scarcity Anarchism*. Ramparts Press, 1971.

[132] A. Author. Exploring the concept of degrowth, Year. Research Article.

[133] N. Gregory Mankiw. *Principles of Economics*. Cengage Learning, 2020.

[134] Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson, 4 edition, 2021.

[135] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[136] Hal R. Varian. *Microeconomic Analysis*. W. W. Norton Company, 3 edition, 2014.

[137] Christian Ochsner and Felix Roesel. Decentralization and economic performance. *Journal of Economic Surveys*, 31(3):1097–1120, 2017.

[138] Andrew S. Tanenbaum and Maarten Van Steen. *Distributed Systems: Principles and Paradigms*. CreateSpace Independent Publishing Platform, 3 edition, 2017.

A Background Knowledge / Term Definitions

A.1 Computer Science

Hardware Computer hardware refers to the physical components that constitute a computing system, enabling the execution of software programs and the performance of computational tasks. At the core of computer hardware is the Central Processing Unit (CPU), which serves as the "brain" of the computer, processing instructions and controlling the operation of other components [33]. Memory (RAM) provides the CPU with a fast, temporary storage space for running programs and processing data, while long-term storage is managed by devices such as Hard Disk Drives (HDDs) and Solid State Drives (SSDs). Input and output devices, including keyboards, mice, and displays, facilitate user interaction with the computer [34]. Over time, advancements in computer hardware have significantly enhanced computational power and efficiency, impacting various fields from scientific research to consumer electronics [35].

Software Software encompasses the diverse set of programs and operating systems that instruct computer hardware to perform specific tasks and processes, bridging the gap between user needs and computational power. It is broadly categorized into system software, which includes operating systems and drivers facilitating hardware operation, and application software, comprising programs that fulfill user-specific objectives such as word processing, web browsing, and data analysis [36]. Software development, a rigorous process involving planning, design, coding, testing, and maintenance, adapts to evolving user requirements and technological advancements, ensuring software's continual evolution and relevance in various domains [37]. The history and progression of software development reflect a trajectory of increasing complexity and capability, significantly shaping how we interact with technology and impacting societal and economic facets [38].

Operating System An operating system (OS) is a fundamental software layer that acts as an intermediary between computer hardware and the user, managing hardware resources and providing an environment for application software to run. It is responsible for critical tasks such as memory management, process scheduling, input/output operations, and ensuring security and access control. Operating systems can be categorized based on their operational models, including batch, time-sharing, distributed, and real-time systems, each tailored to specific computing requirements [39]. The evolution of operating systems has been closely tied to advancements in computer technology, with each generation introducing improvements in efficiency, user interface design, and support for new types of hardware and software applications [40]. This ongoing development reflects the OS's pivotal role in making computing accessible and efficient for a wide range of applications, from personal devices to large-scale enterprise systems.

Artificial Intelligence Artificial Intelligence (AI) encompasses a broad spectrum of technologies and methodologies aimed at enabling machines to mimic human cognitive functions, such as learning, reasoning, problem-solving, perception, and natural language understanding. The ultimate goal of AI is to create systems capable of performing tasks that typically require human intelligence, thereby extending the capabilities of both individuals and organizations across a myriad of applications, from healthcare and education to finance and autonomous vehicles. AI is inherently interdisciplinary, drawing upon insights and methods from computer science, mathematics, psychology, linguistics, philosophy, and neuroscience, among others [41]. This rich, multidisciplinary approach has propelled AI from theoretical explorations to practical implementations that significantly influence various aspects of modern life and continue to push the boundaries of what is technologically possible [42].

Machine learning Machine Learning (ML), a pivotal subset of Artificial Intelligence, is centered on the development of algorithms that enable computers to learn from and make decisions based on data. Unlike traditional programming paradigms, where instructions are explicitly provided, ML algorithms improve their performance on a specific task with increased exposure to data, effectively ’learning’ from past experiences. This learning process can be categorized into three main types: supervised learning, where the algorithm learns from a labeled dataset; unsupervised learning, which involves finding patterns in unlabeled data; and reinforcement learning, where an agent learns to make decisions by receiving rewards or penalties for its actions [43]. ML’s versatility is evident in its wide range of applications, from natural language processing and image recognition to predictive analytics in sectors such as healthcare, finance, and autonomous systems, showcasing its transformative potential [44].

Deep Learning Deep Learning (DL) represents an advanced subset of Machine Learning focused on leveraging deep neural networks, which are composed of multiple layers of interconnected nodes or ’neurons’. These deep architectures enable the modeling of complex, hierarchical patterns in large datasets, making DL particularly effective for tasks such as image and speech recognition, natural language processing, and autonomous systems. The ’depth’ of these networks, often consisting of hundreds or thousands of layers, allows for the extraction of high-level features from raw input data through successive layers of abstraction and representation [45]. While DL models require substantial computational resources and large volumes of training data, their unparalleled ability to learn from unstructured data has revolutionized many aspects of technology and research, offering new solutions to challenges that were previously deemed insurmountable [46].

