NBER WORKING PAPER SERIES

# HARMS OF AI

Daron Acemoglu

Working Paper 29247
http://www.nber.org/papers/w29247

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2021

Prepared for The Oxford Handbook of AI Governance. I am grateful to many co-authors who have contributed to my thinking on these topics and on whose work I have heavily relied in this essay. They include: David Autor, Jonathon Hazell, Simon Johnson, Jon Kleinberg, Anton Korniek, Azarakhsh Malekian, Ali Makhdoumi, Andrea Manera, Sendhil Mullainathan, Andrew Newman, Asu Ozdaglar, Pascual Restrepo and James Siderius. I am grateful to David Autor, Lauren Fahey, Vincent Rollet, James Siderius and Glen Weyl for comments. I gratefully acknowledge financial support from Google, the Hewlett Foundation, the NSF, the Sloan Foundation, the Smith Richardson Foundation, and the Schmidt Sciences Foundation. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

Harms of AI
Daron Acemoglu
NBER Working Paper No. 29247
September 2021
JEL No. J23,J31,L13,L40,O33,P16


## ABSTRACT

This essay discusses several potential economic, political and social costs of the current path of AI technologies. I argue that if AI continues to be deployed along its current trajectory and remains unregulated, it may produce various social, economic and political harms. These include: damaging competition, consumer privacy and consumer choice; excessively automating work, fueling inequality, inefficiently pushing down wages, and failing to improve worker productivity; and damaging political discourse, democracy's most fundamental lifeblood. Although there is no conclusive evidence suggesting that these costs are imminent or substantial, it may be useful to understand them before they are fully realized and become harder or even impossible to reverse, precisely because of AI's promising and wide-reaching potential. I also suggest that these costs are not inherent to the nature of AI technologies, but are related to how they are being used and developed at the moment - to empower corporations and governments against workers and citizens. As a result, efforts to limit and reverse these costs may need to rely on regulation and policies to redirect AI research. Attempts to contain them just by promoting competition may be insufficient.



Daron Acemoglu
Department of Economics, E52-446
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139
and NBER
daron@mit.edu

# 1 Introduction

To many commentators, artificial intelligence (AI) is the most exciting technology of our age, promising the development of "intelligent machines" that can surpass humans in various tasks, create new products, services and capabilities, and even build machines that can improve themselves, perhaps eventually beyond all human capabilities. The last decade has witnessed rapid progress in AI, based on the application of modern machine learning techniques and huge amounts of computational power to massive, often unstructured data sets (e.g., Russell and Norvig, 2009, Neapolitan and Jiang, 2018, Russell, 2019).[1] AI algorithms are now used by almost all online platforms and in industries that range from manufacturing to health, finance, wholesale and retail (e.g., Ford, 2015; Agarwal, Gans and Goldfarb, 2018; West, 2018). Government agencies have also started relying on AI, especially in the criminal justice system and in customs and immigration control (Thompson, 2019; Simonite, 2020).

Whether AI will be everything its enthusiastic creators and boosters dream, it is likely to have transformative effects on the economy, society and politics in the decades to come, and some of these are already visible in AI algorithms' impact on social media, data markets, monitoring of workers and work automation. Like many technological platforms (or "general purpose technologies") that can be used for the development of a variety of new products, services, and production techniques, there are a lot of choices about how AI technologies will be developed. This, combined with the pervasive effects of AI throughout society, makes it particularly important that we consider its potential dark side as well.

In this essay, I will focus on three broad areas in which the deployment of AI technologies may have economic and social costs — if not properly regulated. I want to emphasize at the outset that the arguments I will present are *theoretical*, and currently there is insufficient empirical evidence to determine whether the mechanisms I isolate are important in practice. The spirit of the exercise is to understand the *potential harms* unregulated AI may create so that we have a better understanding of how we should track and regulate its progress.

The areas I will focus on are:

1. **Collection and control of information**. I will argue that the combination of the

[1] The field of "AI" today is dominated by the suite of current artificial intelligence technologies and approaches, mostly based on statistical pattern recognition, machine learning and big data methods. The potential harms of AI I discuss in this paper are relevant for and motivated by these approaches. Nevertheless, I will also emphasize that "AI" should be thought of as a broad technological platform, precisely because the general aspiration to produce "machine intelligence" includes efforts to improve machines in order to complement humans, create new tasks and services, and generate novel communication and collaboration possibilities.

demand of AI technologies for data and the ability of AI techniques for processing vast amounts of data about users, consumers and citizens produces a number of potentially troubling downsides. These include: (a) *privacy violation:* companies and platforms may collect and deploy excessive amounts of information about individuals, enabling them to capture more of the consumer surplus via price discrimination or violate their privacy in processing and using their data; (b) *unfair competition*: companies with more data may gain a strong advantage relative to their competitors, which both enables them to extract more surplus from consumers and also relaxes price competition in the marketplace, with potentially deleterious effects; and (c) *behavioral manipulation:* data and sophisticated machine learning techniques may enable companies to identify and exploit biases and vulnerabilities that consumers themselves do not recognize, thus pushing consumers to lower levels of utility and simultaneously distorting the composition of products in the market.

2. **Labor market effects of AI**. I will argue that even before AI there was too much investment in cutting labor costs and wages in the US (and arguably in some other advanced economies as well). Such efforts may be excessive either because, in attempting to cut costs, they reduce production efficiency, or because they create non-market effects (for example, on workers losing their jobs or being forced to take lower-pay work). AI, as a broad technological platform, could have in principle rectified this trend, for example, by promoting the creation of new labor-intensive tasks or by providing tools for workers to have greater initiative. This does not seem to have taken place. For example, automation is a quintessential example of efforts to cut labor costs, and like other efforts, it may be excessive. Many current uses of AI involve automation of work or the deployment of AI in order to improve monitoring and keep wages low, motivating my belief that AI may be exacerbating the excessive efforts to reduce labor costs. In this domain, I focus on four broad areas: (a) *automation*: I explain why automation, a powerful way to reduce labor costs, can be part of the natural growth process of an economy, but it can also be excessive, because firms do not take into account the negative impact of automation on workers; (b) *composition of technology*: problems of excessive automation intensify when firms have a choice between investing in automation versus new tasks, and I explain why this margin of technology choice may be severely distorted, and how AI-type technologies may further distort this composition; (c) *loss of economies of scope in human judgment*: in contrast to the hope that AI will take over routine tasks and in

the process enable humans to specialize in problem-solving and creative tasks, AI-human interplay might gradually turn humans into worse decision-makers as they hand more and more decisions to machines, especially when there are economies of scope across tasks; and (d) *monitoring*: I also explain how technologies like AI that increase the monitoring ability of employers are very attractive to firms, but may at the same time generate significant social inefficiencies.

3. **AI, communication and democracy.** I will finally suggest that AI has also exacerbated various political and social problems related to communication, persuasion and democratic politics that, once again, predate the onset of this technology. The main concern here is that democratic politics may have become more difficult, or even fundamentally flawed, under the shadow of AI. I focus on: (a) *echo chambers in social media*: how AI-powered social media generates echo chambers that propagate false information and polarize society; (b) *problems of online communication*: I suggest that online social media, which is interwoven with AI, creates additional misalignments related to private communication; (c) *big brother effects*: AI increases the ability of governments to closely monitor and stamp out dissent. All of these effects of AI are, by their nature, damaging, but may have their most consequential effects by impairing democratic discourse; and (d) *automation and democracy:* finally, I suggest that the process of automation may further damage democracy by making workers less powerful and less indispensable in workplaces.

In each one of the above instances, I discuss the basic ideas about potential costs and, when appropriate, I present some of the modeling details (in some cases based on existing work and in others as ideas for future exploration). Throughout, my approach will be informal, attempting to communicate the main ideas rather than providing the full details of the relevant models.

In addition, I will point out the relevant context and evidence in some cases, even though, as already noted, we do not have sufficient evidence to judge whether most of the mechanisms I am exploring here are likely to be important in practice. I then discuss the common aspects of the potential harms from AI and explore their common roots. I also argue that these costs, if proved important, cannot be avoided in an unregulated market. In fact, I will suggest that in many of these instances, greater competition may exacerbate the problem rather than resolving it.

The aforementioned list leaves out several other concerns experts have expressed (AI leading

to evil super intelligence or AI's effects on war and violence), mostly because of space restrictions and also partly because they are further away from my area of expertise. I mention them briefly in Section 5.

The long list of mechanisms via which AI could have negative economic, political and social effects may create the impression that this technological platform is *bound* to have disastrous social consequences, or it may suggest that some of these problems are solely created by AI. Neither is true. Nor am I particularly opposed to this technology. I believe that AI is a hugely promising technological platform. Furthermore, with or without AI, our society has deep problems related to the power of corporations, automation and labor relations, and polarization and democracy. AI exacerbates these problems, *because* it is a powerful technology and, owing to its general-purpose nature and ambition, it is applicable in a wide array of industries and domains, which amplifies its ability to deepen existing fault lines. These qualities make the potential negative effects of AI quite difficult to foresee as well. Perhaps even more than with other technologies and technological platforms, there are many different directions, with hugely different consequences, in which AI can be developed. This makes it doubly important to consider the costs that it *might* create. It also makes it vital to think about the *direction* of development of this technology.

Indeed, my point throughout is that AI's costs are avoidable, and if they were to transpire, this would be because of the choices made and the direction of research pursued by AI researchers and tech companies. They would also be due to the lack of appropriate regulation by government agencies and societal pressure to discourage nefarious uses of the technology and to redirect research away from them. This last point is important: again like most other technologies, but only more so, the direction of research of AI will have major distributional consequences and far-ranging implications for power, politics, and social status in society, and it would be naïve to expect that unregulated markets would make the right trade-offs about these outcomes — especially since, at the moment the major decisions about the future of AI are being made by a very small group of top executives and engineers in a handful of companies. Put differently, AI's harms are harms of unregulated AI. But in order to understand what needs to be regulated and what the socially optimal choices may be, we first need to systematically study what the downside of this technology may be. It is in this spirit that this current essay is written.

