Governing AI to Advance Shared Prosperity

By Katya Klinova, Partnership on AI

DRAFT COPY

This is a draft of a chapter that was prepared for the forthcoming Oxford Handbook of AI Governance.

I am thankful to Anton Korinek, Daron Acemoglu, and Sophia Nevle Levoy for their comments on the drafts of this chapter. Financial support of the AI and Shared Prosperity Initiative (*http://partnershiponai.org/shared-prosperity*) by the Ford Foundation is gratefully acknowledged.

katya@partnershiponai.org

Abstract

I describe a governance approach to promoting AI research and development that creates jobs and advances shared prosperity. Concerns over the labor-saving focus of AI advancement are shared by a growing number of economists, technologists and policymakers around the world. They warn about the risk of AI entrenching poverty and inequality globally. Yet, translating those concerns into proactive governance interventions that would steer AI away from generating excessive levels of automation remains difficult and largely unattempted. Key causes of this difficulty arise from two types of sources: (1) insufficiently deep understanding of the full composition of factors giving AI R&D its present emphasis on labor-saving applications; and (2) lack of tools and processes that would enable AI innovators and policymakers to anticipate and assess the impact of AI technologies on employment, wages and job quality. I argue that addressing (2) will require creating worker-participatory means of differentiating between genuinely worker-benefiting AI and worker-displacing or worker-exploiting AI. To contribute to tackling (1), I review AI innovators' motivations and constraints, such as relevant laws, market incentives, as well as less tangible but still highly influential constraining and motivating factors, including explicit and implicit norms in the AI field, visions of future societal order popular among the field's members and ways that AI innovators define goals worth pursuing and measure success. I highlight how each of these factors contributes meaningfully to giving AI advancement its excessive labor-saving emphasis and describe opportunities for governance interventions that could correct that over emphasis.

Keywords: steering AI, automaton, inequality, job displacement, future of work.

1. Introduction

What governance levers can help to ensure that AI does not lower the demand for human labor, cutting off large swaths of the global population from their only source of income? This chapter examines the forces that currently give AI R&D its labor-replacing focus. I analyze the monetary and non-monetary interests pursued by AI innovators and the constraints their actions are subjected to, categorized into four modalities described by Lessig (1999): those arising from legislation, market conditions, social norms and "built architectures"—implicit and explicit codes by which the AI field operates. This chapter reviews the literature on the current state of each of these four sources of constraints as they apply to the field of AI, identifying gaps that stand in the way of operationalizing theoretical ideas about steering AI advancement towards inclusive economic outcomes. It argues that constructing worker-participatory ways to distinguish between economically sustainable and unsustainable AI investments is a pre-condition for enabling effective governance of AI in service of shared prosperity.

Why use governance to deliberately ensure AI does not diminish global labor demand? An alternative would be to let AI advancement stay on its current trajectory, which, as a growing number of leading economists and technologists agree, is likely to generate a large-scale redistribution of economic and political power towards a concentrated group of winners—a small handful of countries, firms, and individuals—eliminating jobs or lowering wages for an ever-growing share of the population (see, for example, Acemoglu and Restrepo 2019 and 2020, Korinek and Stiglitz 2019, Altman 2021). At present, a globally inclusive system for taxing and redistributing income and ownership of productive capacities does not exist. The political process necessary to enable the creation of such a system, if this process takes place at all, is very likely to lag behind the pace of technological change, resulting in large groups of the population getting left behind, which would undermine social stability and public trust in AI progress. Figuring out how to enable the beneficial advancement of AI while protecting and expanding access to good jobs is therefore a pressing governance challenge. Rising to this challenge does not require imposing a ban on labor automation—the composition of available jobs can continue to evolve, with some jobs getting automated as long as new and better jobs replace them and are not associated with insurmountable skill barriers. An abundance of well-paying, secure and dignified jobs will be necessary at least until humankind has figured out how to robustly and on a global scale decouple the elimination of jobs from the elimination of dignity, status, and access to income.

Is it possible to steer AI away from labor displacement? The view presenting technological progress as an unalterable and unavoidable march of automation is fairly common, but it overlooks a key consideration: the direction of technological change is a function of AI innovators' choices made in pursuit of their interests, subject to constraints imposed—or not imposed—on them by the market, legislation and the norms and structural features of the field. This chapter refers to "AI innovators" broadly (including private and public sector researchers and engineers, entrepreneurs, venture capitalists, and corporations), and systematically examines the incentives and constraints that influence these actors' decisions around investing in and developing labor-saving AI.

For examining the constraints AI innovators are subjected to, the chapter uses Lawrence Lessig's pathetic dot theory as an organizing framework (Lessig 1999). The theory proposes that all actions are constrained, or regulated, by four interacting forces: applicable legislation, the market, social norms and "architecture." "Architecture" refers to "the way the world is" or the "built environment" where the activity of interest is performed—in the case of this chapter, the focus is on the "built-in" features of the AI development field. The four modalities of constraints interact and influence each other: the built environment can change as a result of legislation, and a change in legislation can be prompted by evolving social norms and expectations if those result in mounting political pressure. This chapter will specifically examine the influence of regulatory policy on market constraints and the impact of social norms and shared views in the AI field on defining its architecture and the ways things are done, highlighting the relevance of both of these types of interactions for the resulting direction of AI progress.

In Lessig's framework, legislation and the market are the first two forces that regulate any activity. AI advancement is no exception: the legislative and regulatory environment in which AI innovators operate is determined by the applicable laws and policies which shape the market and influence the relative profitability of investing in various types of AI applications. Many policy decisions which superficially do not appear to direct AI R&D can either boost or reduce the incentives of AI developers to create labor-saving technology. For example, low interest rates make the investment in machines and software more appealing encouraging excessive automation (Stiglitz 2014 as cited in Schindler, Korinek, Stiglitz 2021), while tax regimes heavily favoring capital over labor can make it difficult to justify an investment in labor force expansion over automation even when producing the same output with machines is more expensive pre-tax (Acemoglu, Manera, Restrepo 2020). Labor mobility restrictions limit labor supply and boost the incentives to develop labor-saving AI, generating little-studied cross-border spill overs of automation technology from countries with aging populations and restrictive immigration policies to countries struggling to produce sufficient numbers of formal sector jobs for their young and growing labor forces (Pritchett 2020). But successfully re-orienting policies to steer AI advancement towards an economically inclusive trajectory requires not only a recognition of the labor demand-depressing side effects of current policies but also a construction of a better understanding of how to practically distinguish labor-friendly AI from inequality-producing AI.

