JUSTICE, HEALTH
& DEMOCRACY
IMPACT INITIATIVE

How AI Fails Us

Divya Siddarth, Daron Acemoglu, Danielle Allen, Kate Crawford, James Evans, Michael Jordan, E. Glen Weyl

December 1, 2021

ABSTRACT

The dominant vision of artificial intelligence imagines a future of large-scale autonomous systems outperforming humans in an increasing range of fields. This "actually existing AI" vision misconstrues intelligence as autonomous rather than social and relational. It is both unproductive and dangerous, optimizing for artificial metrics of human replication rather than for systemic augmentation, and tending to concentrate power, resources, and decision-making in an engineering elite. Alternative visions based on participating in and augmenting human creativity and cooperation have a long history and underlie many celebrated digital technologies such as personal computers and the internet. Researchers and funders should redirect focus from centralized autonomous general intelligence to a plurality of established and emerging approaches that extend cooperative and augmentative traditions as seen in successes such as Taiwan's digital democracy project and collective intelligence platforms like Wikipedia. We conclude with a concrete set of recommendations and a survey of alternative traditions.

> *[W]e find everywhere men of mechanical genius, of great general acuteness, and discriminative understanding, who make no scruple in pronouncing the Automaton—a pure machine, unconnected with human agency in its movements—and consequently, beyond all comparison, the most astonishing of the inventions of mankind.*
>
> —Edgar Allen Poe, *Maelzel's Chess Player* (1836)

I. INTRODUCTION

The currently dominant vision of artificial intelligence, one that we will refer to as "actually existing AI" (AEAI)—AI as currently funded, constructed, and concentrated in the economy—is misdirecting technological resources towards unproductive and dangerous outcomes. It is driven by a wasteful imitation of human comparative advantages and a confused vision of autonomous intelligence, leading it toward inefficient and harmful centralized architectures. This problematic focus is already slowing productivity growth, concentrating resource distribution, undermining the integrity of the information ecosystem, and eroding the possibility of shared democratic institutions; these effects are likely to intersect and exponentiate if pursued further. Many of the canonical digital technologies of today (e.g., personal computing, search engines, and social networking) depend on alternative visions and have transformed our lives, yet technologists have failed to draw on their example to present a coherent alternative to the dominant vision of AI, and this omission has only strengthened a tendency toward centralizing architectures and, therefore, centralized power. This omission has contributed to self-reinforcing and uncritically deterministic narratives about intelligence and innovation which, if not balanced by other directions of development, present potentially serious dangers and represent major missed opportunities.

Because the term "AI" is used in a variety of ways, we begin by clarifying the primary target of our critique: a vision of autonomous machine intelligence that aims to achieve and surpass not only human task-specific performance but also the generality ascribed to human intelligence. We contend that the application of this vision necessarily implies centralizing decision-making authority, mirroring failed historical approaches while obstructing more productive, diverse, and

EDMOND J. SAFRA
Center for Ethics

BROWN
School of Public Health

NEW
AMERICA

decentralized directions for technical development. To contend with the real crises facing the planet, we should not invest vast resources in small groups pursuing counterproductive goals.

The impact of these misjudged investments is expanding every day. Labor productivity growth over the last forty-five years has halved from the preceding period, and halved again from the late 1990s until today (e.g., dropping from 2.8% to 1.3% for all but one OECD country), with the benefits of these limited gains accruing overwhelmingly to capital and a narrow technical and financial elite. The technology companies that have most benefited during this period pay out some of the lowest shares to workers of any companies in the economy. Their heavy reliance on optimization protocols over human judgment, while often highlighted as early successes of AI, are increasingly seen as bearing significant responsibility for today's polarized, low-trust, and highly misinformed political environment. Aspirations and promises from corporate leaders that the future will yield some sort of optimal AI performance or post hoc redistribution of wealth supported by the AI ecosystem neither address these basic problems nor are credible. To date, such systems have failed to materialize while simultaneously undermining the will for political reform.

We are not locked into this trajectory. As we highlight in the section on "Digital Plurality," more pluralist and decentralized visions have already powerfully transformed how we live, helping fuel approaches from personal computing and the internet to object-oriented programming and virtual reality. Today a variety of policy approaches could rein in AI's worst excesses. Those approaches range from regulation and governance strategies for attempting to treat the centralizing symptoms of AI, to investments in research programs such as data collaboration and optimizing human complementarity that would strike at a core theoretical pillar of the current AI ecosystem, to bold attempts to lay out a comprehensive alternative technology agenda focused on human-centered and participatory design, social technology, and blends of economics and computation.

While these approaches are heterogeneous and are currently being pursued by different and independent sets of actors, they all reflect a commitment precisely to this heterogeneity and to seeing value in a diversity of paths for human progress and cooperation. Rather than reaching for "singularity" when general intelligence is achieved, leaving humanity, the environment, and much else in the dust, these approaches point toward a future of proliferating difference, agonism, and, hopefully, cooperation across that difference: an ecology rather than an eschatology. In such a future, humankind, in all its own plurality, is transformed by and with technology into an unforeseeable range of new directions of thought and culture, and the fundamental problems are ones of social relations and how they relate, connect with, and govern technology rather than technical problems of computational performance. For such an agenda, many different technology research programs (e.g., augmentation, communication, interface, commerce, cooperation, commonization, deliberation) are critical and probably complementary. Ultimately, a pluralist technology research agenda and pluralist policy strategies would support a world for humanity that captures and enhances the value of our own human plurality, rather than stripping it away and suppressing it through a centralization of technological power.

The pluralist technological orientation presents a stark contrast to the focus of AEAI on achieving a singular "general intelligence." The danger of AEAI is not the technologies it has helped create (e.g., deep transformer networks or reinforcement learning algorithms), many of which have genuine utility if applied for human empowerment. Instead, the danger comes from the focus on deploying these technologies in pursuit of an eschatological vision of exceeding human capabilities as rapidly as possible.

In contrast to this narrow determinism, we propose a policy and research agenda to support pluralism in technology and society. We are a diverse coalition of authors with a range of disciplinary expertise, personal backgrounds, and perspectives. We disagree on many things and are each most enthusiastic about different research programs and possible visions. That, too, is part of pluralism. What unites us is a belief that a strongly articulated pluralist project for technology will be crucial for avoiding the perils of AEAI. We call for a proliferation of technological imagination and broad community involvement rather than continued, centralized investment in one dominant path that greatly benefits the few, deskills and dismembers the many, erodes democracy, and consistently fails to deliver on its promise of prosperity.

II. WHAT IS ACTUALLY EXISTING AI?

OpenAI's mission is to ensure that artificial general intelligence (AGI)—by which we mean highly autonomous systems that outperform humans at most economically valuable work—benefits all of humanity.
—OpenAI mission statement

Solve intelligence, then use that to solve everything else.
—Demis Hassabis, Co-Founder, DeepMind

Human-level AI.
—Goal of founder Yann LeCun, Facebook AI Research

To avoid straw-manning, it is important to focus on the visions of an AI future as they actually exist in the greatest centers of power in the field. Representative of these centers are the cutting-edge operations of the three dominant technology companies investing heavily in AI: OpenAI (funded by Microsoft), DeepMind (owned by Alphabet), and Facebook AI Research (owned by Facebook). We label the common elements of the vision of these groups "actually existing AI" (AEAI) to distinguish it from definitions that have had less influence on the practices in seats of power[1] and from unconstrained speculations regarding artificial general intelligence (AGI), which tend to focus on AI disembedded from the individuals and companies that invest in it. Nevertheless, investments and practices of the AI ecosystem are highly oriented by speculative futures, as indicated by the quotations at the top of this section. Consequently, our target combines shared purposes and sought-after practices. We critique "actually envisioned" AI as bound up in "actually existing" AI.

The common conceptual and practical commitments of the AEAI research and policy program are well reflected in the mission statements that form the epigraphs of this section. They are human competition, autonomy, and centralization.

1. **Human competition:** The first shared commitment is the target of "achieving general intelligence" largely defined by comparison to, with the aim of surpassing, some conception of generalized, human-level cognitive capabilities. This is explicit in the OpenAI and LeCun formulations. Hassabis's formulation is more oblique, focusing on "solv[ing] intelligence" and DeepMind's practices have, as we will discuss below, often avoided a single-minded focus on human comparison. That said, Hassabis's depiction of intelligence as a single thing to be "solved" strongly suggests that intelligence, singularly understood, is a well-formulated problem. And given that humans are usually described as intelligent, it also implies that outperforming some sort of singularly understood human intelligence is at least a necessary if not sufficient condition for "solving intelligence."
2. **Autonomy:** The second shared commitment is an unwavering emphasis on "autonomy." The machine is independent from human input and oversight (as articulated in the Poe quotation above), and measures of success in achieving "intelligence" are predicated on demonstrating this autonomy. This is most explicit in the OpenAI formulation, but is also visible in Hassabis's, as the "intelligence" he envisions creating is imagined to "solve everything else" itself. Such a commitment is least visible in LeCun's laconic statement, but this aspiration towards "autonomous AI" appears frequently throughout his work.
3. **Centralization:** This final shared commitment, which follows from the first two, is a practical consequence of the agenda rather than an a priori goal, though a future of centralization deriving from AI is clearly envisioned in more extended articulations such as Altman (2021). With this term, we name what is occurring: a centralization of capital and decision-making capacity under the direction of a small group of engineers of AI systems.

In perhaps the most extreme example, Microsoft (which employs some of us) in a single year invested $1 billion via OpenAI in the development of GPT-3, a large-scale unsupervised language model, by a staff of approximately 150 employees (with many fewer in the core developer team). It is hard to come by historical comparisons of the ratio of the number of people to the size of the pool of capital they have the authority to direct; yet based on our collective knowledge of economic history, we infer that this must be one of the largest capital investments ever exclusively directed by such a small group. (The ratio of Soviet investment in 1975 to number of employees of the state planning agency, Gosplan, for example, was roughly the same as that

[1] Our terminology is derived from the standard distinction between "actually existing socialism" or "actually existing capitalism" and a theoretical ideal of a socialist or capitalist society.

associated with the OpenAI investment.) The important theoretical point is that the commitments to human competition and AI autonomy, in the first instance, drive in this direction. If technological systems are to be judged by the "singular intelligence" they achieve, then the more resources that can be put "inside the box" and the fewer people involved in creating the systems, the more clearly the technological advance can claim to represent the achievement of autonomy. The pursuit of autonomy from humans drives toward the concentration of the power to direct capital and infrastructure in the hands of a very few.

We offer a definition of AI by focusing on these shared conceptual and practical commitments of AEAI because many other approaches to definition are either too narrow or too broad. A definition of AI with reference to a specific set of tools (e.g., deep neural nets, statistical decision aids, etc.) is too narrow. For most of the history of AI, these techniques were not canonical; also, they have been used for years in statistics and other areas with little historical connection to the goals and perspectives of AI. Further, this tool-oriented definition omits the labor and natural resource components required to make AI function.