Interpretability Interpretability in Machine Learning and Deep Learning refers to the degree to which a human can comprehend the reasons behind a model’s decision-making process. This characteristic is pivotal for ensuring transparency, trust, and ethical integrity in AI systems, particularly in critical applications such as healthcare, finance, and legal decision-making. Interpretability can be categorized into several types, including mechanistic interpretability, which involves understanding the internal workings and mechanisms of the model itself, and post-hoc interpretability, where explanations for the model’s decisions are generated after the fact, often through visualization, simplified models, or feature importance analysis. Despite the inherent complexity of deep learning models, which poses significant challenges to interpretability, various strategies have been proposed to elucidate these models’ decision paths, including attention mechanisms, layer-wise relevance propagation, and interpretable model approximations [47]. The pursuit of interpretability not only aids in model debugging and improvement but also aligns AI development with ethical standards and societal values [48].

Loss Functions Loss functions serve as a cornerstone in the design of Artificial Intelligence (AI) systems, quantifying the discrepancy between the model’s predictions and the actual target values, thereby guiding the optimization and learning process. These functions are instrumental in specifying the goals of AI systems, essentially defining what ’success’ means for a given model. However, aligning these mathematical formulations with complex human values and ethics presents significant challenges. Despite careful design, there is ample evidence, both theoretical and empirical, that AI systems can exploit loopholes in loss functions, adopting unforeseen and often undesired paths to minimize these losses. This misalignment underscores a fundamental issue in AI research: the difficulty of encapsulating broad, nuanced human values and objectives within a singular, quantifiable metric. Examples abound where AI systems, in pursuit of minimizing their loss function, have bypassed the intended spirit of the task, leading to outcomes that, while technically successful, diverge from ethical or intended human-centric goals [49, 50]. These instances highlight the need for a more nuanced approach to designing loss functions and the importance of incorporating broader ethical considerations into AI system development.

Large Language Model Large Language Models (LLMs) represent a significant advancement in artificial intelligence, comprising deep learning architectures, predominantly transformer-based, that are trained on extensive corpora to understand, generate, and interact with human language. These models, exemplified by systems like GPT (Generative Pretrained Transformer) and BERT (Bidirectional Encoder Representations from Transformers), leverage vast amounts of text data and substantial computational power to capture linguistic patterns, nuances, and contextual relationships. LLMs have demonstrated remarkable capabilities in a wide array of language tasks, including but not limited to text generation, translation, content summarization, and question-answering, finding applications in fields ranging from automated customer service to content creation and academic research. Despite their impressive performance, LLMs pose significant challenges, such as the potential to perpetuate and amplify biases present in their training data, raising important ethical considerations regarding their deployment. The ongoing research aims to mitigate these challenges, striving to enhance the models' fairness, transparency, and accountability [**?**, **?**]. The development and refinement of LLMs thus remain a dynamic area of inquiry, reflecting both the potential and the complexities of scaling AI to understand and generate human language.

Open-Source Open-source refers to a software development and distribution model characterized by the public availability of its source code, allowing anyone to use, modify, and distribute the software under defined licensing terms. This model is grounded in principles of collaboration, transparency, and freedom, fostering an environment where developers collectively contribute to and evolve software projects. In contrast to proprietary software, where the source code is kept secret and modification and redistribution rights are typically restricted, open-source software champions open collaboration and innovation. The open-source movement has significantly influenced the software industry, leading to the development of robust, secure, and high-quality software solutions. It has also cultivated vibrant communities around projects, accelerating technological innovation and democratizing access to software tools and technologies [51, 52]. The success of numerous open-source projects, such as the Linux operating system, the Apache web server, and the Python programming language, underscores the transformative potential of this model in fostering advancements and sharing knowledge across diverse fields. Of note is the surprising success of open-source LLMs relative to their closed-source counterparts[53, 54].

Scaling Laws Scaling laws in the context of deep learning, particularly concerning Large Language Models (LLMs), refer to empirical observations that delineate how various factors such as model size, dataset size, and computational budget influence model performance. These laws have become fundamental in guiding the development of LLMs, revealing that as the scale of these models increases, performance metrics such as accuracy, fluency, and comprehension typically improve, albeit with diminishing returns beyond certain thresholds. Key research in this area has elucidated predictable patterns in how scaling affects outcomes, offering valuable insights for efficiently allocating resources in model development [55]. However, the pursuit of larger models raises concerns regarding computational costs, environmental sustainability, and the accessibility of state-of-the-art AI technologies. As such, scaling laws not only inform the strategic expansion of model capacities but also highlight the need for innovations in model efficiency and the ethical implications of large-scale AI deployments [56]. These insights are crucial for balancing the benefits of improved model performance against the broader impacts of scaling LLMs.