The project of governing the direction of AI in service of shared prosperity also needs to pay attention to something more subtle and much less well described than regulatory impacts on automation incentives. The chapter argues that the structural features of the AI field, which include what Lessig defines as norms and architecture, bias its focus towards labor-saving applications. AI innovators often split their time between the public and private sector, but even those who work only in the private sector are driven not only by the profit motive. They are also influenced by the field's features—a web of its norms, shared aspirations, commonly accepted ways to measure progress, and ideas about desirable future for the human society. “The field’s goal has always been to create human-level or superhuman AI,” states Stuart Russell, one of the preeminent AI scientists, in the opening of his latest book (Russell, 2019). The structure of the AI field orients towards this explicit goal of matching and exceeding human performance on all basic capabilities. In addition, tech elites’ dominant political view favors redistribution, which is seen as a solution to an automation-induced loss of labor income, but strongly disfavors government regulation of technology (Broockman 2019). This conjunction of the fields’ goals and political beliefs fuels the labor-saving bias of AI advancement in a multitude of direct and indirect ways.

The rest of this chapter proceeds as follows. First, it reviews the economic literature that raises concerns over AI’s current direction and describes the practical difficulties around differentiating labor-saving and labor-complementing technologies for legislative or market-regulating purposes, exacerbated by the absence of empowered participation by workers in the AI development process. Next, the chapter reviews the current objectives of the leading AI labs in an attempt to understand which socio-technical factors (or, using a term from Jasanoff and Kim, 2015, “imaginaries”) might have shaped those objectives, and how they can be altered by governance mechanisms in service of re-orienting the field towards producing shared prosperity-compatible outputs. Section 4 concludes.

2. Policy, market incentives, and the difficulty of defining labor-friendly AI

The possibility of using AI for automating human labor does not make AI advancement non-labor-friendly by definition. Automation is a phenomenon that long predates AI and that has enabled a broad-based rise in productivity and living standards. What is different about the wave of labor automation enabled by AI is not only that AI allows for a dramatic expansion in automation possibilities, potentially leading to a meaningful acceleration in the pace of automation, but also that AI advancement is happening against the backdrop of rising inequality and worsening labor market outcomes for an ever-increasing share of the population—trends that AI, on its current trajectory, is poised to exacerbate.

In the last four decades, growth in the advanced economies has not been inclusive: labor shares of national incomes have steadily declined, wages of workers without college degrees have stagnated or even declined in real terms, the number of middle-paying jobs has decreased, giving rise to the gap between average and median worker compensation (Autor, Mindell, Reynolds 2020). Acemoglu and Restrepo (2021) document robust evidence suggesting that at least in the case of the US, the majority of the change in the economy’s wage structure is explained by the relative wage declines of workers in industries that experienced rapid automation. In their earlier work, the authors show that unlike in the four decades following WWII, when the volume of waged tasks automated every year was matched by newly created waged tasks for humans, in the last three decades in the US, the pace of automation has measurably accelerated and consistently eclipses the pace with which new waged tasks appear. This has contributed to a decline in labor demand and compensation for workers in automation-exposed occupations (Acemoglu, Restrepo 2019).

AI-induced technological change, like any other type of technological change, can be expected to bring about two effects: an increase in the overall economic output and a redistribution of income between factors of production, where factors of production usually include labor, capital, and sometimes land. Labor can be further disaggregated into groups of workers, for example, by skill level. If labor receives at least some of the gains from technologically induced output increase, technological change is referred to as labor-using, and in the opposite scenario as labor-saving. For example, technological change that powers labor-displacing automation but does not create a compensatory labor demand would be labor-saving. Following Korinek, Schindler, and Stiglitz (2021), this chapter uses the terms “direction,” “focus” and “orientation” of the AI field as a reference to whether the technological change it generates is labor-saving or labor-using.

Further, if labor benefits from technological change relatively more than the other factors of production, technological change is deemed biased in favor of labor. Of course, not all technological change is biased in favor of labor—it can often be biased in favor of capital, or only certain kinds of labor. Technological change can complement certain skill groups and raise their gains while reducing the demand for other skill groups, in which case it is deemed skill-biased. For example, the proliferation of personal computers and the development of word processing and planning software tools complemented the skills of people in managerial and other occupations but reduced the demand for typists and secretaries. Skill-biased change is likely to deepen economic inequality.

While technical change is not the only factor that influenced real wage stagnation and wage declines for workers with lower levels of formal educational attainment in recent decades, the persistence of this trend coupled with accelerating automation of tasks associated with middle-playing jobs suggest that in its latest form technical change has been biased against formally lower-skilled groups. To ensure that this inequality-deepening trend is not accelerated by the advancement of AI, leading economists have been calling to create AI technologies that are labor complementary and not labor-saving, since the latter kind would decrease the demand for human labor, leading to a reduction in either employment, wage levels, quality of jobs, or some combination of all of the above (Korinek and Stiglitz 2019, Acemoglu and Restrepo 2020).

An increase in the demand for human labor is generally associated with an increase in employment, wages and improved working conditions, so technologies that lead to an increase in labor demand would benefit labor, at least in absolute terms, if not relative to other factors of production. But practically executing on the recommendation to develop labor demand-boosting AI is difficult because the effect of new technology on labor demand is highly uncertain ex-ante due to the presence of multiple second-order effects, possible variation in deployment contexts, a variety of ways basic research can be used in final applications and unknowable counter-factual scenarios of technological advancement.

Let us examine each of these sources of uncertainty in turn. Second-order effects of technical change refer to its indirect impacts: outside of jobs directly created or cut by a company that develops or introduces a new technology, often there are jobs created or cut (or wages increased or reduced) elsewhere in the economy, for example, by the company’s suppliers, clients, or competitors (see a detailed categorization of indirect effects in Klinova and Korinek, 2021). If the company's introduction of new technology creates productivity gains and those are passed on to consumers in some form, for example as reduced prices, improved quality, or new products, that can free up consumers' income to be spent on other goods in the economy, creating new labor demand in corresponding sectors. Productivity gain does not always get fully passed on to consumers, especially in monopolistic markets in which leading AI companies tend to operate. Acemoglu and Restrepo (2019) also warn of the recent proliferation of what they refer to as "so-so" technologies—those that displace human labor but do not create a meaningful productivity gain, thus failing to give rise to a compensatory labor demand elsewhere in the economy.

Early-stage AI research can enable a wide range of applications down the road, including both those biased in favor of and against labor. This adds to the difficulty of practically steering AI advancement in service of increasing employability of economically vulnerable workers. Moreover, the same high-level application can be used to replace or augment human labor depending on the deployment context: for example, self-driving car technology can displace workers on ordinary roads, but can also allow humans to reach places where human drivers would not be able to go due to, for example, harsh or dangerous conditions. Lastly, evaluating a tentative impact of new technology on labor demand based on a comparison with the status quo can be misleading, because such comparison does not take into account the impact of technologies the development of which would happen thanks to the creation of the new technology in question, nor the impact of alternative technologies that could be developed by people who are busy with the creation of the technology in question.