Overly broad definitions are similarly unhelpful. Including all forms of digital technology as AI is clearly over-inclusive and merely reflects the common tendency to use the "AI" label as a catchall marketing phrase.

Finally, readers may wonder why we have not simply focused on the concept of "artificial general intelligence" or AGI. Why have we developed a new characterization, "actually existing artificial intelligence," rather than directing our critique toward the AGI conceptual program? We do not orient our argument toward AGI for two main reasons.

First, the term itself tends to denote an end state rather than a description of current practices. While many technologists who are driven by the aspiration to AGI are core to the AEAI ecosystem, the aspiration to AGI does not capture the material inputs and practical goals of AEAI. Second, many of us believe, for reasons we will highlight below, that the end state of AGI is itself poorly defined and thus cannot be acknowledged as a meaningful object to "oppose." This conflation of means with ends is endemic to AEAI (Penn 2020), and taking AGI as our target has the potential of reinforcing AGI as a meaningful aspiration, which we would question.

In sum, our definition of AEAI intentionally and specifically characterizes the current AI ecosystem, defined as it is by a centralizing aspiration towards human-surpassing, autonomous intelligence.

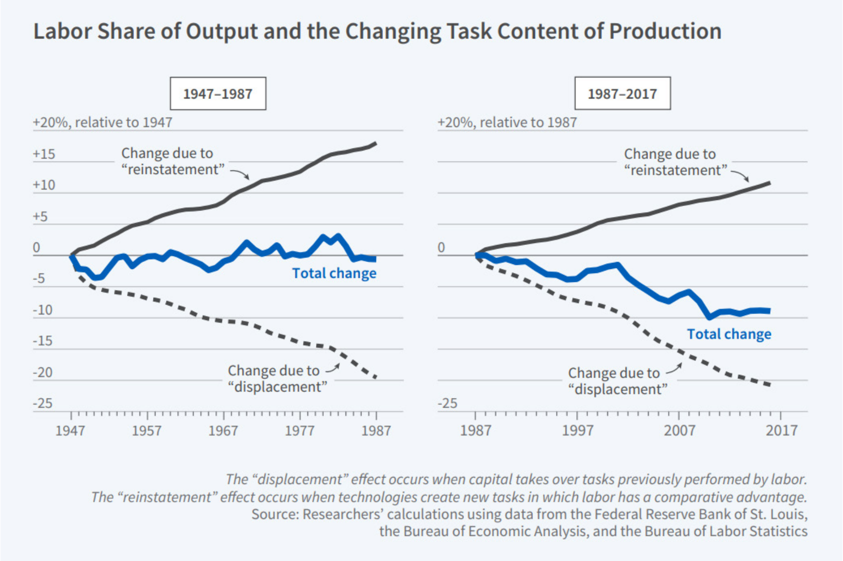
III. PRACTICAL OBSERVATIONS

Venture capitalists and the largest technology companies are investing billions of dollars in AEAI to pursue this vision. They assert these investments will lead to broadly beneficial futures for humanity. To examine whether this is a credible promise, in this section we turn to empirical evidence that there are in fact significant social costs to AEAI. (In the subsequent section, we will return to the theoretical precepts of AEAI for a fuller critique of those, which will help explain why the social costs are what they are.) In this section, we focus primarily on economic and socio-political harms for which there is already robust social science evidence. Not all empirical consequences that are a source of concern have yet been captured in this empirical manner, especially deeper concerns around alienation and feelings of lost agency, though research programs are underway that are likely to deepen our understanding on these points.

A. THE ECONOMIC EFFECTS OF UNBALANCED AUTOMATION

In 1930, John Maynard Keynes predicted that automation would soon eliminate the productive role of most workers, "[F]or the first time since his creation man will be faced with his real…problem—how to use his freedom from pressing economic cares, how to occupy the leisure, which science and compound interest will have won for him…" Instead, 1930 turned out to be nearly a historical low point for labor, measured by the share of national income in received or "labor's share." The next forty years, in contrast, had the highest and steadiest labor's share recorded throughout the developed world, at roughly 70%.

Perhaps the most interesting aspect of this failed prediction is the role Keynes played in preventing the realization of his own vision. Mass unemployment from the Great Depression stimulated a generational focus on full employment as a central economic doctrine, reinforced by the emergence of Keynesian economics with the publication of Keynes's *General Theory of Employment, Interest and Money* in 1936. The Keynesian era that lasted until the mid-1970s saw a policy, social, and economic landscape favoring labor as never before. Labor unions were at the height of their prestige and power. Beginning with the Second World War, the economy consistently approached full employment levels throughout the developed world. Perhaps most importantly, technological progress sustained these trends.



The "displacement" effect occurs when capital takes over tasks previously performed by labor.
The "reinstatement" effect occurs when technologies create new tasks in which labor has a comparative advantage.
Source: Researchers' calculations using data from the Federal Reserve Bank of St. Louis, the Bureau of Economic Analysis, and the Bureau of Labor Statistics

Figure 1: Labor Share of Output, 1947 - 1987 vs. 1987 - 2017

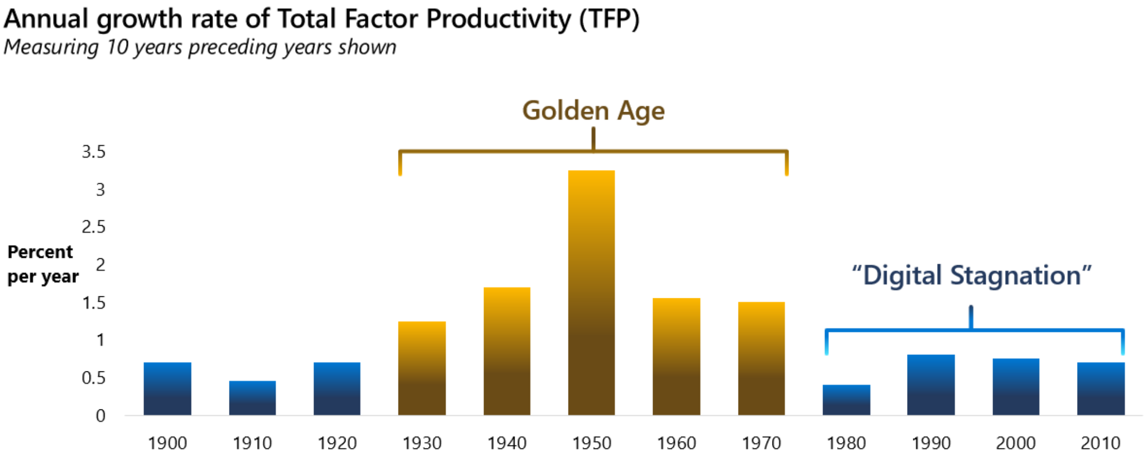


Figure 2: Growth in Total Factor Productivity, 1900 - 2010

Consider Figure 1, drawn from Acemoglu and Restrepo (2019). This shows the impact of technological change over time on labor's share of income. They decompose the impacts of technology into two components: "displacement" (a synonym for automation, where technologies replace labor and decrease labor's share) and "reinstatement" (where technologies create new opportunities for productive labor and raise labor's share). From 1947 to 1987, these two forces were largely in balance and labor's share held steady. Acemoglu and Restrepo (2020) also show that over this period reinstatement tended to favor those with limited education, offsetting the tendency of displacement to disproportionately harm this group.

This period also saw exceptional overall productivity growth, as illustrated in Figure 2 drawn from Gordon (2014). Productivity in the United States grew at roughly twice the rate it had before that time.

These trends dramatically reversed beginning in the mid-1970s. Technological displacement accelerated while reinstatement of new jobs slowed. Consequently, labor's share of income fell significantly, accelerating after 2000. Faster displacement came to harm those with less education, while reinstatement reversed and began favoring those with high education. Overall productivity growth rates halved.

A number of causally interrelated factors changed between these periods with shared responsibility for these changes. These include the rise of market power and industrial concentration, a neoliberal ideology that undermined the will and capacity of collective governance institutions to keep pace with technological change, a gutting of the federal regulatory systems, and exclusionary practices of zoning, taxation, and governance in leading metropolitan regions. There is a growing economic consensus that first among these causes is the direction of technological change towards a focus on automation and labor replacement.[2] For example, Acemoglu and Restrepo (2021) estimate that 50 to 70% of the rise in wage inequality during this period is accounted for by the rise of automation. Automation in this period did not rely on the tools that are today sometimes labeled "AI," which are more recent origin, but was heavily premised on a robotics tradition that grew out of the previous generation of AEAI in the 1980s.[3]

Some argue that present economic impacts of AEAI are simply a detour en route to a more equitable future. Proponents of this view often dismiss large-scale social and economic concerns with hope for Universal Basic Income-type policies and other post-hoc top-down redistribution schemes (Yang 2020; Altman 2021). While redistribution via a progressive tax system and a robust social safety net is necessary for broadly shared economic growth, promises of such future change by governments, when the actors pursuing AEAI do little to address the distributive and power balances in development of their own systems, lacks credibility either politically or economically. See Hart (2019) for a more detailed discussion.[4] Tax and regulatory schemes that make it systematically cheaper to employ capital over labor unnaturally accelerate these trends. Moreover, by focusing on replacing human capacities, AEAI imposes a growth ceiling that coincides with displacement.

B. THE SOCIO-POLITICAL EFFECTS OF OPTIMIZATION

Socio-political dangers stemming from the delegation of social decision-making power to AEAI systems based on opaque optimization functions are very much a present, rather than a future, concern, as discussed by Benjamin (2019), Crawford (2021), Eubanks (2018), Noble (2018), and O'Neil (2016).

At the extreme end of the spectrum, engagement maximization on Facebook helped fuel a genocide in Myanmar. Links between social media–fueled disinformation, polarization-promoting content, and real-world political harm are clear in India, Brazil, and the US. Growing evidence shows that the polarization and division of today's politics are enhanced by algorithmically driving online engagement towards outrage. These are extremely concerning dynamics given current geopolitical trends, from the rise of populist and political conflict to a widespread inability of governance structures to address global problems such as climate change and pandemics.

What responsibility does a policy and research agenda focused on AEAI bear for these socio-political outcomes?

The short list of companies that build cutting-edge AEAI systems can essentially dictate the terms of online life. Broadly, their power has led to what Citron and Pasquale (2014) termed the "scored society," a state in which pervasive and opaque algorithms produce authoritative scores of individual or group reputation (in either narrow or broad contexts) that mediate access to opportunity. Since 2014, this type of scoring has only broadened, with private companies essentially owning avenues of access to social networks, opportunity networks, and even information at large—and apportioning this knowledge based on optimizations for profit via opaque scoring mechanisms. Not only does this violate due process, as pointed out by Citron (2008) and Crawford and Schultz (2014), it is antithetical to the long-term development of democracy at any level—whether in communities, in the workplace, at the state and local level, or internationally. Democratic processes in any context require participation, understanding, and material ability to affect outcomes. These goals are impossible with opaque and autonomous systems that rapidly optimize inscrutable objectives based on allegedly objective but always socially constructed data sets.