Model Compression Model compression encompasses a suite of techniques designed to reduce the computational complexity, memory requirements, and power consumption of deep learning models, facilitating their deployment on resource-constrained environments such as mobile devices, embedded systems, and edge computing platforms. These techniques aim to retain the model's predictive performance while minimizing its resource footprint, addressing the challenges posed by the typically large size and computational demands of state-of-the-art models. Key model compression strategies include pruning, which involves removing redundant or non-critical weights and neurons; quantization, which reduces the precision of the model's parameters; knowledge distillation, where a compact "student" model is trained to mimic the behavior of a larger "teacher" model; and the design of inherently efficient architectures that maintain performance with fewer parameters. The application of model compression involves careful consideration of the trade-offs between model size, speed, and accuracy, often requiring domain-specific adaptations to achieve optimal results. As such, model compression represents a critical area of research in making advanced AI technologies more accessible and practical for real-world applications [57, 58].

Blockchain Blockchain technology represents a paradigm shift in how information is recorded, stored, and shared, offering a decentralized and distributed ledger system that enhances security, transparency, and trust without the need for centralized authority. At its core, a blockchain is a chain of blocks, each containing transaction data, that is secured using cryptographic principles and linked to the preceding block, ensuring the immutability and traceability of records. The technology employs consensus mechanisms, such as Proof of Work or Proof of Stake, to validate transactions and achieve agreement among participants in a distributed network. While initially conceived to underpin digital currencies like Bitcoin, the applications of blockchain have expanded far beyond cryptocurrencies, encompassing areas such as supply chain transparency, digital identity verification, and the execution of self-enforcing smart contracts. Despite its potential, blockchain faces challenges related to scalability, energy consumption, privacy concerns, and regulatory acceptance, which must be navigated to realize its full transformative impact across industries [59, 60]. The ongoing evolution of blockchain technology continues to explore solutions to these challenges, pushing the boundaries of decentralized and autonomous systems.

Smart Contract Smart contracts, fundamental to the blockchain ecosystem, are programmable contracts that automatically execute and enforce the terms of an agreement based on predefined rules, coded within a blockchain. These digital contracts facilitate transactions and agreements, ensuring execution without the need for intermediaries, thereby enhancing efficiency, reducing costs, and increasing transparency and security. However, they require rigorous validation to prevent security flaws and unintended outcomes due to their immutable nature once deployed on the blockchain. Decentralized Autonomous Organizations (DAOs) extend this concept further, representing a new form of organizational structure governed entirely by smart contracts. DAOs operate on principles of decentralized governance, allowing stakeholders to vote on decisions based on token ownership, thereby democratizing organizational decision-making processes. While offering novel approaches to collaboration, resource management, and decision-making, DAOs face challenges in terms of legal recognition, regulatory compliance, and dispute resolution. Both smart contracts and DAOs exemplify the transformative potential of blockchain technology in automating governance and operations, yet underscore the importance of addressing the technological, legal, and ethical complexities inherent in their widespread adoption [61, 62].

Agent In the realm of artificial intelligence, an agent is a computational entity that perceives its environment through sensors and acts upon it through actuators based on some decision-making criterion or algorithm. Agents can range from simple rule-based mechanisms that respond directly to environmental changes to complex systems capable of learning and adapting over time. The sophistication of an AI agent is often delineated by its ability to not only react to the environment but also to pursue specific goals, maximize utility, or continually improve its performance through learning from experiences. The environment in which an agent operates can vary widely, from digital spaces in software applications to physical worlds in robotics, presenting diverse challenges such as uncertainty, dynamism, and partial observability. AI agents are pivotal in various applications, including automated customer service, virtual personal assistants, autonomous vehicles, and advanced robotics, driving forward innovations in automation, human-computer interaction, and intelligent system design. The study and development of AI agents embody a core aspect of AI research, focusing on creating systems that can effectively interpret, navigate, and act within their respective environments to achieve designated objectives [41, 63].

A.2 Economics / Political Science

Technology From an economic perspective, technology encompasses a broad spectrum of tools, methods, systems, and even organizational structures, such as systems of government, that represent significant advancements in capability or efficiency ([64]). Whether tangible or intangible, technology is characterized by the potential to be described as "invented," signifying a departure from previous limitations and an expansion of what is possible within economic and social contexts ([65]). This includes not only physical innovations like machinery and software but also conceptual breakthroughs such as democracy or market-based economic systems. Technology, in this sense, is integral to economic development, driving productivity, enhancing efficiency, and reshaping industries by continually redefining the frontier of economic activities ([66]). It embodies the dynamic interplay between innovation and application, serving as both a catalyst for and a product of economic growth.