The rest of this essay is organized as follows. In Section 2 I start with the effects that AI creates via the control of information. Section 3 moves to discuss AI's labor market implications.

Section 4 turns to the effects of this technological platform on social communication, polarization and democratic politics. Section 5 briefly touches upon a few other potential unintended consequences of AI technologies. Section 6 steps back and discusses the role of choice in this process. I explain why the direction of technological change in general, and the direction of AI research in particular, is vital, and how we should think about it. This section also reiterates that many of the costs mentioned in the preceding sections are the result of choices made about the development and use of AI technologies in specific directions. It then builds on the mechanisms discussed in previous sections to emphasize that unregulated markets are unlikely to internalize AI's costs and how greater competition may sometimes make things worse, and nor are unfettered markets likely to direct technological change towards higher social value uses of AI. In this spirit, this section also provides some ideas about how to regulate the use of AI and the direction of AI research. Section 7 concludes.

# 2 AI and Control of Information

Data are the lifeblood of AI. The currently-dominant approach in this area is based on turning decision problems into prediction tasks and apply machine learning tools to very large data sets in order to perform these tasks. Hence, most AI researchers and economists working on AI and related technologies start from the premise that data create positive effects on prediction, product design and innovation (e.g., Brynjolfsson and McAfee, 2019; Jones and Tonetti, 2020; Varian, 2009; Farboodi et al., 2019). However, as emphasized by several legal scholars and social scientists, data and information can be misused — deployed in exploitative ways that benefit digital platforms and tech companies at the expense of consumers and workers (e.g., Pasquale, 2015, Zuboff, 2019). Zuboff, for example, argues that such exploitative use of data is at the root of the recent growth of the tech industry, which "claims human experience as free raw material for hidden commercial practice process of extraction, prediction, and sales." (2019, p. 8).

In this section, I discuss social costs of AI related to the control of data and information, with a special emphasis on exploring when data can become a tool for excessive extraction and prediction.

## 2.1 Too Much Data

Concerns about control and misuse of information become particularly important when there are benefits to individuals from "privacy". Individuals may value privacy for instrumental or intrinsic reasons. The former includes their ability to enjoy greater consumer surplus, which might be threatened if companies know more about their valuations and can charge them higher prices. The latter includes various characteristics and behaviors that individuals would prefer not to reveal to others. This could be for reasons that are economic (e.g., to avoid targeted ads), psychological (e.g., to maintain a degree of autonomy), social (e.g., to conceal certain behaviors from acquaintances), or political (e.g., to avoid persecution).

Standard economic analyses tend to view these privacy-related costs as second-order for two related reasons: if individuals are rational and are given decision rights, then they will only allow their data to be used when they are compensated for it adequately, and this would ensure that data will be used by companies only when their benefits exceed the privacy costs (e.g., Varian, 2009; Jones and Tonetti, 2020). Secondly, in surveys individuals appear to be willing to pay only little to protect their privacy, and hence the costs may be much smaller than the benefits of data (e.g., Athey et al., 2017). Yet these arguments have limited bite when the control of data has a "social" dimension — meaning that when an individual shares her data, she is also providing information about others. This social dimension is present, by default, in almost all applications of AI, since the use of data is specifically targeted at learning from like-cases in order to generalize and apply the lessons to other settings.

How does this social dimension of data affect the costs and benefits of data? This is a question tackled in a series of papers, including MacCarthy (2010), Choi, Jeon and Kim (2019), Acemoglu et al. (2021) and Bergemann et al. (2021), and here I base my discussion on Acemoglu et al. (2021). The social dimension of data introduces two interrelated effects. First, there will be *data externalities* — when an individual shares her data, she reveals information about others. To the extent that data is socially valuable and individuals do not internalize this, data externalities could be positive. But if indirect data revelation impacts the privacy of other individuals, these externalities could be negative. The second effect is what Acemoglu et al. call *submodularity*: when an individual shares her data and reveals information about others, this reduces the value of others' information both to themselves and to potential data buyers (such as platforms or AI companies). This is for the simple reason that when more information is shared about an individual, the less important the individual's own data become for predicting his or her decisions.

Acemoglu et al. (2021) model this in the following fashion. Consider a community consisting of $n$ agents/users interacting on a (monopoly) digital platform. Each agent $i$ has a type denoted by $x_i$ which is a realization of a random variable $X_i$, where the vector of random variables $\mathbf{X} = (X_1, \ldots, X_n)$ has a joint normal distribution $\mathcal{N}(0, \Sigma)$, with covariance matrix $\Sigma \in \mathbb{R}^{n \times n}$ (and $\Sigma_{ii} = \sigma_i^2 > 0$ denoting the variance of individual $i$'s type). Each user has some personal data, $S_i$, which are informative about her type. Personal data include both characteristics that are the individual's private information (unless she decides to share) and also data that she generates via her activity online and off-line. Suppose $S_i = X_i + Z_i$ where $Z_i$ is a normally-distributed independent random variable, $Z_i \sim \mathcal{N}(0, 1)$.

Although Acemoglu et al. (2021) discuss various metrics, here I suppose that the relevant notion of information is mean square error (MSE). Then we can define *leaked information* about user $i$ as the reduction in the mean square error of the best estimator of her type:

$$\mathcal{I}_i(\mathbf{a}) = \sigma_i^2 - \min_{\hat{x}_i} \mathbb{E}\left[\left(X_i - \hat{x}_i\left(\mathbf{S}_\mathbf{a}\right)\right)^2\right],$$

where $\mathbf{S}$ is the vector of data the platform acquires, $\hat{x}_i(\mathbf{S})$ is the platform's estimate of the user's type given this information, and $\mathbf{a} = (a_1, \ldots, a_n)$ is the data-sharing action profile of users (with $a_i = 0$ denoting no direct data-sharing and $a_i = 1$ corresponding to data-sharing). Then, the objective of the platform is to maximize

$$\sum_i \left[\eta \mathcal{I}_i(\mathbf{a}) - a_i p_i\right],$$

where $p_i$ denotes payment ("price") to user $i$ from the platform, which is made only when the individual in question shares her data directly (i.e., $a_i = 1$), and $\eta > 0$. The price could take the form of an actual payment for data shared or an indirect payment by the platform, for example, the provision of some free service or customization. This specification embeds the idea that the platform would like to acquire data in order to better forecast the type/behavior of users.

User $i$'s objective is different. She may wish to protect her privacy and she obviously benefits from payments she receives. Thus her objective is to maximize:

$$\gamma \sum_{i' \neq i} \mathcal{I}_{i'}(\mathbf{a}) - v_i \mathcal{I}_i(\mathbf{a}) + a_i p_i.$$

The first term represents any positive direct externalities from the information of other users (for example, because this improves the quality of services that the individual receives and

does not fully pay for) and thus $\gamma \geq 0$. The second term is the loss of privacy (capturing both instrumental and intrinsic values of privacy). Hence $v_i \geq 0$ here denotes the value of privacy to user $i$. Finally, the last term denotes the payments she receives from the platform.

This framework allows data to create positive or negative total benefits. To illustrate this point, suppose that $v_i = v$. In that case, data create aggregate (utilitarian) benefits provided that $\eta + \gamma(n-1) > v$. In contrast, if $\eta + \gamma(n-1) < v$, the corporate control and use of data is socially wasteful (it creates more damage than good). But even in this case, as we will see, there may be data transactions and extensive use of data. In general, because $v_i$ differs across agents, data about certain users may generate greater social benefits than the costs, while the revelation of data about others may be excessively costly.

In terms of market structure, the simplest option is to assume that the platform makes take-it-or-leave-it offers to users in order to acquire their data.

A key result proved in Acemoglu et al. (2021) is that $\mathcal{I}_i(\mathbf{a})$ is monotone and submodular. The first property means that when an individual directly shares her data, this weakly increases the information that the platform has about all individuals, i.e., $\mathcal{I}_i(\mathbf{a}') \geq \mathcal{I}_i(\mathbf{a})$ whenever $\mathbf{a}' \geq \mathbf{a}$. Mathematically, the second implies that, for two action profiles $\mathbf{a}$ and $\mathbf{a}'$ with $\mathbf{a}'_{-i} \geq \mathbf{a}_{-i}$, we have

$$\mathcal{I}_i(a_i = 1, \mathbf{a}_{-i}) - \mathcal{I}_i(a_i = 0, \mathbf{a}_{-i}) \geq \mathcal{I}_i(a_i = 1, \mathbf{a}'_{-i}) - \mathcal{I}_i(a_i = 0, \mathbf{a}'_{-i}).$$

Economically, it means that the information transmitted by an individual directly sharing her data is less when there is more data-sharing by others.

I now illustrate the implications of this setup for data sharing and welfare using two simple examples. Consider first a platform with two users, $i = 1, 2$, and suppose that $\gamma = 0$, $\eta = 1$, and $v_1 < 1$ so that the first user has a small value of privacy, but $v_2 > 1$, implying that because of strong privacy concerns, it is socially beneficial not to have user 2's data be shared with the platform. Finally, suppose that the correlation coefficient between the data of the two users is $\rho > 0$. Since $v_1 < 1$, the platform will always purchase user 1's data. But this also implies that it will indirectly learn about user 2 given the correlation between the two users' data. If $v_2$ is sufficiently large, it is easy to see that it would be socially optimal to close off data transactions and not allow user 1 to sell her data either. This is because she is indirectly revealing information about user 2, whose value of privacy is very large. This illustrates how data externalities lead to inefficiency. In fact, if $v_2$ is sufficiently large, the equilibrium, which always involves user 1 selling her data, can be arbitrarily inefficient.