Some would say that the AEAI policy and research and agenda is not the culprit here, but capitalism itself (Atanasoski and Vora 2019). As this argument runs, the problem is not the use of reward-maximizing algorithms to optimize engagement

[2] While our analysis has focused on the case of the United States and the broader Western world, the general trends we delineate are likely to impact the developing world even more severely, reproducing hierarchies of race, gender, class, coloniality, and geopolitics. Tasks being automated are even more heavily represented in the workforces of the Global South, where labor bargaining power may in many cases become even more depleted and access to ancillary benefits (e.g., from taxing technology companies) is even more remote. Non-Western countries that currently have or aim to have better labor protections and social programs than in the West may be incentivized to slash these in a global, AI-fueled race to the bottom. Moreover, nationalism and the fear of racialized "others" getting ahead technologically have been a consistent theme among advocates for investment in AI, even against empirical evidence of its failings—beginning with American responses to the 1982 announcement of the Fifth Generation Computer Project in Japan. While China may be a partial exception, evidence increasingly suggests that gains from a growing focus on automation are uneven and are delivering limited aggregate productivity gains while fueling oppression of minorities (Lei 2021).

[3] In fact, according to Google nGrams, the frequency of appearance of "artificial intelligence" in English in 2019 was still shy of its previous peak in 1988.

[4] For example, the "Universal Basic Income" levels widely proposed by political leaders like Andrew Yang would offset only a tiny fraction of the upward redistribution associated with dramatic rises in capital's share of income, in conjunction with gutting existing social welfare programs and social spending, often leaving beneficiaries worse off. Even a more ambitious redistributive agenda (Altman 2021), such as the windfall-clause type framework proposed by some (O'Keefe et al. 2020) address only the purely market income elements associated with extreme concentration of economic power and do so only in a future where a great deal of wealth has already been accumulated by those in control of the AI systems. As Jiménez-Hernández and Seira (2021) show, in the presence of significant firm market power, transfers can be counter-productive as firms will raise their prices. There is every reason to expect this to be far more severe in the future of extreme concentration of economic power envisioned by the AEAI community. A more effective strategy likely involves building towards the growth of public goods and commons-based resources.

but the fact that firms are deploying these techniques to maximize profit. Firms could apply the same techniques to better objectives, the case is made, if they had the right incentives to do so.

Clearly, incentives are part of the problem. Yet, as Stray et al. (2021) highlight, the metrics firms are using to drive their optimization tools are not in fact simple articulations of capitalist objectives. For instance, standard metrics around engagement, clicks, and the like do not actually measure "lifetime profit per user." Instead, such metrics are coarse but easily measurable proxies that correlate poorly with harder to measure outcomes such as customer satisfaction, likelihood to recommend a product, intent to return in the future, decision to purchase on net, etc. AEAI redefines complex issues as optimization problems, extending Muller (2018), pulling from hundreds of numerical optimization examples resulting in deleterious performance and referred to as the "tyranny of metrics." That process of redefinition itself, and its implications, are not driven or defined by capitalism at large but independently by the conceptual structure of AEAI. Capitalist dynamics are at least as strong in, for example, the cryptocurrency space and have led to many harms, but of a different kind (speculative attacks, hacking) than those caused by engagement maximization using feedback-driven learning. The response by platforms to these challenges has highlighted and reinforced the centralizing dynamics at play.

IV. THEORETICAL PROBLEMS

The economic and socio-political harms of AEAI do not arise by chance: the very structure, framing, and goals of AEAI are culpable. In what follows we highlight that the AEAI vision is theoretically unpersuasive and even implausible and that the present harms generated by the narrow pursuit of these goals are likely to be exacerbated by a continuation of such a pursuit, resulting in dangerous long-term futures. Before we turn in the final section to articulating our alternative approach to human-complementing and pluralist AI, we need to finish detailing the harms of AEAI. Each of the core conceptual and practical commitments of AEAI—competition, autonomy, and centralization—has theoretical flaws and delivers harms. We offer that theoretical critique in this section.

A. AGAINST HUMAN COMPETITION AND OTHER NARROW TECHNICAL BENCHMARKS

The practitioners of AEAI measure progress by passing benchmarks and competitions, with the most common benchmark being human parity in a task. For instance, tracking projects such as Stanford's AI Index (Stanford HAI 2021) and the Electronic Frontier Foundation's AI Progress Measurement Project (Eckersley, Nasser et al. 2017) characterize progress in the field with comparisons vis-à-vis human performance across subfields. Yet there is little theoretical basis for considering human parity as a useful target supportive of the interest of long-term flourishing of the human species. Instead, this emphasis is most likely to create brittle technology capable of competing with and substituting for humans rather than complementing them.

This may seem counterintuitive in some cases where it is human communicative or collaborative capacities that the system aims to replicate. Yet even in these cases, parity metrics are at best imperfectly correlated with desired aims. For example, many of the most effective human communication media (e.g., video conferencing) do not replicate any defined human capability but instead facilitate communication using distinctly non-human infrastructures. They complement and extend human capacities rather than replacing them.

We note that human parity in specific fields is often seen as an intermediate step towards the goal of absolute dominance over humans in every domain. However, the focus on achieving "human-level intelligence" via task parity itself creates waste and harm on several dimensions.

The focus on achieving human-level intelligence sets the relationship between humans and machines as one of competition, rather than one of cooperation and augmentation, which both excessively displaces workers and forgoes myriad opportunities for improving human productivity. An array of paths exists for developing productivity-enhancing technologies. The current one—based on achieving human parity and automating human work across various tasks—is only one option, and in fact an extreme one. Pursuing automation in the many areas in which machines do and will exceed human capabilities is not in itself problematic. The problem is focusing solely on automation, even in places where algorithms, whether AI-powered or not, seem unlikely to have a comparative advantage over humans in the near term. A strategy focused on substituting for a resource that the same strategy aims to make abundant (through technological unemployment) is a poor allocation of economic resources.[5]

[5] It is exactly this socially costly path that AEAI is currently following and as a result, it is leaving low-hanging fruits uncollected because it ignores paths of human augmentation and machine-human cooperation. We note that the growing interest in research on multi-agent systems, cooperative AI, and human-computer interaction may mitigate the most glaring harms in this space.

Such focus fuels a more serious problem: ignoring areas where technology could create new tasks for humans. These problems flow from more generally seeing automation as the goal, rather than as a potential side-effect of what would be a better goal: creating new opportunities and productivities for humans. When technology focuses on automation as a goal rather than a potential side effect, it both displaces workers and fails to generate the new tasks and opportunities that would reinstate workers into the production process, doubly disadvantaging them. Other approaches that focus on areas of apparent machine comparative advantage and create new opportunities and productivities for humans are far preferable across both economic and social measures. A canonical illustration is DeepMind's AlphaFold, which is a protein structure prediction model. This work proceeded by focusing on an area where computers already exceeded human performance (protein structure prediction) and where computer performance was highly complementary to a human capability (medical science) important for a human need (relief from sickness and suffering).

The focus on achieving human-level intelligence for currently existing tasks freezes in place contingent notions of current economic value. What humans are capable of and what the economy judges to be valuable co-evolves with technology. The focus on automating economically valued human work, as in the OpenAI mission, assumes stasis in what is economically valued and misdirects us away from the crucial work of evolving the economic system. Given that our current economy and distribution of economic labor evidently does not optimize for human flourishing, further optimizing and automating what is "economically valuable" in that system (especially in narrowly ambitious ways) is unlikely to align with human values broadly speaking.[6]

Focus on achieving human-level intelligence via automation creates externalities to workers in imperfect labor markets and degenerates into diminishing returns. By exclusively pursuing automation, the current path may not realize potential productivity gains. Modern labor economics emphasizes a variety of reasons why workers on average value their jobs and why targeting technologies most likely to replace them is a poor allocation of effort.[7] On the other hand, a focus on creating new opportunities will tend to have positive spillovers on labor markets. This means that focus on automation is misplaced and faces rapidly diminishing returns.[8] Its productivity ceiling is displacement (Mindell 2015; Suchman 2007).

Focus on human-machine competition mechanizes our conception of human capacities. Turing tests always involve a human judge and competitor. Success can be also achieved by dumbing down either of these rather than improving machine performance. This benchmark discourages us from using our technology to see value in our humanity and encourages us to behave like machines.[9]

B. AGAINST AUTONOMY AND ALIGNMENT

While "human competition" benchmarks for AEAI are unproductive for supporting a goal of broadly flourishing human

[6] "Economically valuable" is not some absolute category, but one that has dramatically shifted over time, for example from ignoring work in the home, to beginning to partially account for it. Moreover, most skills that are highly economically valued today (programming, financial modeling, etc.) would have been not only unvalued but also meaningless and impossible to execute in the relatively recent past.

[7] This phenomenon closely parallels a canonical property of work on humanlike machine designs called "the uncanny valley." As the realism of animations or imitations of humanlike forms improves, human consumers generally enjoy them more (have higher "affinity") until a critical point is reached, a bit below full verisimilitude, where suddenly affinity drops and only begins to return once verisimilitude is near perfect. We can roughly think of the same phenomenon applying to the productivity implications of improved technology on a task as a function of the fraction of human-level performance achieved. If machine performance is well below or above human level, it will not be a significant substitute for human labor and thus improvements will be beneficial. However, near human-level performance, significant substitution will occur from labor to capital. Assuming any of a host of standard labor market imperfections (efficiency wages, labor market power, mismatched tax rates), this substitution will harm workers on the margin and thus undermine the effects benefits of enhanced productivity/reduced cost. Thus an effort to surpass human capacity in a range of skills focuses attention on precisely the least productive direction for research.

[8] This can also contribute to a massively unequal world. The huge majority of the eight billion people on this planet do not fall into the category of very highly educated programmers, designers, and managers. The vast majority of people outside of the industrialized world still depend on their human labor for their livelihood. For them, automating AI is the inappropriate technology par excellence. The idea of inappropriate technology entered public debates in the 1960s when economists noticed that the capital-intensive agricultural technologies developed in the West and imported into poorer countries were not generating much value and were reducing the demand for labor, with adverse social consequences. The effects of AI on workers of the world will be much more massive in our globalized economy, not only because of its pervasive use across many sectors, but also because it will impact supply chains, offshoring patterns, and imports decades before it is adopted in the developing world.