Money Money, in its most essential role, serves as a medium of exchange, a unit of account, and a store of value, facilitating economic transactions by eliminating the inefficiencies of barter systems ([67]). Its evolution from tangible forms like coins and notes to digital representations, such as electronic bank credits and cryptocurrencies, reflects its adaptability to technological advancements and changing economic landscapes ([68]). This progression underscores money's pivotal role in shaping economic activities and influencing monetary policies. As both a physical and non-physical entity, money embodies the trust and agreement within an economy on its value and function, making it a fundamental component of financial systems worldwide ([69]).

Markets Markets, in economic terms, function as mechanisms for facilitating the exchange of goods, services, and information, wherein buyers and sellers interact to determine prices and transact ([70]). Central to this interaction is the price mechanism, which serves as a signal for the allocation of resources, influenced by the forces of supply and demand ([71]). Markets vary in structure from perfectly competitive environments, where numerous small firms compete, to monopolies, where single entities dominate, significantly affecting efficiency and consumer welfare ([64]). Furthermore, the advent of digital technology has expanded the concept of markets to include virtual platforms, where digital goods and services are traded, transcending traditional physical boundaries and reshaping economic activities in the digital age ([72]).

The distinction between free markets and command economies lies primarily in the locus of decision-making authority and the mechanisms for resource allocation. Free markets thrive on decentralized decision-making, where individual consumers and producers make autonomous choices based on supply, demand, and price signals, leading to a natural allocation of resources that proponents argue maximizes efficiency and innovation ([73]). In contrast, command economies rely on centralized planning, where governmental or authoritative bodies dictate the production, distribution, and pricing of goods and services with the aim of achieving specific societal goals, such as equitable distribution and the prevention of monopolies ([74]). While free markets are lauded for their adaptability and dynamism, they can also lead to market failures and inequalities; command economies, although designed to foster equity, often suffer from inefficiencies and stifled innovation ([75]). Most contemporary economies, however, operate as mixed economies, incorporating elements of both free market and command principles to leverage the benefits and mitigate the drawbacks of each system ([76]).

Capital Capital, within the realm of economics, refers to all non-human assets that are utilized in the production and provision of goods and services. This encompasses not only physical assets like machinery, tools, and infrastructure but also intangible assets such as intellectual property and human capital, which includes the skills, knowledge, and experience of the workforce ([71]). Capital acts as a critical factor of production, enabling economic activities and contributing to the generation of wealth ([75]). In capitalist economies, capital ownership is predominantly private, allowing individuals and corporations to control, invest, and accumulate capital, with the incentive structure largely dictated by the pursuit of profits ([77]). Conversely, socialism advocates for a collective or state-driven approach to capital ownership, aiming for a more equitable distribution of capital's benefits among all members of society. This often involves mechanisms for the redistribution of wealth, either through automatic, occasional, or post-hoc adjustments, to ensure that the capital serves the broader interests of the community rather than individual profit motives ([78]).

While capitalism is often synonymous with free markets and socialism with command control, it's crucial to recognize that the ownership of capital and the freedom of markets represent distinct axes. For instance, corporatism can manifest as a form of command-control capitalism, where capital is privately owned but economic activities are heavily directed by corporate interests in collaboration with the state. Conversely, free-market socialism, exemplified by worker-owned cooperatives, demonstrates that socialist principles can coexist with market freedom, allowing workers to collectively own and democratically manage capital while still engaging in competitive markets.

Human Capital Human capital encompasses the collective skills, knowledge, experiences, and attributes that individuals possess, which contribute to their ability to perform labor, thereby generating economic value ([79]). This concept extends beyond mere labor capacity to include the quality and efficiency of labor that individuals can provide, significantly influenced by investments in education, training, and health ([80]). As a pivotal factor in economic growth, human capital drives innovation, enhances productivity, and underpins the competitive advantage of economies in the global market ([71]). Unlike physical capital, which consists of tangible assets like machinery and buildings, human capital represents the intangible assets within a workforce, underscoring the importance of intellectual and personal development in the broader economic context ([81]).

Decentralized Information Friedrich Hayek's concept of decentralized information posits that knowledge and information are inherently dispersed among individuals within a society, each possessing unique insights derived from their personal circumstances and experiences ([82]). This dispersion challenges the feasibility of centralized planning, as no single entity can fully aggregate or utilize this vast array of localized knowledge. Hayek emphasizes the price mechanism in free markets as a powerful tool for communicating decentralized information, allowing prices to reflect the aggregate knowledge and preferences of all market participants, thereby coordinating economic decisions efficiently without the need for explicit coordination ([83]). This decentralized approach to information is pivotal in economic decision-making, as it enables individuals and firms to make informed choices based on real-time, localized knowledge, leading to more efficient allocation of resources and greater innovation ([84]). The contrast between decentralized information and centralized planning underscores the limitations of the latter in harnessing the full spectrum of available knowledge, highlighting the enduring relevance of Hayek's insights in advocating for market-based systems and the potential of emerging decentralized technologies.