Despite the presence of all these uncertainties about the impact of AI applications on labor demand, AI companies have begun to pick up on the growing concern around labor-saving AI and increasingly describe their products as labor-augmenting. The meaning ascribed to the term "human-augmenting AI" is frequently vague, suggestive of AI that "helps" or "assists" workers by, for example, making them more productive or reducing the number of workplace accidents. This broadly matches the economic definition of labor-augmenting technologies as technologies that increase the marginal product of labor. However, "worker-augmenting" AI does not guarantee that workers would receive any gains as a result of its introduction and does not ensure that workers would not be made worse off. A prime example of technologies that are often positioned as worker assisting, productivity increasing and safety improving are workplace surveillance solutions, usually described by their producers as tools for "worker productivity monitoring."

As documented by Nguyen (2021), workplace monitoring technologies enable work speed-ups and intensification, the creation of excessively punitive work environments and the shifting of risks from employer to employee. For example, flexible algorithmic scheduling allows employers to offload the risk of a slow day onto workers whose shifts are scheduled or cancelled last minute based on fluctuating customer demand projections. In the context of the workplace principal-agent problem (where the employer is the principal and the worker is the agent), the principal's imperfect ability to observe the agent's effort is a source of power for the agent that otherwise typically holds very few powers, especially in a non-unionized setting. Imperfect ability to observe an agent's effort gives the employer a motivation to incentivise high performance by treating the agent to dignified working conditions, performance bonuses, etc. When an agent's effort is perfectly observable, the principal has less of an incentive to reward high performance; she can set targets and penalize workers not meeting them (Gerety 2020).

Workers being monitored by assistive AI might or might not be aware that they are training an algorithm for becoming a better substitute for them. Even when workers are fully aware of the training they are participating in, like, for example, human drivers helping self-driving cars navigate difficult and ambiguous situations, they are nearly never recognized as co-creators of the resulting technology (Lanier 2014). They do not hold IP rights for their know-how, do not receive royalties every time the data they generate is used, are generally not granted equity shares and are not allotted the praise and social status of AI innovators.

To avoid a proliferation of technologies that claim to augment workers but in practice enable employer overreach and exploitation and strip workers of privacy and power in the workplace, it is necessary to introduce measurement and disclosure parameters around AI systems' impact on labor. Aside from an analysis of the magnitude of second-order impacts on labor demand referenced above, the disclosures should be required to contain results of independently carried out surveys of workers' experience with the AI system in question, their involvement in decisions around the introduction of AI into the workplace, as well as channels of recourse and contestation available to them. This is necessary to ensure an empowered participation of workers in the design and deployment of AI systems that are poised to affect them, as well as because affected workers are likely to have first-hand insight into whether an AI system in question benefits or hurts them.

In game theoretical terms, today the claim that "our AI augments humans" that is made by a growing number of AI companies amounts only to a cheap signal, and hence it cannot be used by regulators or potential buyers to differentiate AI systems that expand access to good jobs from tools that benefit capital owners but disadvantage workers. Substantiating the claim with tangible and transparently reported metrics around the impact of an organization's AI development and deployment efforts on labor demand and job quality would allow it to meaningfully differentiate itself from others with a credible signal. Organizations genuinely committed to a mission of augmenting and complementing human workers should want their commitment to be seen as credible and thus would be willing to report on their impact on the distribution of good jobs in the economy.

There are efforts currently underway to develop robust ways to measure labor demand impact of a given AI-developing organization (see, for example, Partnership on AI 2021). Such measurement approaches can be used by AI innovators in the private and public sectors to inform their assessment of the economic consequences and social sustainability of new products and services and hopefully to guide their product choices in a prosocial direction. Governments can use such measurement frameworks to inform their R&D investments and industrial policy, as well as introduce a rule to procure AI only from companies that assess and disclose their labor market impact (Seamans 2021). Importantly, governments should also pay attention to how their policies across the board might be incentivizing the development of labor-saving technologies. The next section turns to the interaction between Lessig's first two forces—legislation and markets—to give an overview of how the market incentives faced by the AI industry are shaped by policy choices, biasing the advancement of AI against labor.

Policies incentivizing the development of labor-saving AI

Economic incentives that commercial AI development, including the development of labor-saving AI, is faced with are determined by multiple factors, including the present and expected conditions of the global economy, state of competition and regulation, the demographic situation, and more. Consequently, policy choices of special relevance to shaping the direction of AI development include tax policy, R&D and industrial policy, interest rate policy, government procurement practices, policies around migration and what Korinek, Schindler and Stiglitz (2021) refer to as "rules of the game," or policies that affect the returns on factors of production, such as labor legislation, competition laws, rules regulating corporate governance, etc. All of these can raise or lower the relative commercial appeal of investing in labor-saving technologies. For example, low interest rates stimulate capital investment into technology and equipment and large government orders for machines that substitute for human workers can accelerate the development of such machines, etc. Policies introducing distortions deserve special attention as they can lead to excessively high (or low, depending on the direction of the distortion) levels of automation. This subsection will focus on two policies giving rise to the biggest distortions in the pace of automation: policies around taxation and labor mobility.

In much of the industrialized world labor is taxed more heavily than capital, which makes replacing a higher portion of the workforce with machines financially appealing for businesses, provided that the net present value of technology development costs does not exceed the NPV of the tax payments that will be saved. That said, if taxes are set optimally (which for many countries might mean capital is taxed at a lower rate than labor), the level of automation will also be optimal in absence of other distortions. Acemoglu, Manera and Restrepo (2020) show that at least in the US—a country with outsized influence on AI R&D—the effective tax rate on labor was too high in the 2010s, while the effective tax rate on capital was too low, leading to excessive levels of automation compared to the socially optimal level. They note that even if tax rates were set optimally going forward, it would still be welfare-improving to reduce the resulting equilibrium level of automation because the starting point would be one with already excessive levels of labor-saving technologies brought about by tax distortions present throughout the 2010s.

Since the US plays a dominant role in AI development and because of low marginal costs of deploying AI applications around the world once they have been developed in the US, the country's distorted tax code is effectively "exported" to the rest of the world through excessive levels of automation spilling over to low- and middle-income countries struggling to create a sufficient supply of formal sector jobs for their young and growing labor forces. For example, recent evidence analyzed by Diao et al. (2021) suggests that global technological trends exported to African countries induce local firms to employ capital-intensive technologies, which are inappropriate in the context of those countries' comparative advantage or workforce needs.

The borderless nature of technology deployments prompts a discussion of potential policy-induced incentive distortions not only from the point of view of a single country but also of the entire planet. Pritchett (2020) points out what he describes as "the biggest price distortion ever": the disparity between wages in high- and low-income countries. Clemens, Montenegro and Pritchett (2019) show that labor price differentials between rich and poor countries exceed any current or historic trade tariffs or carbon price distortions by orders of magnitude. This leads to a situation where private sector actors producing technologies in high-income countries face a distorted labor supply curve: the labor supply they effectively respond to does not reflect the global supply of labor, while the technologies they produce do spread globally relatively quickly. As theoretically shown by Acemoglu (2010), labor scarcity encourages strongly labor-saving technological change, which can manifest in lower employment, wages or under-employment domestically. If labor-saving technologies spread across the globe—which they often do—that generates a negative externality for countries struggling to create an adequate provision of good jobs, especially for their youth.