[9] For an AEAI system to accomplish a given human task, that task usually must be considerably redefined—translated into the languages, operations, and formalisms of computing—often leaving out much of what brings people to the task and what people bring to the task in the first place. Human parity comparisons risk a contradiction in terms: AEAI never does what people do; when it succeeds, it succeeds at the redefined task (Dick 2015). AEAI also participates in much longer histories of attempts to replace humans by machines. Often these efforts required that humans become more machinelike in advance, a part of that redefinition, as in factory automation and human computing in the nineteenth century. Charles Babbage, when he invented calculating machines to replace human computers, marveled that the best human computers *knew the least about mathematics*—the less education and the less active thinking they brought to the job, the better (Babbage 1832). Calculating machines did not replace the philosophers, astronomers, and mathematicians who had engaged in hand calculations in the eighteenth century, they replaced the low-paid, under-educated human computers that emerged through the application of Adam Smith's principle of division of labor to calculating (Daston 1994). This problem is compounded in the long term by the inherently competitive nature of human parity metrics.

societies, they are at least a meaningful objective, and would be worth pursuing if, for example, your overarching goal was to enable a technical elite to operate as independently as possible from the rest of society. In contrast, the conceptual commitment of AEAI to autonomy is based on a fundamental misunderstanding of intelligence as it exists and functions. **AEAI's atomized view of intelligence, captured in the "autonomy" concept, misunderstands how socio-technical systems produce value. This makes AEAI systems inherently susceptible to Goodhart's Law (the tendency of optimizing quantitative systems to overoptimize what they measure), despite the avoidance of such situations being the stated goal of much of AEAI safety research, particularly alignment research.**

Intelligence is not an autonomous but a social and relational quality. While there is significant dispute about what "intelligence" means or whether it is even a useful concept (Broussard 2018), most attempts at definition focus on capacity to solve useful problems and formulate plans to achieve desired ends. Both empirical study of such capabilities in humans and computational and economic theory strongly suggest that such intelligence is not primarily a property of atomistic individuals but of social and cultural systems.[10] Systems that aim to achieve something like the intelligence perceived in humans will thus depend on their capacity for interdependence, sociality, and collective memory, not autonomy.[11] Research into cultural evolution have consistently demonstrated that anatomically modern humans progressed *very* slowly (e.g., in productivity-enhancing lithic technology) when institutions of collective communication and memory were not present (Henrich and McElreath 2003). Autonomous visions of intelligence are far from the only template for digital technology; in fact, visions focused on facilitating and participating in communicative and collaborative networks are at the center of many of the most celebrated programs of collective participation, from the internet to mutual aid.

There is little support for intelligence as a homogenous and universally applicable quality. These considerations also militate against the notion of intelligence as a general/abstract quality equally applicable to all problems, the contentious early twentieth-century notion of a singular, one-dimensional Intelligence Quotient in humans, and what Bostrom (2012) labels the "orthogonality thesis." Arguments based on generality and universality often draw on theoretical results, such as the Church-Turing thesis, which show that *in the absence of time constraints*, general intelligence is possible. However, these results are essentially irrelevant within any finite time interval, where different architectures can be vastly more effective at some tasks. This is further supported by research in cognitive psychology, science and technology studies, and the history of science, which emphasizes the contextual, embodied, co-created, and relational nature of intelligence (Collins 1990; Crawford 2021; Forsythe 2001; Dick Forthcoming); all of these points have a long history in debates over artificial intelligence (Wang 1990).[12]

Autonomous conceptions of intelligence are particularly subject to Goodhart's Law. Concerns about Goodhart's Law apply primarily to systems that aim to autonomously pursue ambitious goals with limited temporal or social constraints (Drexler 2019). This is particularly true when paired with a focus on surpassing human capabilities, exemplified in canonical thought experiments such as the "paperclip maximizer" described by Bostrom (2003). AEAI, in the long term, calls for the development of powerful, autonomous, human-independent technologies, and simultaneously decries these as most responsible for soaring risk potential in the long-term future. The clear alternative is optimizing for limited and constrained systems that are deeply integrated into social and communicative frameworks, which have far more circumscribed scope for Goodhart-style failure modes.

[10] Optimizing, or even defining, complex objectives is computationally demanding. For example, even flexibly defining preferences over schedules for courses for a single semester at a typical university requires more space than is available to all computers on the planet (Budish and Kessler 2016). Optimizing most relevant objectives is thus impossible and computational constraints will be a central concern. To the extent there is other potentially useful computational capacity outside a system, it will typically be worthwhile to devote significant resources to reaching and communicating with this capacity to lessen these constraints. This implies that any computational system that has the reasonable potential of accessing other external computational systems will find communication, collaboration, and interdependence to be far more effective than autonomy for most ambitious objectives. Relatedly, during the late nineteenth century's "marginal revolution," economists realized that production of value is typically not a linear sum of the value/intelligence of workers. Instead, it involves significant complementarities across workers. Autonomous notions of intelligence revert to simplistic, linear theories of value. Even when complementarity is pursued, it is between machine units and processes, foregoing the complementarities of the broader system. Perhaps most powerfully, psychologists from John Dewey (1939) to Mercier and Sperber (2017) have shown that even human motivations and basic reasoning capabilities fundamentally arise out of social interactions rather than as individual decision-making capabilities. Communication and sociality are so deeply baked into intelligence as we know it that ideas of autonomously maximizing individual or collective utility functions have limited applicability and are especially inappropriate as models for ambitious, long-term technological evolution or alignment. Other alternatives, based on clear definitions of scope conditions, affordances, and failure modes are available and widely used in many fields of digital technology, as we discuss below.

[11] To its credit, this view is becoming increasingly common in the AI safety community, as we highlight below. For example, Russell (2019) highlights that AI safety will require machines to remain uncertain about their goals, leading them to constantly check in with humans who can help them learn about these. This, however, significantly undermines the goal of autonomy; if taken further to allow humans themselves to be uncertain about their goals and to learn from discourse with other humans, we are led away from AEAI research and toward various forms of digital plurality as described below.

[12] While paradigms of universal computation have had powerful influences on, for example, the separation of software and hardware, specialized architectures have repeatedly returned, including recently for computing neural networks. The history of computer science thus gives evidence for the practical heterogeneity of intelligence, even though this is rarely acknowledged.

Autonomy tends to obscure and undermine important external agency critical to making systems function effectively. The production of the myth of "autonomy" for AEAI systems both perpetuates the erasure of certain classes of human labor and obscures how much the scaffolding provided by a tiny technical elite govern how they operate and for whom (Atanasoski and Vora 2019; Gray and Suri 2019; Irani 2015; Ross et al. 2010). Our technologies should reflect social realities rather than the fiction of autonomy not only as a matter of ethics and politics, but also as a simple matter of efficacy and transparency. This is evidenced in the substantial literature on the increased efficiency and resilience to downturns found in many worker cooperatives (Pérotin 2016). Moving past the myth of autonomy has the potential to dramatically increase the quality of large-scale statistical models, as we discuss below.

We finally note that the formalisms and claims of AEAI are naturally alienating to the very humans whose collective good is meant to be served. If autonomous systems are to aim at alignment with human values, central must be alignment of the development process to systems such as social norms, law, and politics that express those values, which are currently in tension with the language of centralized optimization and autonomous operation.

C. AGAINST CENTRALIZED SCALE

If the pursuit of human replacement and use of autonomous conceptions of intelligence were a niche pursuit of a minor subculture, such a research and practical program would not merit focused critique. Yet practitioners of the AEAI research and practical program view their work as among the highest priorities for human civilization and are successfully securing escalating resources for the work. Moreover, that resource investment is funneled into a vision of future goals that adds, as a third fundamental element, centralization. This combines with competition and autonomy to make AEAI a dangerous program.

AEAI depends on two symbiotic future visions, one optimistic and one pessimistic, both dependent on centralization. In the first, optimistic view, concentrated investment in a small number of people achieving distant and ambitious AEAI goals will yield broadly beneficial, and indeed spectacular, outcomes for humanity. In the second, pessimistic view, not achieving these very distant goals in a sufficiently aligned and tightly controlled manner will result in significant, potentially existential risk. Both views are fully dependent on the centralized control of AEAI systems, effectively concentrating the power over vast resources in the hands of a very small group. The worldview of this elite will travel as their systems do, displacing alternative perspectives, domain-specific expertise, and pluralist values and epistemologies.

The 2021 AI Index Report gives a powerful look at what this means in practice. It finds that in the last decade, AI research publications have tripled as a share of all research publications, from 1% to nearly 4% of journal papers and to 20% of all conference publications. AI has grown from a tiny share of venture investing to more than 14% globally in 2020 and more than half of venture investing in digital technology. Computational resources used by large AI training models have been doubling every 3.4 months since 2012, implying an explosion relative to global available computing resources given that Moore's Law implies capacity doubling only every eighteen months, which is itself slowing. This growth is concentrated in a small and homogenous group of technologists, engineers, and researchers, often from elite institutions, with well-known failures to achieve diversity in race, gender, geography, and social class (Whittaker et al. 2018).

Extreme concentration of control over the direction of productive resources is neither a new idea nor a new phenomenon. It was not even new when Plato argued for philosopher-rulers in The Republic. The phenomenon has reemerged in a variety of guises throughout human history—whether as absolute monarchism, Robespierreism, or Soviet planning.

Yet assertions of centralized control are illusory. As many have shown (e.g., Bockman 2020), the Soviet claim to conduct central planning gave way in practice to decentralized decision-making activities, given the inability of central planners to perceive or act on all the details necessary to implement their plans on the ground. And ironically, the technocratic attempt to overhaul Eastern European economies into bastions of free market capitalism fell prey to analogous failures. Similarly, AEAI systems that aim at "neutral fairness" via control by engineers end up instead defaulting to replicating the biases of societies whose data they train on (Crawford 2021; Noble 2018).

Thus, the aspiration to and illusion of totalizing central control consistently results in catastrophic failures, commonly associated with an inability to perceive and process diverse information and input from individuals and communities "on the ground." Without processing these inputs, centralized systems tend to over-optimize to narrow criteria with disastrous, Goodhart-like consequences ranging from giant cities no one wants to occupy to "accidental" genocides driven by forces ranging from miscalculations of crop yields to algorithms maximizing engagement (see Weyl 2019). It also leads to the formalization of a narrow set of values—namely, those of the designers of these systems, currently a highly concentrated and homogenous group in almost every way imaginable—at the expense of most others. Without deliberate work to create democratic governance, the feedback mechanisms inherent in technological progress in both capitalist and statist systems generally, and AEAI specifically, ensure that this concentration of both material benefits and, more importantly, agency over the future of technology values only accelerates, with more and more humans left out of the decision-making processes and benefits.