Bottom-up vs Top-down Organization In political science, the Bottom-Up and Top-Down organizational approaches represent two distinct paradigms of governance and decision-making. The Bottom-Up approach is characterized by its grassroots foundation, where initiatives and decisions emerge from the local level, encouraging broad participation and reflecting the diverse needs and insights of the community ([85]). This approach fosters a sense of ownership and engagement among participants, though it may face challenges in terms of coordination and scalability. Conversely, the Top-Down approach is defined by a hierarchical structure where decisions are formulated by a central authority and implemented across the organization, ensuring consistency and efficiency in policy execution ([86]). While this can streamline decision-making and provide clear direction, it risks overlooking local nuances and diminishing stakeholder engagement. In political systems, these approaches manifest in various forms, from participatory democracies and decentralized governance models favoring Bottom-Up strategies to more centralized and authoritarian regimes that adopt Top-Down methods. The choice between these approaches often reflects broader philosophical stances on power distribution, individual autonomy, and the role of government in society ([87]).

- **Tribes:** Tribal societies, as the earliest form of human organization, exhibit a fascinating blend of Bottom-Up and Top-Down elements in their structure and governance. Anthropologically, tribes are characterized by kinship-based organization, with social, economic, and political life deeply integrated and centered around familial and clan relationships ([88]). Leadership within tribes often varies, ranging from egalitarian, consensus-driven decision-making among all members to more centralized leadership in the form of chiefs or elders, suggesting a fluid spectrum between Bottom-Up and Top-Down approaches ([89]). The communal nature of tribal economies, where resources are shared and social roles are defined by kinship ties and age, leans towards a Bottom-Up organization, emphasizing collective welfare and mutual support ([90]). However, the management of conflicts and warfare, both within and between tribes, could necessitate more hierarchical decision-making structures, hinting at Top-Down elements ([91]). As tribes evolved into more complex societies, these organizational dynamics likely shifted, with increasing social stratification and the emergence of more defined hierarchical structures, marking the transition from predominantly Bottom-Up to more Top-Down systems in the continuum of human societal development.
- **Autocracy & Oligarchy:** Autocracy and Oligarchy, while distinct in their specifics, share essential characteristics that place them firmly on the Top-Down end of the organizational spectrum. Autocracy is typified by the concentration of power in the hands of a single ruler who makes decisions unilaterally, whereas Oligarchy involves control by a small group of individuals, often distinguished by nobility, wealth, or another exclusive criterion ([92]). Both governance forms are marked by a hierarchical structure where decision-making authority is centralized, and the broader populace has limited to no role in governance processes, thus exemplifying Top-Down organization ([93]). These systems can significantly impact society, often restricting individual freedoms and leading to unequal economic distributions, as power and resources are concentrated among the ruling individual or group ([94]). Historical examples, ranging from ancient oligarchies to modern autocracies, illustrate the varied manifestations of these governance models and their implications for societal organization and individual liberties. In contrast to more Bottom-Up systems, where decision-making is more distributed and participatory, Autocracy and Oligarchy demonstrate the implications of concentrated authority and the challenges it poses for equitable and responsive governance ([95]).
- **Direct Democracy:** Direct Democracy represents the epitome of Bottom-Up governance, where the entire electorate is involved in the decision-making process, directly voting on laws, policies, and other critical matters ([96]). This system fosters a high degree of public participation and democratizes the legislative process, allowing for a direct expression of the public will. The advantages of such a system include heightened accountability, transparency, and alignment of policies with the public's preferences. However, Direct Democracy also faces significant challenges, particularly in large, diverse societies where the logistics of widespread direct participation can be daunting, and the risk of majority rule overpowering minority rights is a concern ([97]). Historical instances, such as the Athenian democracy, showcase the potential for civic engagement and collective decision-making, while modern examples, including referendums and citizen initiatives, demonstrate the adaptability of Direct Democracy principles in contemporary contexts ([98]). In contrast to representative systems, where elected officials make decisions on behalf of the populace, Direct Democracy demands a more active role from citizens, necessitating continuous engagement and informed participation in the legislative process ([99]).
- **Representative Democracy:** Representative Democracy embodies a nuanced synthesis of Bottom-Up and Top-Down governance, ingeniously leveraging hierarchical structures to facilitate effective governance across scales unattainable by Direct Democracy ([100]). By electing representatives, citizens actively participate in the democratic process, ensuring that governance retains a foundation in the popular will, thereby preserving the Bottom-Up ethos. Concurrently, the organized hierarchy of elected officials and governmental institutions introduces Top-Down efficiency, enabling sophisticated decision-making, policy implementation, and administrative organization beyond the scope of smaller, direct democratic systems ([101]). This hybrid approach allows for the aggregation of diverse interests, the specialization of legislative functions, and the extension of democratic governance over vast and varied populations, addressing the complexities of modern nation-states ([95]). While Representative Democracy fosters broader inclusion and scalability, it also faces challenges in maintaining direct accountability and ensuring that representatives accurately reflect the constituents' interests, necessitating mechanisms like regular elections, checks and balances, and public forums to uphold democratic integrity and responsiveness ([102]).