More interesting are the consequences of submodularity, which can be illustrated using this

example as well. To understand these, let us consider the edge case where the information of the two users is very highly correlated, i.e., $\rho \approx 1$. In this example, the platform will know almost everything relevant about user 2 from user 1's data. The important observation is that this data leakage about user 2 undermines the willingness of user 2 to protect her data. In fact, since user 1 is revealing almost everything about her, she would be willing to sell her own data for a very low price. In this extreme case with $\rho \approx 1$, therefore, both the willingness of the platform to buy user 2's data and benefits user 2 receives from protecting her data are very small, and thus this price becomes approximately 0. But here comes the disturbing part for data prices and the functioning of the data market in this instance: once the second user is selling her data, this also reveals the first user's data almost perfectly, so the first user can only charge a very low price for her data as well. As a result, the platform will be able to acquire both users' data at approximately zero price. This price, obviously, does not reflect users' value of privacy. They may both wish to protect their data and derive significant value from privacy. Nevertheless, the market will induce them to sell their data for close to zero price. Imagine once again that $v_2$ is sufficiently high. Then, despite this high value of privacy to one of the users, there will be a lot of data transactions, data prices will be near zero, and the equilibrium will be significantly (arbitrarily) inefficient. These consequences follow from submodularity.

As a second example, consider the case in which again $\gamma = 0$ and $\eta = 1$ but now there is no heterogeneity between the two users, so that $v_1 = v_2 = v > 1$. This configuration implies that neither user would like to sell their data (because their privacy is more important than the value of data to the platform). Nevertheless, it can be shown that so long as $v$ is less than some threshold $\bar{v}$ (which is itself strictly greater than 1), there exists an equilibrium in which the platform buys the data of both users for relatively cheap. This too is a consequence of submodularity: when each user expects the other one to sell their data, they become less willing to protect their own data and more willing to sell it for relatively cheap. This locks both users into an equilibrium in which their data are less valuable than they would normally assume, and partly as a result, there is again too much data transaction.

One final conclusion is worth noting. In addition to leading to excessive data use and transactions, the externalities also shift the distribution of surplus in favor of the platform. To see this, suppose $v_1 = v_2 = v \leq 1$ and $\rho \approx 1$, so that it is now socially optimal for data to be used by the platform. It is straightforward to verify that in equilibrium, data prices will again be equal to zero, and thus all of the benefits from the use of data will be captured by

the platform.

Are data externalities and the inefficiencies they create empirically relevant? Like many of the channels I discuss in this essay, the answer is that we do not know for sure. If, as industry insiders presume, benefits of data are very large, then they will outweigh the costs from data externalities I have highlighted here. Even in this case, the market equilibrium will not be fully efficient, though the use of data by platforms and corporations may be welfare-increasing overall. However, there are reasons to believe that privacy considerations may be quite important in practice. First, many digital platforms have a monopoly or quasi-monopoly situation (such as Google, Facebook or Amazon), and thus their ability to extract rents from consumers can be significant. Second, some of the intrinsic reasons for consumers to care about privacy — related to dissent and civil society activity — are becoming more important, as I discuss in Section 4.

In summary, the general lessons in this case are clear: when an individual's data are relevant about others' behavior or preferences (which is the default case in almost all applications of data), then there are new economic forces we have to take into account, and these can create costs from the use of data-intensive AI technologies. In particular:

1. The social nature of data — enabling companies to use an individual's data for predicting others' behavior or preferences — creates externalities, which can be positive or negative. When negative externalities are important, there will tend to be too much use of data by corporations and platforms.

2. The social nature of data additionally generates a new type of submodularity, making each individual less willing to protect their data when others are sharing theirs. This submodularity adds to the negative externalities, but even more importantly, it implies that data prices will be depressed and will not reflect users' value of data and/or privacy.

3. In addition to leading to excessive use of data, both of these economic forces have first-order distributional consequences: they shift surplus from users to platforms and companies.

If these costs of data use and AI are important, they also call for regulating data markets. Some regulatory solutions are discussed in Acemoglu et al. (2021), and I return to a more general discussion of regulation of AI technologies and data in Section 6.

## 2.2 Data and Unfair Competition

AI technologies amplify the ability of digital platforms and companies using data from these platforms to predict consumer preferences and behavior. On the upside, this might enable firms to design better products for customers (after all, this is one of the main benefits of AI). But the use of such data can also change the nature of competition. These effects become even more pronounced when some firms are much better placed to collect and use data relative to their competitors, and this is the case I will focus on this subsection. Specifically, one firm's collection and use data that others cannot access may create a type of "unfair competition", enabling this firm to capture consumer surplus and relax price competition. I now develop this point in the simplest possible setting, using a Hotelling-type static model with two firms. The main lesson will be that even when data improves product quality, it creates powerful forces that shift the distribution of surplus away from consumers and towards firms.

Suppose that consumers are located uniformly across a line of length 1 and incur a cost — similar to a transport cost — when they purchase a product further away from their bliss point, represented by their location (see, for example, Tirole, 1989). I assume that the utility of consumer $i$ with location (or bliss point) $i$ can be written as

$$\alpha - \beta(x_i^f - i)^2 - p_i^f,$$

where $x_i^f \in [0,1]$ is the product of firm $f \in \{0,1\}$ and $p_i^f$ is its (potentially) customized price for this consumer. Throughout, we normalize the cost of production to zero for both firms (regardless of whether they produce a standardized or customized product).

Let us interpret the two firms as two different websites, which consumers visit in order to purchase the good in question. Before AI, firms cannot observe the type of consumer and I assume that they cannot offer several products to a consumer that visits their websites. Thus they will have to offer standardized products. This description implies that, in terms of timing, they first choose their product, and then after observing each other's product choice, they set prices. Since each firm is offering a standard product and cannot observe consumer type, it will also set the same price for all consumers. This makes the pre-AI game identical to a two-stage Hotelling model, in which firms first choose their product type (equivalent to their location) and then compete in prices. Throughout, I assume that

$$5\beta < 4\alpha, \tag{1}$$

which is sufficient to ensure that the market is covered and the firms will not act as local monopolies. As usual, I focus on subgame perfect equilibria, but with a slight abuse of terminology,

I refer to these as "equilibria".

It is straightforward to see that the unique equilibrium in this model, as in the baseline Hotelling model with quadratic transport costs, is maximal product differentiation (Tirole, 1989). In this setting, this means that the two firms will offer products at the two ends of the line ($x^0 = 0$ and $x^1 = 1$) and set equilibrium prices given by $p^0 = p^1 = \beta$, sharing the market equally. For future reference, I also note that in this equilibrium total firm profits are equal to $\Pi^{\text{pre-AI}} = \pi^0 + \pi^1 = \beta$, and consumer surplus is

$$\begin{aligned} CS^{\text{pre-AI}} &= \alpha - 2\beta \int_0^{1/2} x^2 dx - \beta \\ &= \alpha - \frac{13}{12}\beta, \end{aligned}$$

where the first line of this expression uses the symmetry between the firms and consumers on the two sides of $1/2$.

After advances in AI, one of the firms, say firm 1, can use data from its previous customers (those with $i \geq 1/2$) to predict their type and customize their products and prices.[2] In particular, I assume that in this post-AI environment, firm 1 can observe the type of any consumer $i \geq 1/2$ that visits its website and offers a customized bundle $(x_i^1, p_i^1)$ to this consumer. For simplicity, let us assume that firm 0 cannot do so and also that firm 1 cannot simultaneously offer customized and standardized products. Now in equilibrium, firm 1 will offer each consumer with $i \geq 1/2$ a customized product $x_i^1 = i$. It will also charge higher prices. The exact form of the equilibrium depends on firm 0's product choice, which, given its inability to use the new AI technology, cannot be customized. It is straightforward to see that firm 0 will also change its product, because it no longer needs as much product differentiation (since firm 1 will be charging higher prices). The unique post-AI equilibrium is one in which firm 0 changes its standardized product to $x^0 = 1/4$. It then sets a price that makes the consumers that are farthest away from it indifferent between buying its product and not doing so, i.e.,[3]

$$p^0 = \alpha - \frac{\beta}{16}.$$

It is also straightforward to see that it is optimal for firm 1 to set:

$$p_i^1 = \alpha \text{ for all } i \geq \frac{1}{2},$$

[2] More generally, the fact that AI-intensive firms are using data from and customizing products to their existing customers introduces intertemporal linkages, which could create lock-in effects and rich-get-richer dynamics, as in the switching cost and dynamic oligopoly literatures, such as in Klemperer (1995) and Budd, Harris and Vickers (1993).

[3] Firm 0 could offer a lower price and steal some customers from firm 1, but it can be verified that this would lead to lower profits.

thus capturing all the consumer surplus from the consumers about whom it has data.[4] In this equilibrium, we have $\Pi^{\text{post-AI}} = \alpha - \frac{\beta}{32} > \Pi^{\text{pre-AI}}$ (which is guaranteed by (1)), while consumer surplus is now

$$\begin{aligned} CS^{\text{post-AI}} &= \frac{\alpha}{2} - 2\beta \int_0^{1/4} x^2 dx - \frac{1}{2}\left(\alpha - \frac{\beta}{16}\right) \\ &= \frac{1}{48}\beta. \end{aligned}$$

As a consequence, consumer surplus is much lower in this case. This can be seen most clearly by considering the limit where $\beta \to 0$, in which case the pre-AI consumer surplus is maximal (approaching $\alpha$), while post-AI it becomes minimal (approaching 0). The negative impact of AI technologies on consumer surplus has two interrelated causes. First, firm 1 now uses its better prediction power to capture all the surplus from the consumers, even though it is in principle offering a better product and could have increased consumer welfare. Second, given firm 1's more aggressive pricing, firm 0 is also able to capture more profits, reducing even the surplus of consumers whose data are not being used.