Pritchett (2020) also points out that, aside from giving rise to an excessive spread of labor-replacing technologies, distorted prices of labor can and do lead to the creation of labor-shifting technologies. For example, self-service kiosks widely deployed in restaurants, grocery stores, and airports around the world do not exactly automate the cashier's or check-in agent's jobs; instead, they shift almost entirely the same set of tasks onto the customer, who does not get paid for executing them, and might or might not get any amount of compensating benefit in the form of lower wait time or lower price. In other cases, technology might be partially shifting paid work into unpaid work. For example, calling customer service is increasingly associated with navigating automated self-service voice menus, which can cost the customer a lot more time and frustration with little benefit. Work-shifting applications often fall into the "so-so technologies" category (a term from Acemoglu and Restrepo, 2019): they reduce paid employment but do not compensate for that with any meaningful productivity boost to the economy. Task-based platforms for ride-sharing, delivery, home repairs, or online tasks like data labelling also offer a wealth of examples of jobs being broken down into compensated and uncompensated parts. Workers on those platforms are not paid for their time spent waiting or searching for the next task, for learning how to complete the task, for maintaining and repairing their equipment, let alone for taking a lunch break or a sick day (Gray and Suri, 2019), while all of those activities would be compensated in case of a standard employment contract.

The above discussion suggests that more immigration would benefit both high- and lower-income nations: restricting immigration is a weak strategy for protecting jobs at home. Large and growing wage disparities across the countries incentivize offshoring of jobs to lower-wage countries, shifting of jobs to unpaid work and excessive creation of labor-saving technology that undermines wages and quality of jobs at home and spreads beyond a single country's borders. For example, the US could bring in 160 thousand immigrants to close the truck drivers shortage projected to accumulate by 2028 by the American Trucking Association (Costello, Karickhoff 2019), or it can develop autonomous driving technologies and displace all 3.6 million people employed as truck drivers in the US, and likely many more millions employed in this job abroad. And while autonomous driving might come with important benefits like improved road safety that might make the technology desirable despite the lost jobs, similar dynamics are at play creating distorted incentives for the development and deployment of worker-replacing technologies with much more ambiguous benefits, for example, those replacing nurses (Mani et al. 2021), even though the shortage of care workers could be beneficially reduced by expanding cross-border labor mobility.

Many economists agree with the hypothesis that private sector-driven technological change, responding to societal scarcities as reflected by (non-distorted) market prices, can and has generated incredible innovation which supported the rise in living standards observed over the last two centuries. However, there is no theoretical basis for stating that the market process generates an optimal trajectory for technological change in the long term (Korinek 2019). And if the price signals are distorted not in favor of labor, which they presently are, as has been discussed in this section, the market is likely to deliver excessive automation beyond socially optimal levels, directing AI to economize on false scarcities, eliminating good jobs at a time when much of the world is struggling to generate enough of them. But market incentives distorted by misguided policies are not the only factor fuelling the AI field's dangerous focus on churning out labor-saving and worker-exploiting technologies. The next section will examine the contribution of the other two forces on Lessig's list—social norms and structural features of the environment in which development happens.

3. AI field's orientation, norms and structural features that reinforce it

This section reviews how social norms and architecture of the AI field influence the direction of AI development. AI field's norms arise from a shared understanding of what problems are worth tackling and can bring praise and recognition, explicit and implicit definitions of success and state-of-the-art performance, which this section examines. Architecture, in Lessig's definition, is "the way the world is," the "built environment" in which AI development and deployment happens and its structural features. The architecture of the environment and its norms do interact with each other, just like the legislative and market constraints which were discussed in the previous section, but in the context of the AI field, the norms-architecture interaction is notably tight. Lessig (1999) suggested a useful way to differentiate between them: norms constrain or direct the agent's actions only if the agent is aware of them, while architecture's impacts are not predicated on awareness. For example, in the context of AI, a rare innovator is not aware of Moore's law, but its powerful impact would be experienced by AI innovators with or without that awareness, making it a feature of the "built environment." Similarly, the present homogeneity of the AI field in terms of demographic and socio-economic backgrounds impacts what kind of problems get taken up by the field's actors and what kind of consequences of AI deployments get examined or ignored—that impact takes place whether or not the actors in the field are aware of its homogeneity. Unlike that, an AI entrepreneur who is unaware of the venture capital community's appreciation of exponential user growth is less likely to emphasize it in her business pitch over profit generation, placing this case in the category of adhering to norms and commonly held expectations.

3.1. Governing ideas and norms of the AI field

Definitions of success and choice of problems to tackle

A field's orientation is influenced by which problems are selected to be tackled by its members and how they define progress. Whether the field takes up addressing social or commercial challenges, and what kind of achievements are associated with reputational gains and prestige, influence the impacts the field generates on the society, including its economy and labor market conditions.

In his 1988 Presidential Address to the Association for the Advancement of Artificial Intelligence, Turing Award winner Raj Reddy recounted the following story: "In 1966, when I was at the Stanford AI labs, there used to be a young Ph.D. sitting in front of a graphics display working with Feigenbaum, Lederberg, and members from the Chemistry department attempting to discover molecular structures from mass spectral data. I used to wonder, what on earth are these people doing? Why not work on a real AI problem like chess or language or robotics? What does chemistry have to do with AI?" (Reddy 1988, emphasis mine). "Now we know better," Reddy continued in 1988 before proceeding with a long list of notable applications of AI in various fields. That list grew tremendously since 1988, but the statement remains a testament to the existence of shared notions or norms around what problems are "real" or worth pursuing.

Which problems are considered worthwhile in today's field of AI? One way to ascertain that is by reviewing how the leading AI labs describe their work. Table 1 contains those statements for private and academic labs that were among the top 10 places of affiliation of the authors of papers accepted to the 2020 International Conference on Machine Learning, one of the field's primary conferences (Ivanov 2020).