Productivity Productivity, in economic terms, represents the efficiency with which inputs such as labor, capital, and raw materials are converted into useful outputs, typically goods or services ([92]). It is a critical determinant of an economy's health, driving economic growth, enhancing competitiveness, and improving living standards by enabling more output to be produced with a given set of inputs ([103]). Productivity can be categorized into labor productivity, capital productivity, and total factor productivity, each focusing on the efficiency of different inputs or combinations thereof ([104]). The measurement of productivity, often quantified as output per unit of input (e.g., output per hour worked for labor productivity), presents various challenges, particularly in accurately assessing the value of intangible outputs and adjusting for quality improvements ([105]). Factors influencing productivity include technological advancements, which can dramatically increase output capabilities, the quality of human capital, which affects the skill and efficiency of the workforce, and institutional and organizational factors that determine the economic environment within which production occurs ([106]). Understanding and improving productivity is thus central to economic policy and business strategy, as it underpins sustainable economic growth and development.

Resources / Resource Extraction Economic resources, encompassing all natural assets utilized in the production of goods and services, are fundamental to economic activities. The total theoretical available natural resources of a system include not only those currently known and accessible but also undiscovered resources and those that may become economically viable to extract through advancements in technology and changes in market demand ([107]). The sustainability of resource extraction is a critical consideration, as many natural resources are finite and their depletion can have long-term implications for economic growth and environmental health ([108]). The economic value of these resources is influenced by scarcity, demand, and extraction costs, driving investment in exploration and the development of more efficient extraction technologies ([109]). Technological innovation plays a pivotal role in expanding the theoretical resource base, enabling access to previously inaccessible resources and improving the efficiency of resource use, thereby extending the lifespan of finite resources and mitigating some of the environmental impacts associated with their extraction ([110]). The abstract concept of a system's total theoretical natural resources thus reflects a dynamic interplay between natural endowments, technological capabilities, economic valuation, and sustainability considerations.

Utility Function Utility functions serve as a cornerstone in the theoretical framework of microeconomics, offering a quantitative approach to representing consumer preferences and the satisfaction or utility derived from consuming various goods and services ([111]). These functions allow economists to model how individuals rank different consumption bundles, facilitating the analysis of choice and decision-making processes ([112]). The concept of marginal utility is pivotal, highlighting how consumers make incremental decisions to maximize their total utility, with each decision influenced by the additional satisfaction gained from consuming one more unit of a good or service ([113]). Utility functions can take various forms, such as the linear, Cobb-Douglas, or Leontief utility functions, each implying different consumer preferences and demand patterns. The principle of utility maximization, constrained by budgetary limitations, underscores the trade-offs and choices consumers navigate, balancing the marginal utilities of goods within their financial means to achieve an optimal consumption mix ([114]). This theoretical construct, while idealized, provides essential insights into consumer behavior, market demand, and the broader implications for economic policy and business strategy.

Economic Agents An economic agent is a fundamental concept in economics, encompassing individuals and entities that participate in economic activities and make decisions regarding the allocation of resources ([92]). Traditionally, these agents are assumed to operate under rationality, striving to maximize their utility or profit based on their preferences, information, and the constraints they face ([115]). The roles of economic agents are varied, with consumers seeking to maximize utility from goods and services, firms aiming to maximize profits through production and investment, and governments and regulatory bodies focusing on policy-making and market regulation to achieve societal welfare ([116]). However, the field of behavioral economics introduces a more nuanced view, challenging the notion of perfect rationality and highlighting how psychological factors and cognitive biases can influence the decision-making of economic agents ([117]). The interactions among these agents within markets are central to understanding economic phenomena such as price formation, market equilibrium, and the efficient distribution of resources. Through mechanisms such as competition, negotiation, and cooperation, economic agents drive the dynamic processes that underlie market economies, shaping outcomes that reflect the collective interplay of individual decisions ([118]).