It is worth noting that, in the present model there is no intensive margin of consumer choice and the market is covered (under (1)). As a result, AI does not affect quantity purchased, and even when it reduces consumer welfare, it increases utilitarian welfare — in particular, greater customization reduces "transport costs". The logic of the model highlights that this need not be the case when there is a quantity/intensive margin, because higher markups may inefficiently reduce quantity purchased. We will see in the next subsection that there are other reasons for inefficiency in similar environments.

In summary, the general lessons from this model are complementary to the ones from the previous subsection:

1. The use of AI technologies and detailed consumer data for prediction may improve the ability of firms to customize products for consumers, potentially improving overall surplus.

2. However, it also increases the power of (some) companies over consumers.

3. This has direct distributional implications, enabling AI-intensive firms to capture more of the consumer surplus.

[4] If we had allowed this firm to also market a standardized product, it would additionally compete for consumers $i < 1/2$, about whom it has no data. Our assumption rules out this possibility.

4. The indirect effect of the better collection and processing of data by one firm is to relax price competition in the market, increasing prices and amplifying the direct distributional effects.

Although in this model the overall surplus in the economy increases after the introduction of AI technologies, in the previous subsection we saw that this is not necessarily true in the presence of other data-related externalities, and in the next subsection we will encounter a new economic force distorting the composition of products offered by platforms.

## 2.3 Behavioral Manipulation

The previous subsection discussed how even the beneficial use of improved prediction about consumer preferences and behavior might have a downside. But improved prediction tools can also be put to nefarious uses, with potentially far-ranging negative effects. Platforms that collect and effectively process huge amounts of data might able to predict consumer behavior and biases beyond what the consumers themselves can know or understand. Anecdotal examples of this concern abound. They include the chain store Target successfully forecasting whether women are pregnant and sending them hidden ads for baby products, or various companies estimating "prime vulnerability moments" and send ads for products that tend to be purchased impulsively during such moments. They also include marketing strategies targeted at "vulnerable populations" such as the elderly or children. Less extreme advertising strategies also have elements of the same type of manipulation, for example, when websites favor products such as credit cards or subscription programs with delayed costs and short-term benefits or when YouTube and Facebook use their algorithms to estimate and favor more addictive videos or news feeds for the user group in question. As legal scholars Hanson and Kysar have noted, "Once one accepts that individuals systematically behave in non-rational ways, it follows from an economic perspective that others will exploit those tendencies for gain." (1999, p. 630).

Though these concerns are as old as advertising itself, economists and policy-makers hope that consumers will learn how to shield themselves against abusive practices. The sudden explosion in the capabilities of digital platforms to use AI technologies and massive data sets to improve their predictions undercuts this argument, however. Learning dynamics that had made consumers well adapted to existing practices would be quickly outdated in the age of AI and big data. This issue is explored in Acemoglu et al. (2022), using a continuous-time learning model. Here I outline a similar idea in a much simpler setting.

I consider a dynamic setting with two periods, $t = 0, 1$, and no discounting. Consumers have a choice between two products, $x^1$ and $x^2$, in both periods. They are initially uncertain about which one will yield higher utility. Suppose in particular that the true utility that a consumer gets from the product is either $H$ or $L = 0$. The prior belief of individual $i$ is that these two products will yield high utility for them is, respectively, $q_i^1$ and $q_i^2$.

Both products are produced and offered by a digital platform, which again has zero cost of production and can offer personalized prices. To start with, the platform and the consumer have symmetric information, and thus the platform knows and shares the consumer's prior beliefs. Once an individual consumes one of the two products, she obtains an additional piece of information about her utility from the product. I assume, in particular, that if the true quality is $H$, the consumer receives a positive signal, denoted by $\sigma^H$ (with probability 1). However, if the true quality is $L$, the product might still have deceptively high instantaneous utility (but long-term costs). Thus with probability $\lambda$, the consumer will receive the high signal (and receives the low signal $\sigma^L$ with complementary probability). The most relevant interpretation of this "false-positive" signal is that there are certain types of products that (predictably) appear more attractive to consumers, for example because of their tempting short-term benefits or because of their hidden negative attributes.

Let us assume that the platform perfectly observes the consumer's experience with the product she has consumed, and can change its pricing and product offering in the next period. The game ends at the end of the second period.

The pre-AI equilibrium takes a simple form. The platform will offer whichever product has higher $q_i$ for consumer $i$, say product $j$, and will set the price

$$p_{i,0}^j = q_i^j H,$$

capturing the full surplus. If the signal after consumption is $\sigma^L$, then in the next period, it will offer the other product, $\tilde{}j$, charging the lower price

$$p_{i,1}^{\tilde{}j} = q_i^{\tilde{}j} H,$$

once again capturing the full surplus. If, on the other hand, the signal is $\sigma^H$, then in the second period, the same product will be offered, but now there will be a higher price. I assume that in the pre-AI environment, consumers have sufficient experience with such products and the signals they generate that they can correctly anticipate the likelihood of a high-quality product given a positive signal. As a result, the price following a positive signal will not increase all the

way to $H$. Rather, it will be given by the expected value of the product's quality conditional on a positive signal. A simple use of Bayesian updating gives this price as

$$\begin{aligned} p_{i,0}^j &= \frac{q_i^j}{q_i^j + (1-q_i^j)\lambda} H \\ &= \Xi_i^j H, \end{aligned}$$

which again captures the full surplus from the consumer and also defines the expression $\Xi_i^j$, which is convenient for the remainder of this section. (There is no option value term in prices, because the full surplus is being captured by the platform).

The deployment of AI technologies once again improves the platform's ability to predict consumer preferences and behavior — because it has access to the data from many similar consumers and their experiences with similar products. As pointed out above, I assume that this goes beyond what the consumer herself knows. In particular, I suppose that the platform can now forecast whether the consumer will receive the high signal from a truly low-quality product. This, more generally, captures the ability of the platform to predict whether the individual will engage in an impulse purchase or make other choices with apparent short-term benefits and long-term costs.

Post-AI, therefore, the relevant state for consumer $i$ at time $t = 0$ becomes $\left(\left\{q_i^j, \xi_i^j\right\}_{j=1,2}\right)$, where $\xi_i^j = 1$ designates the event that product $j$ will generate a false-positive signal — which means that in reality it is low-quality for the consumer, but still the signal $\sigma^H$ will be realized if the consumer purchases it. Critically, the platform observes $\xi_i^j$, but the consumer does not. Following Acemoglu et al. (2022), I assume that consumers are "semi-behavioral" and do not fully take into account that in the post-AI world, the platform actually knows $\xi_i^j$. This captures the more general economic force mentioned above: in the pre-AI, business-as-usual world, consumers may have learned from their repeated experiences and purchases, accurately estimating the relevant probabilities. The post-AI world is new and it is less plausible to expect that the consumers will immediately understand the superior information that the platform has acquired. Note also that although it can forecast $\xi_i^j$, the platform cannot observe consumer preferences perfectly, and when $\xi_i^j = 0$, it does not know whether the product is high or low-quality.

What does equilibrium look like in the post-AI world? The key observation is that, while before AI the platform's prediction was aligned with the prior of the household, this is no longer the case in the post-AI world. In particular, suppose that we have $q_i^1 > q_i^2$, but $\xi_i^1 = 0$, while $\xi_i^2 = 1$. Then the platform may prefer to offer the second product. To understand this choice,

let us compute the profits from consumer $i$ when the platform is using these two strategies. When it offers product 1, its total profits are

$$\pi_i^1 = \left[q_i^1 + q_i^1 \Xi_i^1 + (1 - q_i^1) q_i^2\right] H.$$

This expression follows by noting that the platform is at first offering product 1 and charging $q_i^1$. Because $\xi_i^1 = 0$, the consumer will receive a positive signal only if the product is truly high quality, which happens with probability $q_i^1$. However, as indicated by the above discussion, in this case, the consumer does not know whether this was a false-positive or a truly high-quality product, and thus her valuation will be $\Xi_i^1 H$, which explains the second term. Finally, if she receives a negative signal (probability $1 - q_i^1$), in the second period, the platform will offer product 2, charging $q_i^2$.

On the other hand, when it initially offers product 2, the platform's profits are

$$\pi_i^2 = \left(q_i^2 + \Xi_i^2\right) H,$$

because in this case there will be a positive signal for sure.

It is straightforward to see that offering the second good is more profitable for the platform when

$$\frac{q_i^2}{q_i^2 + (1 - q_i^2)\lambda} > q_i^1 (1 - q_i^2) + \frac{(q_i^1)^2}{q_i^1 + (1 - q_i^1)\lambda}. \tag{2}$$

Condition (2) is always satisfied whenever $q_i^2$ is sufficiently close to $q_i^1$. Intuitively, the platform is willing to sacrifice a little bit of revenue in the first period for the certainty of getting the consumer to experience a good that it knows she will like — even though this is not a truly high-quality good.