Among the university-based AI labs mentioned in Table 1, two—MIT CSAIL and CMU—describe themselves with language that could be interpreted as expressing an intent to assist or complement humans, by mentioning building "things that help humans" and pioneering "research in computing that improves the way people work, play, and learn" respectively. Two other academic labs—Princeton Visual AI lab and Stanford AI lab—list goals around improving human-AI collaboration. Among the leading private AI labs, only one (Microsoft Research) declares an explicit intent to strive for human complementarity, describing itself as "[p]ursuing computing advances to create intelligent

Organization	Self-description
Google	[W]e’re conducting research that advances the state-of-the-art in the field, applying AI to products and to new domains, and developing tools to ensure that everyone can access AI. *Source: https://ai.google/about/*
MIT	MIT’s Computer Science and Artificial Intelligence Laboratory pioneers research in computing that improves the way people work, play, and learn. *Source: https://www.csail.mit.edu/*
Stanford University	[Stanford Artificial Intelligence Lab] promotes new discoveries and explores new ways to enhance human-robot interactions through AI. *Source: https://ai.stanford.edu/about/*
UC Berkeley	The Berkeley Artificial Intelligence Research (BAIR) Lab brings together UC Berkeley researchers across the areas of computer vision, machine learning, natural language processing, planning, control, and robotics. *Source: https://bair.berkeley.edu/*
DeepMind	We’re a team of scientists, engineers, machine learning experts and more, working together to advance the state of the art in AI. *Source: https://deepmind.com/about*
Microsoft	Pursuing computing advances to create intelligent machines that complement human reasoning to augment and enrich our experience and competencies. *Source: https://www.microsoft.com/en-us/research/research-area/artificial-intelligence/*
Carnegie Mellon University	CMU has spent decades building a culture where people care about using technology to solve real problems. More than half a century ago, Allen Newell and Herb Simon had a vision for a general problem-solver for the human race. Since then, their vision has become a filter: people attracted to building solutions to real-world problems come here. The result? One of the world’s largest collections of people determined to build things that help humans. *Source: https://ai.cs.cmu.edu/about*
Princeton University	We work on developing artificially intelligent systems that are able to reason about the visual world. Our research brings together the fields of computer vision, machine learning, human-computer interaction as well as fairness, accountability and transparency. *Source: https://visualai.princeton.edu/*
Facebook	We’re advancing the state-of-the-art in artificial intelligence through fundamental and applied research in open collaboration with the community. *Source: https://ai.facebook.com/research*
University of California Los Angeles	The StarAI lab [Statistical and Relational Artificial Intelligence Lab] performs research on Machine Learning (Statistical Relational Learning, Tractable Learning), Knowledge Representation and Reasoning (Graphical Models, Lifted Probabilistic Inference, Knowledge Compilation), Applications of Probabilistic Reasoning and Learning (Probabilistic Programming, Probabilistic Databases), and Artificial Intelligence in general. *Source: http://starai.cs.ucla.edu/*

Table 1. Self-descriptions of the top 10 AI labs by a number of affiliated authors whose papers were accepted to ICML 2020 as identified by Ivanov (2020). Source: collected by the author from organizations’ websites (accessed on June 25, 2021).

machines that complement human reasoning to augment and enrich our experience and competencies.” While that description currently stands in stark contrast with those that explicitly or implicitly declare goals around automating human activities starting with basic abilities, human augmentation language alone is too vague to serve as a strong signal of, or commitment to, a qualitatively different path of AI development, as was discussed in Section 2.

A few of the self-descriptions listed in Table 1 explicitly mention advancing the state-of-the-art in the field of AI. What is considered to be state-of-the-art performance in AI? The “Technical Performance” chapter of the most recent AI Index report (2021) gives a detailed picture: it describes the progress in all of today’s main subfields of AI, namely computer vision, language, speech, concept learning, and theorem proving. In other words, advancing the state-of-the-art in AI means improving machines’ ability to imitate basic human capabilities: to see, speak, hear, write, and reason. In these subfields, performance benchmarks are designed to enable an explicit comparison with human performance. The following subsection gives an overview of key benchmarks.

Orienting benchmarks of the AI field

Benchmark datasets specify the goals the AI research and development community optimizes their work for. Large high-quality datasets are difficult and costly to compile anew, which ensures the enduring popularity of existing publicly available ones. Their use is further incentivized by associated prizes and challenges, some of which happen annually and award not only sizable monetary amounts but the status of the field’s leader. Chasing state-of-the-art performance on benchmark datasets has emerged as a common goal in the subfields that constitute today’s AI R&D (Raji et al. 2020). Examining key benchmarks is therefore instructive for understanding what the AI field is presently building towards and what it defines as success.

This subsection reviews benchmarks used to assess progress in today’s key subfields of AI: vision, speech, language understanding, and reasoning. The benchmarks listed below—ImageNet, SuperGLUE, LibriSpeech and VCR—are among those that the AI Index report (2021) references to describe the state of AI’s technical performance.

ImageNet is an image database of over 14 million labelled images, which, according to the ImageNet website, contains “hundreds and thousands” of image examples for each noun in the English language. The dataset is available for free for non-commercial use and is used to train object detection and image classification algorithms. The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) evaluates the performance of such algorithms and measures the progress of computer vision as a field (Russakovsky et al. 2015). Two metrics are most commonly reported on: Top-1 Accuracy, or the percentage of times when an image label predicted by an algorithm matches the correct label indicated in ImageNet, and Top 5 accuracy, or the percentage of times a correct label is found among algorithm’s 5 best predictions. Top 1 accuracy has grown from just over 60% in 2013 to 86.5% in 2021, Top 5 accuracy has grown from just under 85% in 2013 to 97.9% in 2021, “beating” human performance.

SuperGLUE is a benchmark for evaluating the progress of AI systems in English language understanding. It was introduced in 2019 as a more challenging version of the GLUE benchmark after the performance on GLUE had surpassed the level of non-expert humans. SuperGLUE assigns a single score summarizing the progress in eight language-understanding tasks, each of which comes with publicly available standardized training, development and test datasets (Wang et al. 2019). Best performing models are listed on the SuperGLUE leaderboard. At the time of writing, the leaderboard is topped by a team with a score of 90.4, which is just above SuperGLUE's "human baseline" of 89.8.

LibriSpeech is a dataset consisting of over a thousand hours of public domain audiobooks and is used to train and evaluate speech recognition systems (Panayotov 2015). Similarly to the benchmarks discussed above, LibriSpeech has a corresponding leaderboard that ranks models by how well they perform on the LibriSpeech test set. Models with the lowest transcribed word error rate at the top. As of April 2, 2021, the best-performing models get 1.4% of words wrong in a controlled environment and 2.6% wrong in a noisy environment.

Visual Commonsense Reasoning (VCR) is a dataset consisting of 290 thousand question-answer problems derived from 100 thousand movie scenes. It is used to evaluate how well ML models are able to imitate a basic human ability to discern the context and situation from an image. For each test image and a corresponding question about what is happening in the image, an algorithm is expected to choose a correct answer from a selection of 4 and pick a correct rationale from a selection of 4. For example, a question about an image with two people and a basket of money in front of them is "How did person 2 get the money that's in front of her?" The correct answer is: "Person 2 earned this money playing music," and the correct rationale is: "Person 1 and Person 2 are both holding musical instruments and were probably busking for that money" (Zellers 2019). According to the VCR Leaderboard, no ML model has so far crossed the "human performance" benchmark at 85 points. That said, improvement in model performance has been impressive: the best model's score has gone from 44 in 2018 to 70.8 in 2020 (AI Index 2021).