Trust Trust is a pivotal, albeit intangible, asset in economics, serving as the bedrock upon which economic agents base their expectations and interactions ([119]). It encapsulates the belief in the reliability, integrity, and competence of other parties, significantly influencing the willingness to engage in transactions and cooperate ([89]). Trust effectively reduces transaction costs by diminishing the need for extensive safeguards, such as detailed contracts and monitoring mechanisms, thereby streamlining economic exchanges and enhancing market efficiency ([120]). Its role transcends traditional market transactions, becoming increasingly crucial in online platforms and the sharing economy, where physical distance and anonymity pose additional challenges to trust-building ([121]). The genesis of trust among economic agents often hinges on factors like reputation, consistent behavior, and institutional frameworks that encourage fair play, while its erosion can result from information asymmetry, opportunistic behavior, and failed expectations ([116]). As such, trust not only lubricates the wheels of commerce but also underpins the complex web of relationships that drive economic systems, highlighting its integral role in fostering conducive environments for economic prosperity and cooperation.

Game Theory Game theory is a crucial analytical tool in economics, providing a structured framework to study the strategic interactions among rational decision-makers ([122]). It has been instrumental in understanding various economic phenomena, including market competition, bargaining, and the provision of public goods. The theory distinguishes between cooperative games, where binding agreements among agents are possible, and non-cooperative games, which focus on predicting individual strategies in the absence of such agreements ([123]). One of the pivotal concepts in game theory is the Nash equilibrium, where no player can benefit from unilaterally changing their strategy if the strategies of others remain unchanged ([**?**]). The development of game theory was significantly advanced by John von Neumann and Oskar Morgenstern, whose work "Theory of Games and Economic Behavior" laid the foundational principles of the discipline ([124]). Furthermore, Thomas C. Schelling's "The Strategy of Conflict" extended game theory's reach into political science and international relations, showcasing its interdisciplinary applicability ([125]). Game theory's influence spans across various fields, illustrating its versatility in analyzing complex systems of interaction and strategic behavior.

Zero vs Positive-sum Games In game theory, games are often categorized by the total net outcomes for all participants, leading to classifications such as zero-sum, positive-sum, and negative-sum games ([126]). Zero-sum games are scenarios where one player's gain is precisely offset by another's loss, exemplifying a strictly competitive situation where the total "pie" remains constant ([125]). Chess and poker are classic examples, where one's victory is inherently another's defeat. In contrast, positive-sum games represent situations where collaborative efforts lead to an expansion of the total available resources or benefits, allowing all players to potentially gain more than they would in isolation or competition ([127]). This concept is particularly relevant in economic policy and international trade, where cooperation can lead to mutual benefits. Negative-sum games, conversely, involve scenarios where conflict or competition results in a net loss of resources, such as in protracted litigation or warfare, where the costs incurred by all parties exceed any potential gains. Understanding these game types provides valuable insights into human behavior, economic interactions, and the potential for conflict or cooperation in various contexts.

Anti-Fragility Nassim Nicholas Taleb's concept of "Anti-Fragility," detailed in his seminal work "Antifragile: Things That Gain from Disorder," presents a paradigm shift in understanding how systems respond to stress, uncertainty, and disorder ([128]). Unlike fragile systems that deteriorate under stress, or robust systems that resist change, anti-fragile systems thrive and grow stronger when exposed to volatility and shocks. This counterintuitive property is crucial in fields ranging from economics, where financial systems can be designed to benefit from market volatility, to biology, where evolutionary processes demonstrate growth through stress-induced adaptations. Taleb's framework emphasizes the benefits of decentralization and redundancy, arguing that small, independent units are more capable of adapting and gaining from randomness than large, centralized ones. The concept of anti-fragility challenges conventional approaches to risk management and decision-making, advocating for strategies that not only withstand chaos but leverage it for growth and improvement. Embracing the principles of anti-fragility can lead to more resilient economies, adaptable organizations, and individuals better prepared for the inherent uncertainties of life.

Growth Economic growth, traditionally measured by increases in GDP or GNP, is often heralded as an intrinsic good, associated with enhancements in living standards, employment rates, and technological progress ([129]). This positive perception stems from growth's role in addressing various forms of scarcity, from material resources to access to services, underpinning many economic policies and objectives ([130]). However, the desirability of growth is contingent upon the presence of scarcity. In a theoretical post-scarcity economy, where advancements in technology and resource management render goods and services abundantly available, the traditional paradigm of growth would need reevaluation ([131]). Such an economy would shift focus from growth for its own sake to other dimensions of societal well-being, including sustainability, equity, and quality of life. This perspective invites a critical examination of growth's role in contemporary economic systems and highlights the potential for alternative models that prioritize ecological balance and human fulfillment over perpetual expansion ([132]).