What about consumer welfare? Perhaps paradoxically, in the first case, the consumer actually has a positive welfare. This is because in this case we have a high-quality product (and the platform indirectly recognizes this following the realization of signal $\sigma^H$, because it knows that $\xi_i^1 = 0$), and hence the positive signal can come only from a truly high-quality good. It is then straightforward to compute the user's welfare as

$$\begin{aligned} U_i^1 &= q_i^1 \left(1 - \frac{q_i^1}{q_i^1 + (1 - q_i^1)\lambda}\right) H \\ &= \frac{q_i^1 (1 - q_i^1)\lambda}{q_i^1 + (1 - q_i^1)\lambda} H > 0. \end{aligned}$$

This positive surplus may appear as the good side of our behavioral assumption. But the platform's second strategy shows the dark side. With this strategy, the consumer will overpay

in the second period (because, given $\xi_i^2 = 1$, the product is in reality low-quality). Hence her utility is

$$U_i^2 = -\frac{q_i^2}{q_i^2 + (1 - q_i^2)\lambda} H < 0.$$

Therefore, the ability of the platform to predict the consumer's preferences and vulnerabilities leads to a situation in which the platform can increase its profits by marketing low-quality products that are likely to appeal to the consumer in the short run.

In contrast to the pattern in the previous subsection, this not only increases platform profits at the expense of consumers, but it also distorts consumption as it lures consumers towards lower-quality products, reducing utilitarian welfare.

The general lessons in this case are complementary but different from the ones I highlighted in the previous two subsections:

1. AI technologies can enable platforms to know more about consumers' preferences than they themselves do.

2. This opens the way for potential behavioral manipulation, whereby the platform can offer products that may temporarily appear as higher-quality than they truly are.

3. This type of behavioral manipulation tends to do more than just shift surplus from consumers to the platform; it also distorts the composition of consumption, creating new inefficiencies.

# 3 Labor Market Effects of AI

US labor markets have not been doing well for workers over the last 40 years. Wage growth since the late 1970s has been much slower than during the previous three decades, while the share of capital in national income has grown significantly (Acemoglu and Autor, 2011; Autor, 2019). Additionally, wage growth, such as it is, has been anything but shared. While wages for workers at the very top of the income distribution — those in the highest tenth percentile of earnings or those with postgraduate degrees — have continued to grow, workers with a high school diploma or less have seen their real earnings fall. Even college graduates have gone through lengthy periods of little real wage growth.

Many factors have contributed to this sluggish average wage growth and real wage declines at the bottom of the distribution. The erosion of the real value of the minimum wage, which

has fallen by more than 30 percent since 1968, has been clearly important for low-wage workers (Lee, 1999). The decline in the power of trade unions and much of the private sector may have played a role as well. The enormous increase in trade with China also likely contributed, by forcing the closure of many businesses and large job losses in low-tech manufacturing industries such as textiles, apparel, furniture, and toys (Autor, Dorn and Hanson, 2013).

My own work with Pascual Restrepo (2019, 2021) emphasizes and documents the importance of the direction of technological progress in this process. While in the four decades after World War II automation and new tasks contributing to labor demand went hand-in-hand, a very different path of technological development emerged starting in the 1980s, exhibiting more automation and much slower advances in human-friendly technologies, such as those involving new tasks (Acemoglu and Restrepo, 2019). Automation eliminated routine tasks in clerical occupations and on factory floors, depressing the demand and wages of workers specializing in blue-collar jobs and clerical functions. Meanwhile professionals in managerial, engineering, finance, consulting, and design occupations flourished — both because they were essential to the success of new technologies and because they benefited from the automation of tasks that complemented their own work. As automation gathered pace, wage gaps between the top and the bottom of the income distribution magnified. In Acemoglu and Restrepo (2021), we estimate that automation has been possibly the most important factor in reshaping the US wage structure, explaining somewhere between 50 to 70% of the variance of changes in wages by demographic group between 1980 and 2016.

All of this predates AI. In Acemoglu et al. (2021), we find that AI activity across US establishments picks up speed only after 2016. Nevertheless, this background is useful because AI may be the next phase of automation, and there is evidence that it is already being used both for automation and for tighter monitoring of workers, further depressing wages and the labor share. In this section, I first explain how automation works and why we may be concerned about excessive automation in general, and how AI may exacerbate these concerns. I then discuss how AI could be used for generating new tasks and technologies that complement humans, but whether this will be the case or not depends on technology adoption and research and development choices of companies. In this context, I suggest reasons for being concerned that the composition of AI research may be heavily distorted. I also discuss why the most benign view of AI's role in the labor market — automating routine jobs, so that workers have time for more creative, problem-solving tasks — may need to be qualified. Finally, I explore how AI may have pernicious effects when it is used for monitoring.

## 3.1 Excessive Automation and AI

In order to situate the role of AI in the broader context of automation technologies, I start with a review of the framework from Acemoglu and Restrepo (2018, 2019), which models the automation of tasks (as well as the creation of new tasks). Suppose there is a single good in the economy, $Y$, whose production requires the combination of a measure 1 of tasks:

$$Y = \left( \int_{N-1}^{N} Y(z)^{\frac{\sigma-1}{\sigma}} dz \right)^{\frac{\sigma}{\sigma-1}}, \tag{3}$$

where $Y(z)$ denotes the output of task $z$ and $\sigma \geq 0$ is the elasticity of substitution between tasks. The key economic decision is the allocation of tasks to factors. Let me focus on just two factors, capital and labor, and suppose as in Acemoglu and Restrepo (2018, 2019) that each factor has task-specific productivities, determining its comparative advantage, and only tasks $z \leq I$ can be automated given the current level of automation technology. This implies:

$$Y(z) = \begin{cases} A^L \gamma^L(z) l(z) + A^K \gamma^K(z) k(z) & \text{if } z \in [N-1, I] \\ A^L \gamma^L(z) l(z) & \text{if } z \in (I, N]. \end{cases}$$

Here $l(z)$ and $k(z)$ denote the total labor and capital allocated to producing task $z$. The state of technology is captured by the following: factor-augmenting terms, $A^L$ and $A^K$, which increase the productivity of the relevant factor uniformly in all tasks; task-specific productivities, $\gamma^L(z)$ and $\gamma^K(z)$, which increase the productivity of a factor in a specific task; the threshold for tasks that are feasible to automate, $I$; and the measure of new tasks, $N$. Let us assume that $\gamma^L(z)/\gamma^K(z)$ is increasing in $z$, so that labor has a *comparative advantage* in higher-indexed tasks. Suppose that capital is produced from the final good, with marginal cost $R$, which also gives its rental rate. Labor is inelastically supplied, with total supply given by $L$, and the equilibrium wage is denoted by $w$.

Acemoglu and Restrepo (2018, 2019) characterize the competitive equilibrium in this economy. Here I allow both competitive and rigid labor markets, by assuming that the wage cannot fall below some level $\underline{w}$. In this case, the equilibrium wage can be written as:

$$w = \max \left\{ \underline{w}, \text{MPL}(L) \right\},$$

where $\text{MPL}(L)$ is the marginal product of labor when there is full employment at $L$. The wage floor may be a consequence of regulations, such as minimum wages and union-imposed minima, or may result from other labor market imperfections, such as efficiency wage considerations.

Let us first focus on how the marginal product of labor changes (without any wage floor). Following Acemoglu and Restrepo (2018), this is given by

$$\begin{aligned}\frac{\partial \ln \text{MPL}(L)}{\partial I} =& \frac{\partial \ln Y(L,K)}{\partial I} && \text{(Productivity effect)} \\ &+ \frac{1}{\sigma}\frac{1-s^L}{1-\Gamma(N,I)}\frac{\partial \ln \Gamma(N,I)}{\partial I} && \text{(Displacement effect)}\end{aligned} \tag{4}$$

where $s^L$ denotes the labor share and $\Gamma(N,I) = \frac{\int_I^N \gamma^L(z)^{\sigma-1}dz}{\int_{N-1}^I \gamma^K(z)^{\sigma-1}dz + \int_I^N \gamma^L(z)^{\sigma-1}dz}$ is a measure of the *labor's task content of production* (capturing what fraction of tasks are assigned to labor). In the special cases where $\sigma = 1$ or where $\gamma^K(z) = \gamma^L(z)$, we have $\Gamma(N,I) = N - I$, but more generally, $\Gamma(N,I)$ is always increasing in $N$ and decreasing in $I$. The first line of (4) represents the *productivity effect,* which is driven by the fact that automation reduces costs and thus increases productivity — by an amount equivalent to the cost difference between producing the marginal tasks by labor vs. capital:

$$\frac{\partial \ln Y(L,K)}{\partial I} = \frac{1}{\sigma - 1}\left[\left(\frac{R}{A^K\gamma^K(I)}\right)^{1-\sigma} - \left(\frac{\text{MPL}(L)}{A^L\gamma^L(I)}\right)^{1-\sigma}\right].$$

The second line is the *displacement effect* created by automation: as tasks are allocated away from labor towards capital, the marginal product of labor declines. This displacement effect, which reduces the range of tasks employing workers, is always negative.

When we are at full employment, (4) gives the impact of automation on wages. When, instead, the wage floor at $\underline{w}$ is binding, then the same effects now impact employment. The only differences are that on the left-hand side of (4), we now have the proportional change in employment, and on the right-hand side, $\text{MPL}(L)/A^L\gamma^L(I)$ is replaced by $\underline{w}/A^L\gamma^L(I)$.

Let us first consider full employment. What happens to the labor market equilibrium following additional automation? Equation (4) first shows that the labor share will always decline — because of the displacement effect, the wage will increase less than proportionately with productivity. Equally importantly, the wage level may fall as well. This is because the displacement effect can be larger than the productivity effect. In particular, when the productivity effect is small, for example, in the edge case where $\underline{w}/A^L\gamma^L(I) \approx R/A^K\gamma^K(I)$, there is no productivity effect, and the equilibrium wage will necessarily decline. When there is more than one type of labor, the same argument also implies that the average wage may fall, though the wage of some groups may increase (see Acemoglu and Restrepo, 2021).