All the benchmarks described above unambiguously map to the goals of automating or imitating basic human abilities. Corresponding leaderboards explicitly recognize which AI models managed to reach or cross a "human baseline." This normalizes and incentivizes a "competition" between humans and AI models in which the world's best AI researchers chase the goal of creating systems that excel in tasks which most of the 7.8 billion people living today can perform and use every day to earn a living. Of course, one could argue that it is possible to construct an AI system that leverages computer vision, speech recognition and other technologies to complement human workers, raise their productivity and make them more valuable to the labor market and not to displace them from jobs. But that will not be achieved without the field exchanging its current goals—focused on beating people at their basic abilities— for goals aimed at increasing peoples' productivity (Siddarth et al. forthcoming). I will turn next to a discussion of why goals in the AI field are being set the way they are, and what alternative benchmarks have been suggested.

The tradition of evaluating progress in AI in terms of matching and "beating" human performance on basic tasks is a long-standing one: in his 1988 Presidential Address to AAAI mentioned above, Raj Reddy notes that at any given point, the AI field's accomplishments can be measured by "assessing the capabilities of computers over the same range of tasks over which humans exhibit intelligence." Mani et. al (2021), reviewing the substantial progress made by the field of AI since 1988 on the grand challenges posed by Raj Reddy in his presidential address (such as building a world champion chess machine and an accident-avoiding car), provide examples of different goals the AI field could set for itself. For instance, Frank Chen proposes to shift the focus of the goal of AI research away from "surpassing" humans and towards "human + AI = better together" vision, combining the strengths of machines and humans. He suggests judging the progress based on the ability of human+AI teams to perform better than either humans or machines on their own. In the same volume, Steve Cross proposes a "Reddy test" for teams, judging progress based on appropriate high performing team criteria of any given domain—for example, a human+AI team of pilots and air traffic personnel being able to successfully handle an unprecedented situation (Mani et al. 2021).

Adoption of benchmarks and goals around human+AI teams and their elevation into a category of coveted and prestigious goals to chase by the field can be meaningfully helped by structuring government-funded challenges around them, since, as noted by Mani et al. (2021), grand challenges act as important "compasses" for AI researchers, practitioners and especially young researchers looking for worthwhile problems to work on. Government R&D challenges are known for having been able to kick-start entire fields in AI and prompt significant private sector investment, amounting in some cases to hundreds of billions of US dollars. One of the most famous ones is the DARPA Grand Challenge that first ran in 2004 "with the goal of spurring on American ingenuity to accelerate the development of autonomous vehicle technologies that could be applied to military requirements" (DARPA 2005). In a little over a decade since the first challenge, autonomous driving has attracted tens of billions in private investment and counting (Kerry, Karsten 2017), while DARPA continued to pursue the Grand Challenges model to spur on R&D in areas such as robotics and fully automated network defence (DARPA 2014). A well-designed challenge around labor-enhancing AI could similarly serve as a powerful catalyser for the advancement of worker-empowering AI applications. As applicable, challenges can come with benchmark datasets deliberately designed to evaluate the ability of an ML model to assist a human worker and boost her productivity in non-exploitative ways, instead of trying to outperform humans on their basic abilities.

Sociotechnical imaginaries driving AI development

Aside from grand challenges and chasing state-of-the-art performance targets, influential ideas about the desirable technological future for society also serve as a powerful orienting force for AI innovators. Those ideas are heavily shaped by the technologists themselves, as well as by cultural artifacts, such as hit science fiction stories. Notable among these stories was Star Trek, a cult film franchise and television series that aired in the US from the mid-sixties through the mid-2000s, and Isaac Asimov's Robot stories which Star Trek is partially based on. Asimov was the first science fiction author to present robots favorably; in his stories engineers are enlightened characters. This vision inspired many to pursue robotics. Marvin Minsky, a Turing Award recipient who is widely regarded as a key figure in the history of AI, is said to have decided to become a computer scientist and work on robotics after reading Asimov's science fiction stories (Saadia 2016). To this day, start-ups successfully raise capital to build devices "from Star Trek" (see, for example, Sweeney 2019), and even executives of the largest companies openly declare replicating Star Trek technology as their orienting vision—for example, Amit Singhal, Google's former executive in charge of Search from 2000 till 2016, wanted Google Search to work like his "dream Star Trek computer" (Luckerson 2016).

The fact that the public relations department of a major publicly traded corporation thought it desirable to use a childhood dream of a technologist in charge of their core product as a focal point of this product's public positioning suggests that contemporary society commonly views the intentions to build Star Trek technology—as well as people and organizations declaring such intentions—very favorably. In the 2015 book "Dreamscapes of Modernity" edited by Jasanoff and Kim, desirable visions of the future that do not simply represent individual vanguard visions, but become collective reference points and anchors for future projects are referred to as "sociotechnical imaginaries." Sheila Jasanoff noted that imaginaries are "performed" collectively and that they co-produce the reality of not only the world we live in today but also of the "known, the made, the remembered and the desired worlds." The co-production entails that sociotechnical imaginaries are simultaneously "instruments and products" of a collective understanding of what the world and society should look like (Jasanoff, Kim 2015). Star Trek has for decades been serving as a collective reference point used by Silicon Valley engineers, entrepreneurs and VCs. If it hints at a contour of an imaginary performed by the US tech industry professionals—a community with outsized influence on the direction of AI—it is suggestive not only of what kind of technological tools they dream of building but what kind of social order they aspire to on behalf of human society.

In Star Trek's view of the future, members of the United Federation of Planets, founded on the principles of peace, justice and equality, live in a post-scarcity world, liberated from the necessity to work. They possess a "Replicator"—a device capable of producing almost any desired good instantly at no cost. Star Trek was one of the very first stories in science fiction that painted a picture of a positive, utopian future, one in which technological progress and automation of human labor were beneficial underpinning forces. But, as pointed out by Saadia (2016), the real miracle of Star Trek was not its technology, but the social order that enabled everyone to share in the fruits of the progress.

Notably, the Star Trek series contains very little on what policy choices enable and sustain the social order the Federation enjoys. We obviously do not know if in real life the automation of human labor would ever lead to an egalitarian social order where everyone is entitled to share in the abundance. If history is any indication, the likelihood of economic and political power getting shared through voluntary redistribution from winners to losers, unprompted by long and difficult power struggles with highly uncertain outcomes, is very low. And yet, the technology industry leaders, at least in the US, seem to place a lot of hope on monetary redistribution, while feeling deeply opposed to measures and institutions aiming to redistribute political power more evenly, which is necessary to guarantee the stability and continuity of not only the monetary redistribution but of democratic governance as well.