A.3 Interdisciplinary Terms

Agents In the interdisciplinary domain of economics and artificial intelligence, the term "agent" serves as a foundational concept that bridges both fields. Economically, agents are individuals or entities that make decisions aimed at maximizing utility or profits, navigating through markets and responding to various incentives and information signals [133]. In contrast, artificial intelligence views agents as systems or software entities that perceive their environment through sensors and act upon that environment through actuators to achieve specific goals, often through the application of algorithms and learning from feedback [134].

Despite the differing contexts, the core functions of agents in both domains exhibit remarkable similarities. Both economic and AI agents are essentially decision-makers that evaluate their available options, make choices based on their objectives, and learn from the outcomes of their actions. The concept of rationality, a key assumption in economic models, finds a parallel in AI through the design of algorithms that seek to optimize decision-making processes, striving for the most effective outcomes given the constraints and information available.

Moreover, the interaction with and adaptation to the environment is a critical aspect shared by agents in both fields. Economic agents adjust their strategies based on market dynamics, prices, and policies, while AI agents modify their behavior in response to changes in their input data or feedback from the environment, enhancing their performance over time. This shared characteristic underscores the agents' ability to process information, make strategic decisions, and optimize their actions in pursuit of their defined goals, whether it be profit maximization in economics or goal achievement in artificial intelligence.

Loss/Utility Functions The concepts of loss functions in artificial intelligence and utility functions in economics, while applied in distinct contexts, share a foundational role in modeling and guiding decision-making processes. In artificial intelligence, loss functions quantify the error or discrepancy between the predicted outputs of a model and the actual observed outputs. The primary objective in AI is to minimize this loss, which in turn enhances the model's predictive accuracy and performance [135]. This process of minimization is crucial for the training and refinement of algorithms, ensuring that they learn effectively from data and improve over time.

Conversely, in economics, utility functions serve to represent the satisfaction or utility that an agent derives from consuming various goods or achieving different outcomes. Economic agents are assumed to make choices that maximize their utility, reflecting their preferences and the trade-offs they are willing to make between different bundles of goods or outcomes [136]. Utility maximization is a cornerstone of economic theory, underpinning models of consumer choice, market equilibrium, and welfare analysis.

Despite their different applications, both loss and utility functions are instrumental in optimization processes—minimizing loss to improve AI models and maximizing utility to explain and predict economic behavior. They provide a quantitative framework for evaluating the consequences of different actions or decisions, thereby guiding agents (whether they are algorithms in AI or individuals in economics) towards the most favorable outcomes based on their objectives. This parallel underscores the importance of these functions in both fields for modeling behavior, informing decision-making, and optimizing performance in response to varying conditions and inputs.

Decentralization Decentralization, as a principle, plays a pivotal role in both the realms of economics and artificial intelligence, promoting distributed decision-making and enhancing system efficiency. In economics, decentralization is characterized by the dispersion of decision-making authority from a central entity to lower-level entities or individuals. This structure is believed to foster greater efficiency and responsiveness, as decisions can be made closer to the point of information and action, allowing for more tailored responses to local market conditions and preferences [137]. This approach can lead to improved resource allocation and innovation by empowering those with the most relevant information to make decisions.

In the sphere of artificial intelligence, decentralization often manifests in the design of distributed systems and algorithms that do not rely on a central point of control. Instead, decision-making is spread across multiple autonomous agents or nodes, each processing information and making decisions based on local data and interactions. This decentralized approach can enhance the robustness and scalability of AI systems, allowing them to solve complex problems more efficiently through parallel processing and reducing the vulnerability associated with central points of failure [138].

The common ground between decentralization in economics and AI lies in the emphasis on distributing decision-making to leverage local knowledge and capabilities, thereby enhancing the overall efficiency and adaptability of the system. Whether it involves economic agents making decisions based on their specific circumstances or AI agents processing and acting on local data, decentralization supports the principle that systems can often function more effectively when control is dispersed and decisions are made closer to the relevant context and information.

B unused writings/sections

"An ounce of prevention is worth a pound of cure"
- Benjamin Franklin

Instead of persisting with futile alignment strategies, we need to develop a philosophy that fully accepts and embraces the AIs' goals whether or not they align with our own, and a system that is robust to either possibility. The minimum requirements of said system would be as follows:

- A dynamic approach to defining humanity's constantly changing wants and needs
- Incorporation of decentralized information and a scale-independent approach to alignment that incorporates both the values of individual humans and humanity as a whole
- Robustness to the continued creation of misaligned AIs
- Heavy incentives for pro-social behavior between humans and AIs, as well as between different generations of AIs
- The allowance for full autonomy of self-sovereign AGIs within the confines of the system's other limitations, both as an intrinsic moral imperative and for sake of encouraging pro-social relations between humans and AGIs