This framework further clarifies why automation could reduce employment. Suppose the wage floor $\underline{w}$ is binding and again take the edge case where $\underline{w}/A^L\gamma^L(I) \approx R/A^K\gamma^K(I)$, so that

the productivity effect is approximately zero. Then, automation necessarily reduces employment. By continuity, the same happens when the productivity effect is positive but not too large — the case that Acemoglu and Restrepo (2019) refer to as "so-so technologies", because they are good enough to be adopted, but not so good as to have a meaningful impact on productivity.

What are the welfare consequences of employment-reducing automation? In a perfectly competitive market, where workers are at the margin indifferent between leisure and work, and when there are no other distributional concerns, an automation-induced decline in employment does not have first-order welfare consequences. In fact, it is straightforward to see that the competitive equilibrium would always maximize net output (defined as total production minus what is used up for producing capital). However, when there are labor market imperfections, such as those captured by the wage floor $\underline{w}$, then low-productivity automation reduces welfare — thus motivating the term "excessive automation". This can be seen with the following argument: because the productivity effect is approximately zero, gross output and profits do not increase (workers in marginal tasks are replaced by machines, but total costs have not changed). Yet, capital usage increases, and this reduces net output. At this point, reallocating marginal tasks away from capital towards labor — thus reducing automation — would increase net output.

Why is the equilibrium misaligned with social welfare maximization? The answer is related to the wage floor. Firms, when making their hiring and automation decisions, are responding to the market wage, $\underline{w}$, whereas a utilitarian social planner — seeking to maximize net surplus — should take into account the opportunity cost of labor, which is zero. This argument establishes that when productivity effects are limited, there will be excessive automation. It also pinpoints one of the channels for this type of inefficiency: in economies with labor market imperfections, firms base their automation decisions on the higher wage rate, rather than the lower social opportunity cost of labor.

This argument also clarifies that automation is likely to be excessive and potentially welfare-reducing *especially when* it generates small or negligible productivity effects. If the productivity gains from automation had been large, net output would have increased, even if it displaced workers. Moreover, equation (4) highlights that with a large productivity effect, there may not have been a decline in labor demand in the first place.

The case for excessive automation is strengthened if there are other considerations favoring higher levels of employment. For example, if employed individuals generate positive external

effects (on their families and communities or for democracy) relative to the unemployed, then the social planner may want to increase employment beyond the equilibrium level. Distributional concerns would also weigh in the same direction, since, in general, automation helps firms and firm owners, while reducing the labor share. In addition, as shown in Autor, Levy and Murnane (2003) and Acemoglu and Restrepo (2021), automation boosts inequality across worker groups, creating another distributional cost.

What does this imply for AI? AI is a broad technological platform, and can be used for developing many different types of technologies. Automation, especially automation of various white-collar tasks and jobs with a significant decision-making component, is one of these applications. If AI is used for automation, then the arguments outlined above would also imply that low-productivity AI may reduce welfare. Two key questions are thus whether AI technologies are likely to be deployed for substituting capital and algorithms for labor in various tasks and whether this will generate small or large productivity gains. The evidence in Acemoglu et al. (2021) suggests that there has been a significant uptick in AI activities since 2016, and much of this has been associated with task displacement. That paper also finds reduced hiring in establishments adopting AI technologies, so the evidence is consistent with, though does not prove, the idea that new AI technologies may not be improving productivity sufficiently. There are other reasons why productivity gains from AI may be small. Most importantly, AI technologies are being used in some tasks in which humans are quite good (natural language processing, facial recognition, problem-solving; see Acemoglu, 2021).

In summary, the general lessons from this section are:

1. Automation reduces the labor share and may also reduce the (average) wage and/or employment, and this latter outcome is more likely when productivity gains from automation are small.

2. When labor market imperfections create a wedge between the market wage and the social opportunity cost of labor, automation tends to be excessive and welfare-reducing, particularly when it impacts employment negatively as well. This too is more likely to be the case when its productivity effects are small. The same considerations apply when there are non-market reasons for preferring high levels of employment (for example, because employed workers contribute more to their families, communities or society in general).

3. Because it increases the capital share and reduces the labor share and because it boosts

inequality among workers, automation may also be excessive from a welfare point of view due to distributional concerns

4. If AI is used predominantly for automation, it will have similar effects to other automation technologies, and depending on its productivity effects and relevant welfare criteria, it may have a negative impact on social welfare.

## 3.2 Direction of AI Technology and its Labor Market Consequences

The previous subsection discussed some implications of AI used for automation, but it also noted that AI, as a broad technological platform, can be used for creating new tasks or increasing labor productivity as well. In the framework of the previous subsection, this would correspond to an increase in $N$. The framework presented in the previous section additionally implies that new tasks increase the labor share and raise wages or employment (or both). In particular, as in Acemoglu and Restrepo (2018, 2019), we now have

$$\begin{aligned}\frac{\partial \ln \mathrm{MPL}(L)}{\partial N} =& \frac{\partial \ln Y(L,K)}{\partial N} && \text{(Productivity effect)} \\ &+ \frac{1}{\sigma}\frac{1-s^L}{1-\Gamma(N,I)}\frac{\partial \ln \Gamma(N,I)}{\partial N}. && \text{(Reinstatement effect)}\end{aligned} \tag{5}$$

The productivity effect is positive as usual (even if the exact sources of productivity gains from new tasks are different than those from automation). In addition, the reinstatement effect is also positive, because it is driven by the fact that new labor-intensive tasks are reinstating labor back into the production process. As a result, new tasks always increase employment and/or wages. Moreover, the presence of the reinstatement effect implies that the wage bill increases proportionately more than the productivity gains, pushing up the labor share — the converse of the impact of automation. Acemoglu and Restrepo (2019) have argued that the reason why wages grew robustly during the decades following World War II was that rapid automation in certain tasks went hand-in-hand with the introduction of sufficiently many new tasks, counterbalancing the labor market implications of automation.

Returning to the implications of AI, with the same argument, using AI for new tasks would be welfare-improving, especially when there are labor market imperfections or other considerations favoring higher levels of employment than in equilibrium. Furthermore, if AI boosts the creation of new tasks and improves human productivity, it could counterbalance some of the adverse effects of automation based on other technologies (such as robotics or specialized software).

When AI can be used both for automation and for new tasks, the pivotal question becomes how the balance between these two activities is determined — that is, the direction of technological change. Acemoglu and Restrepo (2018) provide a framework for the analysis of the equilibrium direction of technology. Their framework emphasizes the role of factor prices and the labor share and highlights that there are reasons for optimal and equilibrium allocations to differ. In particular, labor market imperfections not only promote too much automation — as we saw in the previous subsection — but could further lead to an unbalanced composition of AI research between automation and new tasks.

There may also be potential distortions in the direction of technological change that go beyond the purely economic. In Acemoglu (2021) I emphasize that the direction of technology is partly shaped by the business models of leading firms and the aspirations of researchers, and if these favor automation, the equilibrium may involve too much automation, even absent economic distortions. Another related argument is that US corporations may have become too focused on cost-cutting, which might also encourage excessive automation. Acemoglu, Manera and Restrepo (2020), on the other hand, show that the US tax code imposes a much higher marginal tax rate on labor than equipment and software capital, thus favoring automation. This policy channel triggers both excessive adoption of automation technologies and disproportionate emphasis on automation in research and development.

AI as a technological platform could in principle boost efforts to create new tasks. Take education as an example. Current investments in this area are focused on using AI technologies for automated grading and the development of online learning tools to replace various tasks performed by teachers. Yet, AI can be deployed for creating new tasks and directly increasing teacher productivity as well. It can be used for adapting teaching material to the needs and attitudes of diverse students in real time, overcoming a major problem of classroom-based teaching — the fact that students have diverse strengths and weaknesses and find different parts of the curricula more challenging (see the discussion in Acemoglu, 2021). Likewise, AI has many diverse applications in health that can personalize care and empower nurses and general practitioners to make more and better decisions in care delivery. These potentially promising directions notwithstanding, AI may be more likely to aggravate excessive automation. The current trajectory in AI research is shaped by the visions of large tech companies, who are responsible for the majority of the spending on this technology. Many of these companies have business models centered on substituting algorithms for humans, which may make them focus excessively on using AI for automation. At the same time, many AI researchers focus

on reaching "human parity" in narrow tasks as the main metric of success, which could create another powerful force towards automation, rather than using this platform for creating new tasks. Like other automation technologies, AI may also appeal to many executives intent on cost-cutting, and if there are additional tax breaks and favorable treatments for software in general and AI-related technologies specifically, these may exacerbate the focus on automation (Acemoglu and Johnson, 2022).

Overall, even though there is no definitive evidence on this question, it is possible that the direction of technological change was already tilted too much towards automation before AI, and AI may have exacerbated these trends. If so, its labor market implications could be one of the major harmful effects of AI.

The general lessons from this discussion are therefore:

1. AI could in principle be used for increasing worker productivity and expanding the set of tasks in which humans have a comparative advantage, rather than focusing mainly on automation. If it is used in this way, it may counterbalance some of the negative effects of automation on labor and may generate more positive welfare effects and beneficial distributional consequences

2. But there is no guarantee that the composition of technological change in general and the balance of AI between automation and more human-friendly activities should be optimal. In fact, there are many possible distortions, some of them economic and some of them social, encouraging excessive automation using AI.