Brookman et al. (2019) surveyed 600 US technology companies' founders and executives, most of them millionaires, who have raised more than $19.6 billion in venture capital investment. The majority of the study participants (62.1%) chose "Don't Regulate and Do Redistribute" as the best description of their views: they do expect the government to tax and redistribute wealth, they are supportive of universal healthcare and programs benefiting the poor, but they do not feel favorably about government regulation, seeing it as doing more harm than good especially when it comes to the regulation of technology product markets and of the labor market. An overwhelming majority of respondents would like to see the strength of unions decrease and consider it to be at present excessively hard to fire workers.

Whether the "Don't Regulate Do Redistribute" worldview subscribed to by the outsized share of US technology leaders is based on a sincere belief that unregulated technological advancement and monetary redistribution are sufficient for attaining an equitable distribution of power in society, or it is simply a posture deliberately adopted for self-serving reasons, this worldview is likely to be material for any effort of governing AI in service of shared prosperity. There is much more to be understood though about the "mechanics" of how it influences the direction of AI. As Jasanoff and Kim wrote in "Dreamscapes of Modernity" (2015), "...particularly empty of theoretical guidelines is the domain that connects creativity and innovation in science, and even more technology, with the production of power, social order, and communal sense of justice." The development of theoretical guidelines and frameworks that illuminate this crucial connection is important for society's ability to take control of governing AI in a democratic way.

3.2. Structural features of the AI field

What structural features of the AI field prompt and sustain its focus on matching and exceeding human performance on basic abilities? To achieve recognition in the field that celebrates state-of-the-art performance, one needs to demonstrate that performance by running a high number of experiments, which is one of the features papers accepted to top conferences in AI are characterized by. This puts academic researchers at a disadvantage compared to their industry peers, because the former group, with rare exceptions, is much more restricted in its access to computing power. This is a major factor contributing to the present state of the field in which the industry, and not academia, is central to AI research activity (Reich 2021). Between 2010 and 2019, the share of graduating PhDs in AI in the US and Canada going to the industry grew from 44.4% to 65% (AI Index 2021, cited by Reich 2021). This "brain drain" away from the academic AI research centers can have far-reaching implications for what real-world problems get tackled by AI researchers and what problems remain without sufficient attention.

The significance of restrictions placed on AI practitioners working in the for-profit sector is highlighted by Rakova et al. (2021). The study presents the results of 26 semi-structured interviews with the AI industry practitioners who advance responsible AI practices in their organizations either as a part of their formal job description or voluntarily. They report having to distill what they do into standard metrics used by the industry, such as number of clicks, acquired users, and churn rate, as one of the key barriers to achieving progress in their work. They are commonly asked to measure the impact of the responsible AI efforts in terms of revenue generated. Product teams they are frequently embedded into are usually pressured to deliver within fast-paced development cycles that incentivize the use of success criteria that are easier to measure and discourage paying attention to long-term societal outcomes. Industry organizations surveyed by Rakova et al. (2021) all lacked structures of accountability for AI's societal impacts, making the analysis of societal impacts of AI likely to be neglected without consequences.

Also relevant in the discussion of the AI field's "architecture" is Conway's law, which states that any organization designing a system "will produce a design whose structure is a copy of the organization's communication structure" (Conway 1968). Conway's law has been shown to be supported empirically, including in the software industry. Documented natural experiments in the software industry highlight that an organization's governance model, approach to problem-solving, and communication patterns "constrain the space in which it searches for new solutions" (MacCormack, Baldwin and Rusnak 2012).

Since organizations in the industry and academia are structured quite differently, we can expect the centrality of the industry to AI R&D to leave an imprint on the structural design features of the systems created in the course of AI progress. For example, universities typically include sets of faculty with deep expertise in a well-rounded set of fields. The frequency and depth of collaborations between the faculty from different departments vary by university and individual faculty members, but at least in theory, they can always solicit an opinion of a colleague from a very different discipline working within the walls of the same university. This is not so within industry. Not only do companies tend to be much more homogenous in terms of represented disciplines, but soliciting an opinion of an academic from a different discipline often requires signing a non-disclosure agreement, which many academics can be wary about, making a case for the collaboration, securing a budget, etc. These difficulties in communication flows can end up leading to AI systems created without sufficient multi-disciplinary consideration of the impact of the design choices on society and the prosperity of its members.

4. Conclusion

Dozens of organizations have published "Responsible AI" principles in the last few years. Those commonly include declarations of intent to make AI transparent, accountable and beneficial to all (Fjeld 2020). And while many organizations have begun dedicating some staff time and resources to substantiate their promises to make AI fair, explainable and safe, very little is being done to move beyond on-paper principles when it comes to ensuring that AI does not exacerbate inequality and lead to concentration of economic power and productive capacities. On the contrary, the expectation that AI advancement will generate large, left behind groups who will need to be supported by expanded retraining programs and social safety benefits like Universal Basic Income is increasingly broadly shared, adding to the troubling normalization of thinking of the current direction of AI advancement as the only one possible.

Viewing the path of AI progress as unalterable might serve certain economic and political interests or simply be a result of misguided beliefs. However, this chapter underscores that adopting this view would deny the objective importance of key factors shaping the direction of AI advancement, such as economic incentives and the policy environment the AI industry is faced with, as well as the ideas, norms and structural features of the field. There is a growing volume of research on how policy decisions influence the direction of AI by altering economic incentives faced by AI innovators, but much remains to be understood about the impact of norms and architecture of the AI field on its direction.

"The AI community has historically fetishized beating or replacing humans," Frank Chen wrote in Mani et al. (2021). Understanding the drivers of this "fetishized" focus is critical for our ability to govern the direction of AI. The redirection of AI towards human complementarity will not be achieved without developing ways to practically measure the impact of AI on labor demand and job quality. Transitioning to human complementarity also requires closing the major gaps in our understanding of how Lessig's four forces — legislation, market, as well as norms and structural features of the AI field — enable and fuel AI's present jobs-destructing focus.

Bibliography

Acemoglu, D., 2010. When does labor scarcity encourage innovation?. Journal of Political Economy, 118(6), pp.1037-1078.

Acemoglu, D., Manera, A. and Restrepo, P., 2020. Does the US Tax Code Favor Automation? (No. w27052). National Bureau of Economic Research.

Acemoglu, D. and Restrepo, P., 2019. Automation and new tasks: How technology displaces and reinstates labor. Journal of Economic Perspectives, 33(2), pp.3-30.

Acemoglu, D. and Restrepo, P., 2020. The wrong kind of AI? Artificial intelligence and the future of labour demand. Cambridge Journal of Regions, Economy and Society, 13(1), pp.25-35.

Acemoglu, D. and Restrepo, P., 2021. Tasks, Automation, and the Rise in US Wage Inequality (No. w28920). National Bureau of Economic Research.