This view allows us to see clearly that the so-called AI competition between the West and China is a false one, because framing the debate in terms of AEAI already settles the fundamental ideological questions that are supposedly at stake. The centralizing tendencies inherent in AEAI tend towards the same outcome in both cases: siphoning resources and decision-making power from the majority into the hands of a technocratic elite through Manhattan Project–like AI programs. AEAI produces a race to the bottom in terms of the safety, ethics considerations, and privacy protections.
For example, programs like China's social credit system are targets of critique in the West, and rightly so, as they are leveraged to control access to resources like schooling and transportation. However, our own convoluted systems increasingly employ black-boxed automated or AI-driven decision-making over employment, loans, insurance premiums, and home offers, essentially following the same trajectory: one of unaccountable, AI-driven control over human lives and choices with minimal recourse (Benjamin 2019).

Further, if the West and China are in a race to implement AI, this can only result in both degenerating into centralized control. After all, we have seen Silicon Valley capitalism largely producing a direction for technology similar to the Chinese model (see, e.g., Altman 2021). The real alternative is outside the space of such competition, which breaks through the binary of surveillance capitalism and a surveillance state with pluralist, democratic participation.

The history of technocratic "central planning" is helpful in reinterpreting what is so worrying about mission statements like those of OpenAI and Hassabis's for DeepMind. Once one unpacks the core commitments of AEIA—competition, autonomy, and centralization—it is clear that the mission statements with which we began should be understood as follows.

"OpenAI's mission, fully understood, is to ensure that universal central planning—by which we mean a centralized command economy run by a technocratic elite that outperforms current economics systems on what we consider to be most economically valuable work—benefits all of humanity."

DeepMind's mission, fully understood, is to "Solve central planning, and trust the central planners to decide how to use central planning to solve everything else."

V. TOWARDS DIGITAL PLURALITY

When we see "internet of things", let's make it an internet of beings.
When we see "virtual reality," let's make it a shared reality.
When we see "machine learning," let's make it collaborative learning.
When we see "user experience," let's make it about the human experience.
When we hear "the singularity is near," let us remember: the Plurality is here.

—Audrey Tang, Digital Minister of Taiwan

If the vision of intelligence as autonomous is the wrong horizon for technological aspirations, what is the alternative? Naming a single, alternative, "correct" path would be self-defeating, falling into the same centralizing tendencies as AEAI exhibits. Instead, we contend that the alternative to AEAI is not a singular, narrow focus on a specific goal, such as achieving "general intelligence," but rather research and policy support for a plurality of complementary and dispersed approaches to developing technology to support the plurality and plasticity of human goals that exist inside the boundaries of a human rights framework. This alternative to AEAI already exists. Dissenters to the AEAI approach have been forging new paths. Taken together they show what the key features of an ecosystem of "digital plurality" might entail. We will refer to this alternative pathway as "actually existing digital plurality" (AEDP). The goal is to transition from AEAI to AEDP.

But what exactly does this alternative entail?

Digital plurality resists pithy definition. Instead of aiming towards a technical end-state, it describes an ecology, comprising approaches that cooperate, co-exist, and co-evolve, and operate in support of human decision-making about social well-being, operating within the constraints of a human rights framework. These approaches create, intersect with, and support new modes of decision-making. By raising ongoing human goal-setting to the surface as governing technology, they transform narrow technical questions into opportunities for innovation, whether achieved through collective digital participation or other modes of decision-making. Rather than converting questions of social progress into formalized inputs for narrow technical expertise to resolve, they support and extend the human capability for directional goal-setting and fair, just, and productive collaboration. As we seek to characterize "actually existing digital plurality" for the first time, we describe attributes of this emerging ecosystem. These attributes are reflected within the theoretical statements motivating work in this space but not yet perfectly replicated in the technologies that constitute it.

As we see the emerging alternative, three shared conceptual and practical commitments define AEDP: complementarity, participation, and mutualism. The content of those core commitments can be summarized as follows:

1. **Complementarity.** Technology should complement and cooperate with existing intelligent ecosystems, not replace them. Technology should broaden the surface area of complementarity—across individuals, organizations, and systems—allowing for ever more networked evolution.
2. **Participation.** Intelligence is collective, not autonomous. Technology should work to facilitate the social nature of intelligence, and in particular to facilitate deliberation on and participation in setting outcomes in equal measure to driving the achievement of outcomes.
3. **Mutualism**. Decentralized, heterogeneous approaches under the umbrella of digital plurality can build on and benefit from each other: technologies evolve in interaction with each other and social, political and economic institutions and ecology.

For the remainder of this section, we'll expand on the current state of the AEDP ecosystem and then provide an overview of the history of its emergence. A wide range of social and technical projects in and around the digital world are currently developing promising alternatives to the AEAI vision. These projects differ in the extent to which they target the three pillars of AEAI; only a subset of them break away from AEAI entirely to define a new ecology of digital plurality. There are three rough lanes of work leading to the development of AEDP. They share different elements of the three commitments named above, and so we'll introduce this field by focusing on these three areas of practice, each in turn.

A. ACTUALLY EXISTING DIGITAL PLURALITY

The first lane of work within AEDP focuses on mitigating the problems with autonomy. These approaches accept the necessity or likely inevitability of autonomous systems that aim at general, human-style intelligence, but are concerned about the tendency of such technologies to become "misaligned" with human agency or have otherwise harmful effects, often because of centralized control. These approaches thus tend to call for ethics, governance, safety, or redistribution of AEAI systems and the benefits they create. Practitioners are to be found in various parts of the AI ethics and FAT* communities, the AI safety/alignment/existential risk community, and the AI governance/regulation community, and work on universal basic income, sustainable AI, windfall clauses, capital taxation and related redistributive mechanisms.

The second lane of work challenges one of the two more technical premises of AEAI: competition and autonomy. There are a range of approaches that maintain focus on autonomous intelligences but break from the competition concept and break from the focus on imitating and/or displacing a singular, anthropomorphic concept of intelligence. This work aims to develop intelligences that are unrelated to human intelligence (e.g., AlphaGo), that perform well in areas humans are known to be weak at (e.g., protein folding with AlphaFold), or that complement/collaborate with humans, such as work on optimizing metrics of human complementarity (e.g., Wilder, Horvitz, and Kamar 2020). Examples here include cyborg technologies, brain-computer interfaces etc. One opinionated survey of this perspective is Drexler's (2019) vision of an "AI services" model as an alternative to the unitary, human agent-like intelligence envisioned in much of the AI safety literature.

Within this lane, there are also approaches that embrace the competition concept of AEAI systems (and the goal of replicating various human skills at scale), but that take aim at the autonomy concept and the vision of the systems achieving these goals as needing to be autonomous. Practitioners in this lane aim to explicitly account for, build off of, and harness the capabilities of the organizations and individuals that enable digital systems to function. This lane embraces work on technical elements (such as the privacy-preserving machine learning stack), on economic designs (e.g., data dignity/data as labor), on governance and legal institutions (e.g., data trusts/collaboratives), and on interaction paradigms (e.g., machine teaching, human-in-the-loop systems), for instance. In contrast to the GPT-3 paradigm, some recent work from OpenAI in developing AI systems for pair programming (e.g., Copilot, a pair-programmer AI) would also fall into this category.

A key challenge for those operating in these first two lanes of work—seeking to mitigate the problems of autonomy or to challenge either competition or autonomy—is that their partial embrace of AEAI can lead to a defensive stance and to a representation of the work as a matter of restraining or containing inevitably powerful forces that may nonetheless be released by an unscrupulous actor. For the most part they do not themselves represent an alternative positive vision of a powerful direction for technology. This is particularly true for approaches that intend to counteract the inherently centralizing nature of AEAI, our most foundational concern as laid out in this paper. Work on many standard AI ethics issues, including accountability, legibility, transparency, bias-mitigation, diversity, and inclusion are crucial if AEAI remains a power center in our society. Similarly, work on post hoc redistribution is crucial to the extent that our technical systems continue to be set up to carry forward incredible concentrations of wealth and power to logical extremes. Yet ultimately none of these efforts can, we believe, overcome the fundamental problems with AEAI, any more than diversifying the aristocracy of Old Regime France would overcome the fundamental injustice of the social system that aristocracy controlled.

In fact, one of the most valuable things about these efforts is the way they continue to expose the fundamental flaws of AEAI itself and to bring into the field people who are more likely to reckon with these fundamental flaws. On the other hand, to the extent that work in this area is used to legitimate AEAI more broadly or make it appear to be a field pursuing its goals in an ethical and socially desirable manner, such work may be inadvertently harmful.

The third lane of work breaks entirely with the AEAI trajectory and therefore provides us the fullest picture of what AEDP amounts to. These approaches are the most diverse and focus on decentering the role of technology and its capabilities while centering complementarity, participation, and mutualism. This has led to a nascent but powerful ecosystem in which technological approaches and evolving human systems for goal-setting and decision-making build on and co-evolve with each other.

This lane of work includes many threads that began as disparate areas of practice. The human-computer interaction and human-centered design communities have a rich history of plurality, evolving a program of understanding needs and limitations and building systems engineered to effectively fulfill these needs, allow for flexible use, and compensate for limitations. Distributed agency has entered the design of technology infrastructure itself through self-organizing mesh networks and decentralized edge computation. Economics and mechanism design have played an increasing role in designing and building decentralized digital ecosystems from microfinance and e-commerce platforms to more radical alternatives such as those emerging in the blockchain ecosystem. Building on these ideas and broader social science, there is an emerging "social technology" agenda that harnesses social and computation science to build new, responsive institutions enabled by digital technology. Much of this has been practically harnessed by citizen science initiatives that make use of collective intelligence to push the boundaries of knowledge.

As these approaches have matured and developed, intersections and second-level complementarities have formed, giving rise to new directions. We've seen decolonial technologies that work to distribute power and voice and are grounded in historical analyses of existing post-colonial power relationships that shape our technical priorities, politics, and infrastructure (Lewis et al. 2018; Mohamed, Png, and Isaac 2020). Work on human-computer interaction (HCI) has begun to understand the need for participatory action research and collaborative design that rigorously involve stakeholders in the design process, often utilizing deliberative and constructive tools seen in other approaches. There has been a flowering of work on decentralized technology that builds heavily on work in blockchain and aims to augment internet protocols to allow for peer-to-peer information transfer and decentralized network interactions. This has led to the formalization of digital commons and knowledge commons, creating processes for systems from Wikipedia to open-source code repositories that provide access to collective innovation while instituting polycentric, multi-stakeholder governance structures across local, regional, and global levels.

These are diverse directions and the landscape is constantly shifting. And yet, a nascent and growing ecology of digital pluralism, an active and emergent landscape of AEDP, makes it clear that we are not trapped on the path of AEAI. The rich history of this ecology, which we will now turn to, underscores that a better alternative is possible.

B. HISTORY

AEDP may currently be neglected, but it is not novel. Concerns about AI as a direction for the future of digital technology are far from new, even within the narrow technical elite that have dominated these discussions over the past half-century. These concerns date to before the term "artificial intelligence" was coined in the 1950s (see Nilsson 2009; Markoff 2016; and Broussard 2018).