## 3.3 AI and Human Judgment

The arguments in the previous two subsections are partly predicated on the notion that AI-based automation may not generate sweeping productivity gains, which could compensate for or even undo the displacement effects it creates. AI's most enthusiastic boosters, on the other hand, believe that AI can bring huge productivity gains. One of the most powerful arguments in this respect is that as AI helps automate and improve (both cognitive and noncognitive) tasks that do not require human judgment and creativity, it will increase the demand for problem-solving tasks that require creativity and judgment and also free workers to focus on these tasks. Although seemingly plausible, I now suggest a potential reason why this expectation may be too optimistic and argue that, even when such reallocation takes place, AI-based automation may be excessive.

Suppose that there are two tasks to be performed, 1 and 2. Overall output in the economy is given by

$$Y = \min\left\{y^1, y^2\right\},$$

where $y^i$ is the output of task $i$, and the Leontief production function imposes that these tasks are strongly complementary.

Before AI, both tasks have to be performed by humans. Suppose that there is a measure 1 of humans, each with 2 units of time. Suppose also that, for reasons I will explain below, we start in an allocation in which each human allocates half of their time to task 1 and the other half to task 2. In this case, they have equal productivity in both tasks, which I normalized to 1. As a result, before AI, the economy produces a total of one unit of the final good. We can think of each worker as a "yeoman-producer", consuming his or her production. Equivalently, we can think of this economy as consisting of firms hiring workers in a competitive labor market. In this case, the per hour wage of each worker in each task will be $1/2$, ensuring that the entire output is paid to workers.

Now imagine that there are advances in AI algorithms that produce the first task at per unit cost $c < 1/2$. This cost is paid in terms of the final good, and the fact that it is less than the equilibrium wage before AI implies that these algorithms are cost-saving and will be adopted. If this were the end of the story, AI would improve net output, because workers would be reallocated from task 1 to task 2, enabling the economy to increase its total output.

However, suppose that there are also economies of scope: individuals learn from performing both tasks at the same time (and that is why the pre-AI allocation involved each worker devoting half of their time to each task). Suppose, in particular, that if a worker does not learn from task 1, his or her productivity in task 2 declines to $1 - \beta$. The post-AI allocation will involve all workers working in task 2, and whatever their total production is in this task, the economy will also produce exactly the same amount of task 1, using AI algorithms. As a result, net output in this economy will be

$$2(1-\beta) - \text{spending on AI} = 2(1-\beta)(1-c).$$

It can be verified that in the special case where there are no economies of scope ($\beta = 0$), $c < 1/2$ is sufficient for net output to increase — in particular, from 1 to $2(1-c) > 1$. However, as soon as $\beta > 0$, this is no longer guaranteed. For example, when $c \approx 1/2$, even a small amount of economies of scope implies that the use of AI would reduce net output.

This simple example captures a more general phenomenon: a finer division of labor and the reallocation of some tasks away from humans can be cost-reducing, but to the extent that human judgment improves when workers gain experience from dealing with a range of problems and recognize different aspects of the problem, it may also come at a cost. When some aspects of the problem are delegated to AI, humans may start losing their fluency with and ability to understand the holistic aspects of relevant tasks, which can then reduce their productivity, even in tasks in which they specialize. An extreme example of this phenomenon can be given from the learning of mathematical reasoning. Calculators are much better than humans in arithmetic. But if students stopped learning arithmetic altogether, delegating all such functions to calculators and software, their ability to engage in other type of mathematical and abstract reasoning may suffer. For this reason, most mathematical curricula still emphasize the learning of arithmetic. If delegating certain tasks to AI becomes similar to students ceasing to learn arithmetic, it may have significant costs.

Would the market adopt AI when there are such economies of scope? The answer depends on the exact market structure. Because the adoption of AI technologies is associated with a finer division of labor, there is no guarantee that firms will internalize the economies of scope. For example, in the pre-AI equilibrium, the cost of one unit of task 1 is $1/2$. So a new firm can enter with the AI technology and make profits in this equilibrium. The entry of these firms would then create a pecuniary externality, discouraging other workers from working on task 1. In particular, even though there are economies-of-scope benefits from performing task 1, workers may not be allocated to task 1, because the price of this task is now lower due to the use of AI technologies.[5] This type of entry could then destroy the pre-AI equilibrium and would drive the economy to the post-AI equilibrium characterized above, even when it is inefficient because $\beta$ is large.

In summary, the general lessons from this short discussion are:

1. In addition to the costs of worker displacement discussed earlier in this section, economies of scope across tasks may create additional costs from the use of AI technologies. In particular, the deployment of AI in various cognitive tasks that do not require a high degree

[5] There are some market structures and pricing schemes that may prevent the adoption of AI technologies when they are inefficient in this case. For example, if workers can take a very low or even negative wage in order to work in task 1 (so as to increase their productivity in task 2), this may outweigh the cost advantage of new firms that enter and specialize in using AI in task 1. The issues are similar to the ones that arise in the context of firm-sponsored general training, and as in that case, labor and credit market imperfections would typically preclude the possibility that workers fully pay for all the benefits they receive by taking wage cuts (see Acemoglu and Pischke, 1999).

of human judgment and creativity may enable workers to reallocate their time towards tasks that involve judgment and creativity. But if economies of scope are important for human productivity, AI may have additional costs.

2. Cost-minimization incentives of firms may encourage them to use AI technologies in inefficient ways, when there are such economies of scope.

## 3.4 AI and Excessive Monitoring

Another use of AI-powered technologies is in worker monitoring, as exemplified by Amazon's warehouses and new monitoring systems for delivery workers. Here, too, employers' incentives to improve monitoring and collect information about their employees predates AI. But once again, AI may magnify their ability to do so. Some amount of monitoring by employers may be useful by improving worker incentives. However, I argue that increasing employer flexibility in this activity can also lead to inefficiently high levels of monitoring for a very simple reason: at the margin, monitoring is a way of shifting rents away from workers towards employers, and thus is not socially valuable.

I now develop this point using a model based on Acemoglu and Newman (2002). Consider a one-period economy consisting of a continuum of measure $N$ of workers and a continuum of measure 1 of firm owners, each with a production function $AF(L_i)$ where $L_i$ denotes the number of workers employed by firm $i$ who exert effort (the alternative, exerting zero effort, leads to zero productivity). Firms are large, so the output of an individual worker is not observable. Instead, employers can directly monitor effort in order to determine whether an employee is exerting effort and being productive.

Specifically, as in Shapiro and Stiglitz (1984) a worker exerting effort is never mistakenly identified as a shirker, and a shirking worker is caught with probability $q_i = q(m_i)$ where $m_i$ is the extent of monitoring per worker by firm $i$, with cost $Cm_iL_i$, where $C > 0$. Suppose that $q$ is increasing, concave and differentiable with $q(0) = 0$ and $q(m) < 1$ for all $m$. Suppose also that because of the limited liability constraints, workers cannot be paid a negative wage and will simply receive a zero wage. This implies a simple incentive compatibility constraint for workers,

$$w_i - e \geq (1 - q_i)w_i,$$

where $e$ denotes the cost of effort. Rearranging this equation we obtain:

$$w_i \geq \frac{e}{q(m_i)}. \tag{6}$$

In addition, the firm has to respect the participation constraint,

$$w_i - e \geq \underline{u}, \tag{7}$$

where $\underline{u}$ is the worker's *ex ante* reservation utility, given by what he could receive from another firm in this market.

The firm maximizes its profits, given by $\max_{w_i, L_i, q_i} \Pi = AF(L_i) - w_i L_i - C m_i L_i$ subject to these two constraints. As shown in Acemoglu and Newman (2002), the solution to this problem takes a simple form because the incentive compatibility constraint always binds — if it did not, the firm would reduce monitoring, increasing its profits. By contrast, the participation constraint (7) may or may not bind.

The main result from this framework relevant for my discussion is that, regardless of whether the participation constraint is binding, the equilibrium never maximizes utilitarian social welfare (total surplus), given by $Y = AF(L) - CmL - eL$, because there is always too much monitoring. The intuition is straightforward: at the margin, monitoring is used to transfer rents from workers to firms, and is thus being used excessively. Mathematically, this can be seen in the following way: start from the equilibrium and consider a small decline in monitoring coupled with a small increase in the wage so that (6) remains binding. This will have only a second-order impact on the firm's profits, since the firm was already maximizing them. But it will have a first-order benefit for workers, whose wage will increase. Thus, a utilitarian social planner would like to increase wages and reduce monitoring. This is true a fortiori if we care about income inequality (presuming, of course, that firm owners are richer than workers).

How does AI affect things in this equilibrium? Suppose that before AI, there was an upper bound on monitoring, so that $m \leq \bar{m}$, and AI lifts this constraint. If the equilibrium level of monitoring without this constraint, $m^*$, is above $\bar{m}$, then improvements in AI will lead to higher monitoring. If, in addition, $\bar{m}$ is not too low, then AI would reduce welfare (the qualifier that $\bar{m}$ should not be too low is included because, if it were very low, the initial equilibrium could be very inefficient).

The economic force here is much more general than the simple model I presented: AI, by enabling better control and use of information, provides one more tool to employers to shift rents away from workers to themselves, and this generally leads to inefficiently high levels of rent-shifting activities.

Is this potential inefficiency relevant? Once again, there is little systematic evidence to suggest one way or another, but the fact that the US labor market has a "good job" problem, and wages at the bottom of the distribution have fallen in real terms over the last several decades (Acemoglu, 2019; Acemoglu and Restrepo, 2021) suggests that it may be.

In summary, the general lessons from this analysis are:

1. AI technologies also create new opportunities for improved monitoring of workers. These technologies have first-order distributional consequences, because they enable better monitoring and thus lower efficiency wages for workers.

2. Because at the margin the use of monitoring technologies transfers rents from workers to firms, monitoring will be excessive in equilibrium. By expanding monitoring opportunities, AI may thus create an additional social cost.