AI Index Annual Report, 2021. https://aiindex.stanford.edu/report/. Accessed on April 15 2021.

Altman, S., 2021. Moore's Law for Everything. https://moores.samaltman.com/ Accessed on August 4, 2021.

Autor, D., Mindell, D. and Reynolds, E., 2020. The Work of the Future: Building Better Jobs in an Age of Intelligent Machines. MIT Work of the Future.

Broockman, D.E., Ferenstein, G. and Malhotra, N., 2019. Predispositions and the political behavior of American economic elites: Evidence from technology entrepreneurs. American Journal of Political Science, 63(1), pp.212-233.

Clemens, M.A., Montenegro, C.E. and Pritchett, L., 2019. The place premium: Bounding the price equivalent of migration barriers. Review of Economics and Statistics, 101(2), pp.201-213.

Conway, M.E., 1968. How do committees invent. Datamation, 14(4), pp.28-31.

Costello, B. and Karichkoff. A., 2019. Truck driver shortage analysis 2015. Arlington, VA: The American Trucking Associations.

Defense Advanced Research Projects Agency (DARPA), 2005. The Grand Challenge. Available at: https://www.darpa.mil/about-us/timeline/-grand-challenge-for-autonomous-vehicles. Accessed on June 25, 2021.

Defense Advanced Research Projects Agency (DARPA), 2014. The DARPA Grand Challenge: Ten Years Later. Available at: https://www.darpa.mil/news-events/2014-03-13. Accessed on June 25, 2021.

Diao, X., Ellis, M., McMillan, M. and Rodrik, D., 2021. Africa's Manufacturing Puzzle: Evidence from Tanzanian and Ethiopian Firms.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A. and Srikumar, M., 2020. Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. Berkman Klein Center Research Publication, (2020-1).

Gerety, R., 2020. Expert Interview.

Gray, M.L. and Suri, S., 2019. Ghost work: How to stop Silicon Valley from building a new global underclass. Eamon Dolan Books.

Ivanov, S., 2020. ICML 2020. Comprehensive analysis of authors, organizations, and countries. Available at: https://medium.com/criteo-engineering/icml-2020-comprehensive-analysis-of-authors-organizations-and-countries-c4d1bb847fde Accessed on: April 15, 2021.

Jasanoff, S. and Kim, S.H., 2015. Dreamscapes of Modernity: Sociotechnical imaginaries and the fabrication of power. Chicago.

Kerry, C. and Karsten, J., 2017. Gauging investment in self-driving cars. Brookings Report. Available at: https://www.brookings.edu/research/gauging-investment-in-self-driving-cars/ Accessed on: June 25, 2021.

Klinova, K. and Korinek, A., 2021. AI and Shared Prosperity. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES'21), May 19—21, 2021, Virtual Event, USA. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3461702.3462619

Korinek, A., 2019. Integrating Ethical Values and Economic Value to Steer Progress in Artificial Intelligence (No. w26130). National Bureau of Economic Research.

Korinek, A., Schindler, M. and Stiglitz, J.E., 2021. Technological Progress, Artificial Intelligence, and Inclusive Growth. IMF Working Paper.

Korinek, A. and Stiglitz, J.E., 2019. Artificial intelligence and its implications for income distribution and unemployment (No. w24174). National Bureau of Economic Research.

Lanier, J., 2014. Who owns the future? Simon and Schuster.

Lessig, L., 1999. Code: And other laws of cyberspace. Basic Books.

Luckerson, V., 2016. How a Big Promotion at Google Reveals the Future of Search. Available at: https://time.com/4206532/amit-singhal-google-future-of-search/. Accessed on: June 25, 2021.

MacCormack, A., Baldwin, C. and Rusnak, J., 2012. Exploring the duality between product and organizational architectures: A test of the "mirroring" hypothesis. Research Policy, 41(8), pp.1309-1324.

Mani, G., Chen, F., Cross, S., Kalil, T., Gopalakrishnan, V., Rossi, F. and Stanley, K., 2021. Artificial Intelligence's Grand Challenges: Past, Present, and Future. AI Mag., 42(1), pp.61-75.

Nguyen, A., 2021. The constant boss: work under digital surveillance. Data and Society.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S., 2015. Librispeech: An ASR corpus based on public domain audio books, 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 2015, pp. 5206-5210, doi: 10.1109/ICASSP.2015.7178964.

Partnership on AI, 2021. Redesigning AI for Shared Prosperity: an Agenda. Available at: partnershiponai.org/shared-prosperity. Accessed on June 21, 2021.

Pritchett, L., 2020. The future of jobs is facing one, maybe two, of the biggest price distortions ever. Middle East Development Journal, 12(1), pp.131-156.

Raji, I.D., Bender, E., Paullada, A., Denton, E., Hanna, A., 2020. AI and the Everything in the Whole Wide World Benchmark. ML Retrospectives NeurIPS Workshop proceedings.

Rakova, B., Yang, J., Cramer, H. and Chowdhury, R., 2021. Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1), pp.1-23.

Reich, R., 2021. Why AI Needs Academia. Available at: http://bostonreview.net/forum/redesigning-ai/rob-reich-why-ai-needs-academia. Accessed on: June 25 2021.

Reddy, R., 1988. Foundations and grand challenges of artificial intelligence: AAAI presidential address. AI Magazine, 9(4), pp.9-9.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A., Li., F. ImageNet Large Scale Visual Recognition Challenge. IJCV, 2015.

Russell, S., 2019. Human-Compatible: Artificial Intelligence and the Problem of Control. Penguin.

Saadia, M., 2016. Trekonomics: The Economics of Star Trek. Inkshares.

Schindler, M.M., Korinek, M.A. and Stiglitz, J., 2021. Technological Progress, Artificial Intelligence, and Inclusive Growth (No. 2021/166). International Monetary Fund

Seamans, R., 2021. Comments at the "Redesigning AI for Shared Prosperity" seminar held by the Centre for the Governance of AI. Available at: https://www.youtube.com/watch?v=EvcCx_qEkrs Accessed on July 2, 2021.

Siddarth, D., Weyl, G., Dick, S., Crawford, K., forthcoming. *Don't Let's Talk About AI*.

Stiglitz, J., 2014. Unemployment and Innovation. NBER Working Paper no. 20670.

Sweeney, K., 2019. Cambridge 'Star Trek' technology startup hooks millions to attack $206bn market. Available at: https://www.businessweekly.co.uk/news/hi-tech/cambridge-%E2%80%98star-trek%E2%80%99-technology-startup-hooks-millions-attack-206bn-market. Accessed on: June 25, 2021

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O. and Bowman, S.R., 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. arXiv preprint arXiv:1905.00537.

Zellers, R., Bisk, Y., Farhadi, A. and Choi, Y., 2019. From recognition to cognition: Visual commonsense reasoning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6720-6731).