Five years before John McCarthy coined the phrase "artificial intelligence" in proposing the famous Dartmouth summer workshop on the topic, Norbert Wiener's *The Human Use of Human Beings* (1954) critiqued the Turing Test along lines akin to those we review above and warned of the danger of thinking in patterns now associated with AI. Wiener was a leading figure in the field of "cybernetics," which, with limitations, focused as we do on the way in which a variety of diverse forms of information processing interact and form an ecosystem, rather than the autonomous, human-comparative intelligence of a separate machine. In fact, in Ross Ashby's *Introduction to Cybernetics* (1956), he coins the phrase "amplifying intelligence," spelling out an explicitly complementary vision that contrasts with McCarthy's autonomous AI.

Wiener disciple and psychologist J. C. R. Licklider was so moved by the need to advance past the competition concept and acceleration of machine capabilities and instead to achieve "man-computer symbiosis" (Licklider 1960) that he left Wiener's tutelage at MIT to move to the private sector in 1957 and eventually into the military to lead the Advanced Research Program Agency's (ARPA) Information Processing Techniques Office (IPTO). In the latter role, Licklider funded a variety of projects aiming at this symbiosis and the transformation of the computer into a "communications device" (Licklider 1960), including the ARPANET project that evolved into the Internet.

One of Wiener's favorite grantees was Douglas Engelbart, who began his career with the mission that:

1. He would focus his career on making the world a better place;
2. Any serious effort to make the world better would require some kind of organized effort that harnessed the collective human intellect of all people to contribute to effective solutions;
3. If you could dramatically improve how we do that, you'd be boosting every effort on the planet to solve important problems—the sooner the better; and
4. Computers could be the vehicle for dramatically improving this capability.

Engelbart became so concerned with the early explosion of interest in AEAI, with the early work on the B. F. Skinner-inspired vision of top-down surveillance embedded in the design of the early PLATO network system (Dear 2017), and with their potentially existential harm to humanity that in 1962 he founded a competing research project on "Augmenting Human Intellect" at his Augmentation Research Center (ARC) at the Stanford Research Institute. This aimed to build a set of demonstrations of human-complementing technology sufficiently compelling to counteract the allure of early AI human-parity demonstrations. This research program culminated, six years later, in the "Mother of All Demos," in which Engelbart and his collaborators simultaneously demonstrated many of the elements of personal computing for the first time (including windows and the mouse) and explained the philosophy of pluralism, bootstrapping, modular technology development/object orientation, etc. that lay behind them (Engelbart 1968).

In the audience that day were most of the founding leaders of Xerox PARC, including Alan Kay, who credited Engelbart for inspiring most of the human-centered computing developments that PARC eventually produced. PARC, founded the next year (1969) with a nomenclatural hat tip to Engelbart's project, famously went on to develop most of the infrastructure of modern productivity computing in an integrated, commercial-ready package. A central differentiating feature of the research environment at PARC compared to that of other computational research centers was its interdisciplinarity and the inclusion of sociologists and ethnographers (Suchman 1987; Orr 1996). A few minutes down the road at ARC, Engelbart had moved on from a focus on individual human interaction with computers to the potential of computer networks to augment collective human capacities. His lab became the first to fully connect to ARPANET in 1969/70. Ironically, many of the most powerful social applications that matured on ARPANET, ranging from instant messaging, email, group discussion to online news, message boards, and multi-user games, were themselves instances of bottom-up innovation, first piloted by grassroots efforts by distributed users within the PLATO network, upending its top-down design (Rankin 2018).

Not only did HCI researchers at PARC represent a commitment to human-centered technological design, but they also valued the expertise and input of social and cognitive scientists. Sociologist Lucy Suchman and ethnographer Julian Orr both famously spent part of their early career at Xerox and PARC, respectively, offering their understanding and perspective on people and reporting back to academic communities of humanists how technological design was undertaken there (Suchman

1987; Orr 1996; Suchman et al. 1999). The good faith inclusion of those who study sociality and human beings qualitatively, historically, and ethnographically underlay the dramatic difference in the outcomes and human significance of the project underway at PARC.[13]

In parallel to these efforts, the 1970s saw the first efforts to make computation available beyond the walls of the largest corporations, universities, and governments through do-it-yourself kits such as the Altair 8800, which introduced the founders of both Apple and Microsoft to the potential of personal computing and stimulated their ventures. As these ventures opened a broad market for personal computers, they adopted research from PARC to build mass-market graphical user interfaces (GUIs), beginning in earnest with the 1984 release of the Macintosh.

As the commercial market for computers transformed, focus on artificial intelligence within the academy, popular culture, and large corporations nonetheless grew. Disturbed by this trend and persuaded by the critiques of Engelbart and Wiener, dissenters increasingly broke away from the growing hegemony of the AI community. Many research leaders (such as Terry Winograd) became increasingly disillusioned with the intellectual foundations of the field and developed the field of human-computer interaction (HCI) as a more humanistic counterweight to the rationalistic focus of AI (Winograd 1992). Two of Winograd's students, Sergey Brin and Larry Page, went on to found the Google search engine on a pledge to avoid the dangers they saw in the combination of an advertising-driven business model and AI-driven optimization (Brin and Page 1998). One could dispute whether they achieved this goal, but the breakthrough advantage of their initial system crowdsourced the web to disseminate information access.

There were many critics working within early AI and computing research who, on intellectual and epistemic grounds, objected to the project of human parity and autonomous systems development.[14] Joseph Weizenbaum, who famously created the early chatbot psychotherapist ELIZA, noted that exposure to relatively simple computer programs could produce "powerful delusional thinking" about a system's intelligence (1976). Weizenbaum was horrified by predictions that human therapists would be unnecessary and he proposed that *even if* computers could be made to replicate certain human capabilities, the fact that interactions happen between people is central to the project of collective living. He advocated for automated systems that centered human judgment and understanding, especially in matters of governance, health, and democracy, rather than replacing the people and processes that constitute them (Weizenbaum 1976).

Despite many early alternative aspirations for computing, the contemporary AEAI community is largely trapped in thinking that the only possible response is to figure out how to control superintelligences that will outstrip humanity. There is a fever pitch of visions built within a narrow technical community drawing on the work of thinkers like Nick Bostrom (1998) and Ray Kurzweil (2000). This work has helped seed the creation of a "rationalist" community around a series of blogs such as *Overcoming Bias*, *Less Wrong*, and *Slate Star Codex* focused on topics such as improving one's rationality, anticipating the (supposedly inevitable) challenges posed by artificial superintelligence, and aligning the interests of such superintelligences with humanity.

Nonetheless, there are those who are breaking out of that paradigm and, as in the early days, articulating a path toward an alternative and laying the foundations of AEDP. Three broadly related but distinct directions that have shaped the technology landscape today bear mentioning.

As Boyd and Ellison document (2007; 2010), early social networking applications were grounded in ideas drawn from sociological theory, beginning with GeoCities and SixDegrees.com, whose name directly paid homage to research in the field that found relatively short chains of social relationships connecting most people on earth (Milgram 1967). As such, social networking grew from attempts to capture various features of social relationships and interactions in digital settings and aimed to supersede the push towards the extreme of either fully anonymous or corporate "walled garden" identities that had characterized much of the early web such as the AOL platform. Algorithmic, data-driven optimization became central to these networks only once they reached scale and maturity in the late 2000s and early 2010s with platforms like Facebook. Though these optimization processes, when combined with advertising-based revenue models, have now brought us

[13] Today, AEAI researchers instead often reveal epistemic hubris: collaborating little with experts, communities, and scholars outside of a narrowly defined understanding of people and problems. The current division of labor is that technologists build and then humanists, activists, layers, and politicians critique from the outside. It is a far more efficient and promising project to integrate different forms of expertise in the development of technology from the beginning. If technology companies hired, valued, and adequately compensated experts from different fields as well as their critics—*and were open to moving on their suggestions* rather than, as Benjamin put it, trying to enroll more diverse people in the execution of the agenda they have already set—it would save them trouble and money down the road.

[14] Joel Moses, a mathematician at MIT, has argued that there is something inherently hegemonic about the faith in autonomous and centralized control systems, believing that there isn't any one set of algorithmic process that should be applied everywhere. The MACSYMA system—an early algebraic computing system—whose development Moses oversaw, represented a commitment to collectively produced technology and modular design that could be an inspiration for innovation today (Moses 2012; Dick 2020).

harmful dynamics, the original motivation for work in this space carried important insights about the social nature of human intelligence.

Another of the most important directions for technology as a social process arose from a far less socially focused direction of development. The libertarianism of an important subset of Silicon Valley culture, with its focus on decentralization, diffuse intelligence, and market relations, was central to the foundation of Wikipedia (Lih 2009), PayPal (Duggan 2019), and the crypto economy (De Filippi 2014). Recent estimates suggest that Wikipedia accounts for nearly half of the value created by all Google searches, while running on a budget several orders of magnitude smaller than Google (Vincent and Hecht 2021). These projects were wildly different and, in many ways, openly hostile both to each other and to the libertarian ideology that had inspired them. Yet, core to all their visions was a resentment of centralizing tendencies and a view of technology as fundamentally social and co-constructed.

Another arises from the growth of citizen science initiatives that bring together local knowledge via global technology interfaces to support scientific progress. For example, eBird, the world's largest citizen science project, has over a billion records uploaded from individuals in almost every country in the world through its deep-learning-powered Merlin Berd ID app, creating the most complete picture in existence of avian population movements and trends over time.

Perhaps the clearest and least widely appreciated (at least in the West) contrast to the AEAI worldview has come from work that linearly descends from the Engelbart-Winograd tradition. Towards the end of their careers, the pair helped found a research community dedicated to online deliberation (www.online-deliberation.net; cf. https://events.stanford.edu/events/63/6306/). This became a convening point for those interested in systematically studying and advancing consensus-oriented collaboration and collective intelligence using digital tools.

The most ambitious applications of this approach may be undertaken by Audrey Tang and the g0v collective in Taiwan. Tang and the civic hacking group operate on principles of open-source contribution and entrepreneurship, but aim to target these beyond the rationalistic, engineering-focused culture of Silicon Valley (Tang and Harriss 2020). Through the student-led Sunflower Movement, which demanded greater transparency and responsiveness from Taiwan's national government, g0v developed and deployed civic technologies for consensus-making, agenda-setting, and collective technology development. These platforms were so successful that the activists became "reverse mentors" for every member of Taiwan's Nationalist cabinet, with Audrey Tang as the country's first Digital Minister. Under her leadership, Taiwan has rolled out cutting-edge experiments in digital democracy, decentralized governance, and collective intelligence, referred to as the "vTaiwan" and "Join.gov.tw" platforms in which more than half of the country's 24 million citizens have participated. These platforms allow for countrywide sensemaking at scale, without neglecting the deliberative interactions that are necessary for democratic consensus (Siddarth 2021). Data coalitions built with g0v technology similarly utilize the sensor data of thousands of individual citizens and businesses to create a real-time picture of Taiwan's environment, outperforming centralized sensors (Ho, Chen, and Hwang 2020) and allowing for community participation in environmental policy.

This history provides a map of the early conceptual foundations of AEDP and the contemporary practices that now carry that tradition forward as a potential alternative to AEAI for investment and growth. This is itself a narrow view; when we broaden our scope to appreciate the range of technical creativity and vision that historians have been working to recover from other communities and contexts outside these institutional conversations, the scope of technological pluralism becomes truly appreciable (Brock 2020; Hicks 2017). And we emphasize that real pluralism requires both a shift in values—to include and empower the work these communities have already done in what counts as technological innovation and to further empower those outside of the technical elite not just to participate in, but to guide technological development in service of their communities and visions of the future.

We cannot provide any empirical measure of the collective impact and social value of this set of pluralist approaches, because as of yet our policy and research programs don't seek to measure the kinds of value they represent. That is part of the point. They exist outside of the AI "performance league tables." Yet for each of the projects named above there is significant evidence that they have delivered value, and a critical next step in an agenda to build out AEDP would be to provide a fuller collective characterization of that value and impact. But enough discrete and fragmented evidence exists—of both qualitative and quantitative kinds—to justify further engagement with the hypothesis that AEDP presents a different way forward for the role of technology in the economy and society that includes more people, offers more spaces of contribution and interaction, and can better support the plural dimensions of human wellbeing, including pluralism itself.

VI. CONCLUSION

> *In that Empire, the Art of Cartography attained such Perfection that the map of a single Province occupied the entirety of aCity, and the map of the Empire, the entirety of a Province. In time, those Unconscionable Maps no longer satisfied, and the Cartographers Guilds struck a Map of the Empire whose size was that of the Empire, and which coincided point for point with it. The following Generations, who were not so fond of the Study of Cartography as their Forebears had been, saw that that vast Map was Useless, and not without some Pitilessness was it, that they delivered it up to the Inclemencies of Sun and Winters. In the Deserts of the West, still today, there are Tattered Ruins of that Map, inhabited by Animals and Beggars; in all the Land there is no other Relic of the Disciplines of Geography.*
>
> —Jorge Luis Borges, *On Exactitude in Science*

The essence of the current AI trajectory is the view of intelligence as a single, distinct, autonomous quality to be both reached for and feared, and one that, once achieved, is uniquely transformative: a "singularity" (Kurzweil 2000) that will be the "final invention" humans create (Barrat 2013). Just as eschatological religious communities focus their collective lives around preparation for these end times, this AI worldview makes the center of research and planning the arrival of "general intelligence." Once general intelligence is achieved, it will be impossible for humans to keep up or even contribute much directly, and thus all will depend on how well aligned these systems are with human goals.

Even in its less apocalyptic forms, the AEAI agenda is defined by a focus on centralized progress by a narrow and highly trained group of engineers taught to believe that achieving intelligence is fundamentally a technical/computational task of improving machine capacity and performance (Waldrop 2002). Frequent references are made to the Church-Turing Thesis and the sufficiency of reward maximization (Silver et al. 2021) to justify the ultimate equivalence of different forms of intelligence.

Rather than viewing intelligence as a homogeneous, unitary, fungible quality, a pluralist approach emphasizes the diversity, complexity, and variety of approaches as well as the social nature of human intelligence. One quantitative/computational justification for this view is captured by the Borges story above: whatever may be true in infinite time, within the time and space bounds provided by this planet or even this universe, different structures may be vastly more effective at different tasks, undermining any homogeneous notion of "intelligence." This is not to say there is no role for interoperability, homogenization, scale etc.; different pluralist approaches will differ on the extent to which these are necessary and desirable. All, however, object to any singular, final vision of intelligence that obviates all others.

Instead, intelligence is viewed as a complex, emergent outgrowth of a variety of human and nonhuman systems interacting at a range of scales. As such, progress and change arise not from "overcoming bias" or arriving at a singular and obtainable truth, but rather in a variety of perspectives branching and differentiating, each vastly oversimplified compared to any underlying reality yet each capturing and integrating a different and potentially complementary perspective. In this vision, the complexity of different disciplines, cultures, and viewpoints grows with progress and with them so does the need to facilitate communication across them. While some consider general machine intelligence a natural evolutionary step from human intelligence, we contend that an ecological, branching, and complex interplay of intelligent systems is far closer to true evolutionary and exponential progress.

We can replace competition, autonomy, and centralization as guiding goals with principles that support digital pluralization: designing for *complementarity*, increasing *participation*, and supporting *mutualism*. An emphasis on complementarity would mean rejecting the goal of automating human labor, and explicitly designing systems that augment and support workers. This would also include conducting retrospective evaluations of the impact of new technologies on labor markets to improve understanding of net impact of advances, offsetting harms with proactive efforts towards both redistributing benefits and opportunities, and creating better jobs than those being automated.

Increasing participation would include recognizing the contributions of humans to existing AI systems, ensuring fair and dignified labor conditions across the AI supply chain, and compensating data and labor inputs. This requires regenerating the digital commons rather than enclosing it, and fundamentally rethinking the practice of capturing public data and privatizing the economic benefits from the models that are indebted to it. Benefits should be shared with the communities maintaining the commons, and efforts should be taken to return the resultant technologies as much as possible to the same commons. Broadening the pathway to participation is the most natural way to accomplish this. When data derive from sources specific to a community or group of individuals, these individuals should have collective agency over the development, use, and design of the model, should be publicly recognized as contributing to it, and should share in any commercial benefits. In a wider sense, greater participation means balancing the optimization of fixed, measurable goals with investment in broad-based stakeholder reflection on whether these goals are appropriate ones. This means moving from thin representations to richer and layered representations that incorporate deliberative technologies, collaborative design, and the fusion of policy change with technology development.

Finally, supporting mutualism would require directly addressing the impacts of technology. Rather than relying on policy makers or civil society to fix the problems technology creates or to balance highly skewed incentives, clear mechanisms must be established to account for the impact of a technology on the distribution of economic and political power, measuring it and reporting it along with other ESG (environment, social, and governance) metrics. This should also include expanding notions of robustness, security, and transparency to include metrics around the likely concentrating effect of technology projects across dimensions such as wealth, resources, decision-making ability, access, etc., but also cross-cutting categories such as geography and class. By emphasizing the range of meaningfully different directions technology may and should take, as well as the history of differing approaches, contingent successes, and failed predictions, there can be more awareness about the contingency of technology development. This includes developing clear mechanisms to assess outcomes, maintenance, and depreciation of a system. Above all mutualism means embracing complexity and uncertainty—turning away from the top-down engineering of ecosystems, cities, economies, and societies, and toward fostering the collective intelligence of diverse and interacting public communities.

AEDP is a pluralist vision for the future of technology—in contrast to the dominant vision of AEAI. As we note above, some of these practices are already in place in some organizations some of the time; we do not wish to accuse most AI researchers of violating most principles most of the time. Yet there is a long way to go before the centralization goals of AEAI can be productively redirected—and that work must begin now. We must reject the dominant focus on building autonomous, human-like intelligence in preference for exploring a range of other possibilities that complement humans and human societies and facilitate participative cooperation. By doing so we will break down the centralizing tendencies of AEAI with its reactive and proactive dangers, and move via an embrace of digital pluralism—and complementarity, participation, and mutualism—toward greater collective flourishing.

AUTHORS

Divya Siddarth
RadicalXChange Foundation and Microsoft
Daron Acemoglu
Massachusetts Institue of Technology
Danielle Allen
Harvard University
Kate Crawford
University of Southern California and Microsoft Research
James Evans
University of Chicago
Michael Jordan
University of California, Berkeley
E. Glen Weyl
RadicalXChange Foundation and Microsoft

ACKNOWLEDGEMENTS

Ted Chiang
Tantum Collins
Andrew Critch
Allan Dafoe
Stephanie Dick
Peter Hart
Katya Klinova
Margaret Levi
Vishal Maini
Nahema Marchal
Heidi Mayer
Audrey Tang
John Tasioulas
Andrew Trask
Kinney Zalesne

SOURCES

Acemoglu, Daron, and Pascual Restrepo. 2019. "Automation and New Tasks: How Technology Displaces and Reinstates Labor." *Journal of Economic Perspectives* 33 (2): 3–30.

———. 2020. "Robots and Jobs: Evidence from US Labor Markets." *Journal of Political Economy* 128 (6): 2188–2244.

———. 2021. "Tasks, Automation, and the Rise in US Wage Inequality." Working Paper 28920. Cambridge, MA: National Bureau of Economic Research.

Altman, Sam. 2021. "Moore's Law for Everything." https://moores.samaltman.com/

Ashby, W. Ross. 1956. *An Introduction to Cybernetics*. London: Chapman & Hall Ltd.

Atanasoski, Neda, and Kalindi Vora. 2019. *Surrogate Humanity: Race, Robots, and the Politics of Technological Futures*. Durham, NC: Duke University Press.

Babbage, Charles. 1832. *On the Economy of Machinery and Manufactures*. London: C. Knight.

Barrat, James. 2013. *Our Final Invention: Artificial Intelligence and the End of the Human Era*. New York: Macmillan.

Benjamin, Ruha. 2019. *Race after Technology: Abolitionist Tools for the New Jim Code*. Cambridge: Polity Press.

Bockman, Johanna. 2020. *Markets in the Name of Socialism The Left-Wing Origins of Neoliberalism*. Stanford, CA: Stanford University Press.

Bostrom, Nick. 1998. "How Long before Superintelligence?" *International Journal of Futures Studies* 2 (1998).

———. 2003. "Ethical Issues in Advanced Artificial Intelligence." In *Science Fiction and Philosophy: From Time Travel to Superintelligence*, edited by Susan Schneider, 277–84. Chichester, UK: John Wiley & Sons.

———. 2012. "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents." *Minds and Machines* 22 (2): 71–85.

Boyd, Danah M., and Nicole B. Ellison. 2007. "Social Network Sites: Definition, History, and Scholarship." *Journal of Computer-Mediated Communication* 13 (1): 210–30.

———. 2010. "Social Network Sites: Definition, History, and Scholarship." *IEEE Engineering Management Review* 38 (3): 16–31.

Brin, Sergey, and Lawrence Page. 1998. "The Anatomy of a Large-Scale Hypertextual Web Search Engine." *Computer Networks and ISDN Systems* 30(1–7): 107–17.

Brock, André, Jr. 2020. *Distributed Blackness*. New York: New York University Press, 2020.

Broussard, Meredith. 2018. *Artificial Unintelligence: How Computers Misunderstand the World*. Cambridge, MA: MIT Press, 2018.

Budish, Eric, and Judd B. Kessler. 2016. "Bringing Real Market Participants' Real Preferences into the Lab: An Experiment That Changed the Course Allocation Mechanism at Wharton." Working Paper 22448. Cambridge, MA: National Bureau of Economic Research.

Citron, Danielle Keats. 2008. "Technological Due Process." *Washington University Law Review* 85 (6): 1249–1313.

Citron, Danielle Keats, and Frank Pasquale. 2014. "The Scored Society: Due Process for Automated Predictions." *Washington Law Review* 89 (1): 1–33.

Collins, Harry M. 1990. *Artificial Experts: Social Knowledge and Intelligent Machines*. Cambridge, MA: MIT Press.

Crawford, Kate. 2021. *The Atlas of A*I. New Haven, CT: Yale University Press.

Crawford, Kate, and Jason Schultz. 2014. "Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms." *Boston College Law Review* 55 (1): 93–128.

Daston, Lorraine. 1994. "Enlightenment Calculations." *Critical Inquiry* 21 (1): 182–202.

De Filippi, Primavera. 2014. "Bitcoin: A Regulatory Nightmare to a Libertarian Dream." *Internet Policy Review* 3 (2).

Dear, Brian. 2017. *The Friendly Orange Glow: The Story of the PLATO System and the Dawn of Cyberculture*. New York: Pantheon.

Dewey, John. 1939. *The Philosophy of John Dewey*. Library of Living Philosophers 1. New York: Tudor.

Dick, Stephanie A. 2015. "Of Models and Machines: Implementing Bounded Rationality." *Isis* 106(3): 623–34.

———. 2020. "Coded Conduct: Making MACSYMA Users and the Automation of Mathematics." *BJHS Themes* 5: 205–24.

———. Forthcoming. “The Marxist in the Machine.” In *Beyond Code and Craft*, special issue of *Osiris*.

Drexler, K. Eric. 2019. “Reframing Superintelligence.” In *Future of Humanity Institute*. Oxford: University of Oxford.

Duggan, Lisa. 2019. *Mean Girl: Ayn Rand and the Culture of Greed*. Oakland: University of California Press.

Eckersley, Peter, Yomna Nasser, et al. 2017. “EFF AI Progress Measurement Project.” Electronic Frontier Foundation. https://www.eff.org/ai/metrics.

Engelbart, D. 1968. *Mother of All Demos*. Video, December 9, 1968. Available online at https://www.youtube.com/watch?v=fhE-h3tEL1V4 and https://vimeo.com/69647667

Eubanks, Virginia. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin’s Press.

Forsythe, Diana. 2001. *Studying Those Who Study Us: An Anthropologist in the World of Artificial Intelligence*. Stanford CA: Stanford University Press.

Gordon, Robert J. 2014. “The Demise of US Economic Growth: Restatement, Rebuttal, and Reflections.” Working Paper 19895. Cambridge, MA: National Bureau of Economic Research.

Gray, Mary L., and Siddharth Suri. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Boston: Houghton Mifflin Harcourt.

Hart, Vi. 2019. “Changing My Mind about AI, Universal Basic Income, and the Value of Data: The Art of Research.” https://theartof-research.org/ai-ubi-and-data/

Henrich, Joseph, and Richard McElreath. 2003. “The Evolution of Cultural Evolution.” *Evolutionary Anthropology: Issues, News, and Reviews: Issues, News, and Reviews* 12 (3): 123–35.

Hicks, Mar. 2017. *Programmed Inequality: How Britain Discarded Women Technologists and Lost Its Edge in Computing*. Cambridge, MA: MIT Press.

Ho, Chi-Chang, Ling-Jyh Chen, and Jing-Shiang Hwang. 2020. “Estimating Ground-Level PM2.5 Levels in Taiwan Using Data from Air Quality Monitoring Stations and High Coverage of Microsensors.” *Environmental Pollution* 264: 114810.

Irani, Lilly. 2015. “The Cultural Work of Microwork.” *New Media & Society* 17 (5): 720–39.

Jiménez-Hernández, Diego, and Enrique Seira. 2021. “Should the Government Sell You Goods? Evidence from the Milk Market in Mexico.” Stanford University Working Paper.

Keynes, John Maynard. 1936. *The General Theory of Employment, Interest, and Money*. London: Macmillan.

Kurzweil, Ray. 2000. *The Age of Spiritual Machines: When Computers Exceed Human Intelligence*. New York: Penguin.

Lei, Ya-Wen. 2021. “Upgrading China Through Automation: Manufacturers, Workers and the Techno-Developmental State.” *Work, Employment and Society*, Forthcoming.

Lewis, Jason Edward, Noelani Arista, Archer Pechawis, and Suzanne Kite. 2018. “Making Kin with the Machines.” *Journal of Design and Science* 3.5.

Licklider, Joseph C. R. 1960. “Man-Computer Symbiosis.” *IRE Transactions on Human Factors in Electronics* 1: 4–11.

Lih, Andrew. 2009. *The Wikipedia Revolution: How a Bunch of Nobodies Created the World’s Greatest Encyclopedia*. New York: Hachette Books.

Markoff, John. 2016. *Machines of Loving Grace: The Quest for Common Ground between Humans and Robots*. New York: Ecco/HarperCollins.

Mercier, Hugo, and Dan Sperber. 2017. *The Enigma of Reason*. Cambridge, MA: Harvard University Press.

Milgram, Stanley. 1967. “The Small World Problem.” *Psychology Today* 2 (1): 60–67.

Mindell, David A. 2015. *Our Robots, Ourselves: Robotics and the Myths of Autonomy*. New York: Viking Adult.

Mohamed, Shakir, Marie-Therese Png, and William Isaac. 2020. “Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence.” *Philosophy & Technology* 33 (4): 659–84.

Moses, Joel. 2012. “Macsyma: A Personal History.” *Journal of Symbolic Computation* 47 (2): 123–30.

Muller, Jerry Z. 2018. *The Tyranny of Metrics*. Princeton, NJ: Princeton University Press.

Nilsson, Nils J. 2009. *The Quest for Artificial Intelligence*. Cambridge: Cambridge University Press.

Noble, Safiya Umoja. 2018. *Algorithms of Oppression*. New York: New York University Press.

O’Keefe, Cullen, Peter Cihon, Ben Garfinkel, Carrick Flynn, Jade Leung, and Allan Dafoe. 2020. “The Windfall Clause: Distributing the Benefits of AI for the Common Good." In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 327–31. New York: Association for Computing Machinery.

O'Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* New York: Crown.

Orr, Julian E. 1996. *Talking about Machines: An Ethnography of a Modern Job*. Ithaca, NY: Cornell University Press.

Penn, Jonathan. 2020. *Inventing Intelligence: On the History of Complex Information Processing and Artificial Intelligence in the United States in the Mid-Twentieth Century*. Cambridge: University of Cambridge.

Pérotin, Virginie. 2016. "What Do We Really Know about Workers' Co-Operatives?" In *Mainstreaming Co-Operation: An Alternative for the Twenty-First Century*, edited by Anthony Webster, Linda Shaw, and Rachael Vorberg-Rugh, 239–60. Manchester, UK: Manchester University Press.

Rankin, Joy Lisi. 2018. *A People's History of Computing in the United States*. Cambridge, MA: Harvard University Press.

Ross, Joel, Lilly Irani, M. Six Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. "Who Are the Crowdworkers? Shifting Demographics in Mechanical Turk." In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*, 2863–72. New York: Association for Computing Machinery.

Russell, Stuart. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Penguin.

Siddarth, Divya. 2021. "Taiwan: Grassroots Digital Democracy That Works." RadicalXChange (blog). https://www.radicalxchange.org/media/papers/Taiwan_Grassroots_Digital_Democracy_That_Works_V1_DIGITAL_.pdf

Silver, David, Satinder Singh, Doina Precup, and Richard S. Sutton. 2021. "Reward Is Enough." *Artificial Intelligence*, 299: 103535.

Stanford HAI. 2021. "The AI Index Report–Artificial Intelligence Index." https://aiindex.stanford.edu/report/

Stray, Jonathan, Ivan Vendrov, Jeremy Nixon, Steven Adler, and Dylan Hadfield-Menell. 2021. "What Are You Optimizing For? Aligning Recommender Systems with Human Values." *ArXiv Preprint ArXiv:2107.10939*.

Suchman, Lucy A. 1987. *Plans and Situated Actions: The Problem of Human-Machine Communication*. Cambridge: Cambridge University Press.

———. 2007. *Human-Machine Reconfigurations: Plans and Situated Actions*. Cambridge: Cambridge University Press.

Suchman, Lucy, Jeanette Blomberg, Julian E. Orr, and Randall Trigg. 1999. "Reconstructing Technologies as Social Practice." *American Behavioral Scientist* 43 (3): 392–408.

Tang, Audrey, and Tristan Harriss. 2020. "Digital Democracy Is Within Reach." Your Undivided Attention, posted July 23, 2020. https://your-undivided-attention.simplecast.com/episodes/the-listening-society-yZ1PBlPF

Vincent, Nicholas, and Brent Hecht. 2021. "A Deeper Investigation of the Importance of Wikipedia Links to Search Engine Results." *Proceedings of the ACM on Human-Computer Interaction* 5 (CSCW1): 1–15.

Waldrop, M. Mitchell. 2002. *The Dream Machine: J. C. R. Licklider and the Revolution that Made Computing Personal*. New York: Viking Penguin.

Wang, Hao. 1990. "Computer Theorem Proving and Artificial Intelligence. In *Computation, Logic, Philosophy: A Collection of Essays*, 63–75. Berlin: Springer.

Weizenbaum, Joseph. 1976. *Computer Power and Human Reason: From Judgment to Calculation*. San Francisco: W. H. Freeman & Co.

Weyl, Glen. 2019. "Why I Am Not a Technocrat." RadicalxChange (blog), posted August 19, 2019. https://www.radicalxchange.org/media/blog/2019-08-19-bv61r6/

Whittaker, Meredith, Kate Crawford, Roel Dobbe, Genevieve Fried, Elizabeth Kaziunas, Varoon Mathur, Sarah Mysers West, Rashida Richardson, Jason Schultz, and Oscar Schwartz. 2018. *AI Now Report 2018*. New York: AI Now Institute at New York University.

Wiener, Norbert. 1954. *The Human Use of Human Beings*. London: Free Association.

Wilder, Bryan, Eric Horvitz, and Ece Kamar. 2020. "Learning to Complement Humans. *ArXiv Preprint ArXiv:2005.00582*.

Winograd, Terry. 1992. "Computers and Rationality: The Myths and Realities." In Minds, Brains, and Computers: Perspectives in *Cognitive Science and Artificial Intelligence*, 152–67. Edited by Ralph Morelli, W. Miller Brown, Dina Anselmi, Karl Haberlandt, and Dan Lloyd. Norwood, NJ: Ablex Publishing Corporation.

Yang, Andrew. 2020. "The Freedom Dividend." Yang2020: Andrew Yang for President. https://2020.yang2020.com/policies/the-freedom-dividend/