# 4 AI, Political Discourse and Democracy

The 1990s witnessed rapid strengthening of democracy around the world, in a pattern the political scientist Samuel Huntington (1991) called "The Third Wave". During this process, many Latin American, Asian and African countries moved from nondemocratic regimes towards democracy and several others strengthened their democratic institutions (Markoff, 1996). The last decade and a half has witnessed a pronounced reversal of this process, however. Several countries moved away from democracy, and perhaps even more surprisingly, Democratic institutions and norms have come under attack in numerous Western nations (Levitsky and Ziblatt, 2017; Snyder, 2017; Mishra, 2017; Applebaum, 2020) and the population appears to be more polarized than in the recent past (Abramowitz, 2010; Judis, 2016). Some have pointed to social media and online communication as major Contributing factors (e.g., Marantz, 2020). I now turn to a discussion of these issues. I focus on the effects of AI on communication, political participation and democratic politics. As indicated in the Introduction, I will highlight several distinct but related mechanisms via which AI might degrade democratic discourse.

## 4.1 Echo Chambers and Polarization

Social media is often argued to promote echo chambers, in which individuals communicate with others who are like-minded, and this might prevent them from being exposed to counter-attitudinal viewpoints, exacerbating their biases. Cass Sunstein noted the potential dangers

of echo chambers as early as 2001. He stated that encountering individuals with opposing opinions and arguments is "important partly to ensure against fragmentation and extremism, which are predictable outcomes of any situation which like-minded people speak only with themselves", and emphasized that "many or most citizens should have a range of common experiences. Without shared experiences, a heterogeneous society will have a much more difficult time in addressing social problems." (Sunstein, 2001, p. 9). The recent documentary *The Social Dilemma* describes the situation as "The way to think about it is as 2.5 billion Truman Shows. Each person has their own reality with their own facts. Over time you have the false sense that everyone agrees with you because everyone in your news feed sounds just like you." AI is crucial to this new reality on social media. For example, the algorithms that sites such as Facebook and Twitter use in deciding what types of news and messages individuals will be exposed to is based on applying AI techniques to the massive amount of data that these platforms collect (Alcott and Gentzkow, 2017; Guriev, Henry and Zhuravskaya, 2020; Mosleh et al., 2021). Recent studies document that these algorithmic approaches are exacerbating the problem of misinformation on social media, for example by creating algorithmic "filter bubbles", whereby individuals are exposed much more frequently to news that agrees with their priors and biases (Levy, 2021). As a result, Vosoughi et al. (2018) conclude that on social media there is a pattern of "falsehoods diffusing significantly farther, faster, deeper, and more broadly than the truth in all categories of information".

In this subsection, I discuss these issues building on the approach in Acemoglu, Ozdaglar and Siderius (2021), which suggests that social media platforms may endogenously create such echo chambers, and do so more when there is more extremist content around. Consider an online community in which each individual receives a news item that is circulating (either from an outside source or from other members of the community). The news item may contain misinformation (which could be downright fake news or some misrepresentation of facts). The individual decides whether to share the news item she receives.

Each agent receives utility from sharing news items. However, if she is found out to have shared misinformation, she incurs a cost. To avoid this, she may instead decide to inspect the news item to check for misinformation and can then decide to kill it. Inspecting is costly. Thus, all else equal, the individual is more likely to inspect and less likely to share a news item if she thinks it contains misinformation.

Suppose that there are two sub-communities, one is left-wing and the other is right-wing, meaning that one community consists of individuals with priors to the left of some reference

point, normalized 0, and the other one consisting of individuals with priors to the right of 0. For simplicity, I suppose that the prior of each left-wing community member is the same, some $b^L < 0$, and the prior of each right-wing community member is also the same, $b^R > 0$. Each member of one of these communities is much more likely to be connected to and sharing items with other members of her community, but there are also some cross-community links. The extent of these cross-community links determines how much "homophily" there is. With high homophily, there are almost no cross-community links. With low homophily, cross-community links are more common.

Each news item has a type consisting of: (1) a news-related message, which is simply taken to be $m \in \{L, R\}$, and (2) a reliability score $r \geq 0$, whereby a high reliability score means that the news item is very unlikely to contain misinformation, while a low reliability score means misinformation is quite likely. Although this setup is a simplified version of the one in Acemoglu, Ozdaglar and Siderius (2021), a full analysis of the model is still involved. Here I summarize the main features of the equilibrium.

Suppose that individuals do not observe the provenance, source and history of the message and do not know whether the message was inspected by others in the past.

Consider a right-wing agent receiving a left-wing message. Given her prior, she is much more likely to think that this message contains misinformation than a right-wing message. Therefore, all else equal, she is much more likely to inspect it. This effect is further magnified when she expects to share it with other fellow right-wingers, because they are also more likely to inspect a left-wing message, and if it contains misinformation, it is likely to be found out, with costly implications for our focal agent.

In contrast, consider this right-wing agent receiving a right-wing message. Now she is less suspicious of the message, and all else being equal, she will use a lower threshold — this means that she will choose a lower threshold $\bar{r}$ such that she inspects it only if the reliability score is $r < \bar{r}$. Homophily now has the opposite effect: if she is in a homophilic network, she expects other right-wingers not to inspect the news item either, and as a result, she is less likely to inspect it herself. When an item is not inspected, then it is more likely to become viral — everybody starts sharing it.

These observations lead to the first basic result in the setup: a news item is more likely to become viral when (i) it reaches individuals who have congruent beliefs, and (ii) when this concordant news item is being shared with others with similar beliefs.

Now consider the problem of the platform on which this community is situated. Suppose

that, via its choice of algorithms, the platform influences the degree of homophily (as in Acemoglu, Ozdaglar and Siderius, 2021). Suppose also that the platform's objective is to maximize engagement and hence, greater virality of the article is better for the platform. First suppose that most of the news items arriving from the outside have high reliability scores and are also being distributed roughly equally between the two sub-communities (so that it is not only left-wing messages going to the left-wing community and likewise for the right-wing community). Then the platform's engagement-maximizing policy is likely to be to introduce between-community links, thus exposing each individual to news items from the other side, since these are all reliable and thus unlikely to be killed, even if they were inspected.

In contrast, suppose we have a situation in which there are many low-reliability news items, and moreover left-leaning news items from the outside go to the left-wing community and right-leaning news items go to the right-wing community. If the platform were interested in stopping misinformation, it could choose low homophily (so that right-wing articles go to the left-wing community as well, for example), and this would induce less sharing among the right-wing agents (expecting inspection from the left-wingers). It would also cause interruptions to virality when left-wingers discover the right-wing news item to contain misinformation (which is likely since it has low-reliability). These considerations imply that when news items have lower reliability, the platform would prefer to induce extreme homophily via its algorithms and propagate misinformation in order to maximize engagement.

This result, mimicking the main finding of Acemoglu, Ozdaglar and Siderius (2021), is in some ways quite striking. It highlights that the platform has incentives to create endogenous echo chambers (or filter bubbles). Worse, this happens precisely when there are low-reliability news items likely to contain misinformation and distributed within the community in a "polarized fashion" (left-wing messages going to left-wing groups, etc.).

The role of AI technologies is again crucial. Without these technologies, the platform would not be able to determine users' biases and create relatively-homogeneous communities. It would also not be able to support the rapid dissemination of viral news items.

Broader social implications of these types of filter bubbles are easy to see as well. Suppose that individuals also update their beliefs on the basis of the news items. When left-wingers receive right-wing news items and the relevant news items have reasonable reliability scores, they will tend to moderate their beliefs, exactly as Sunstein (2001) envisaged in the quote above. In contrast, when left-wingers receive only left-wing news items in a filter bubble, this might lead to further polarization. On the basis of these considerations, Acemoglu, Ozdaglar

and Siderius (2021) suggest various interventions, including regulation and outside inspections, in order to discourage such filter bubbles and reduce polarization, and I return to the issue of regulation in Section 6.

In summary, this section leads to the following general lessons:

1. AI-powered social media presents a variety of new opportunities for connecting individuals and information sharing.

2. However, in the process it may also distort individuals' willingness to share unreliable information. When social media creates echo chamber-like environments, in which individuals are much more likely to communicate with like-minded others, they become less careful in inspecting news items that are consistent with their existing views and more willing to allow the circulation of misinformation.

3. Centrally, social media platforms that are focused on maximizing engagement have an incentive to create echo chambers (or "filter bubbles"), because inspection and interruptions of the circulation of news items with unreliable messages reduces engagement. As a result, especially when there are more items with misinformation, platform incentives are diametrically opposed to social objectives.

## 4.2 Perils of Online Communication

The previous subsection argued that political discourse may be hampered because of the algorithmic policies of digital platforms. In this subsection, I suggest that there might be reasons more endemic to the nature of online communication that disadvantages political communication (see also Lanier, 2018; Tirole, 2021). The main argument is a version of the ideas developed in Acemoglu, Kleinberg and Mullainathan (2022), and here I will provide a simple illustration.

Consider first a pre-AI world in which large-scale social media communication is not possible. Suppose that in such a world, individuals communicate bilaterally in a social network. For simplicity, we can imagine a social network that takes the form of a directed line, in which we start with individual 0, who can communicate with individual 1, and all the way up to the end of the line, individual $n$. Suppose that each individual has a piece of gossip which they can share with their neighbor, which will give utility $v^G$. In addition, the individual may have some news relevant for the political/social beliefs of the community. If she decides to

share this and this news item alters the beliefs of her neighbor, she also receives utility from such persuasion, as I specify below (for simplicity, I am ignoring potential utility benefits from non-neighbors indirectly changing their beliefs as this information is shared further). Because there are "channel constraints" (for example, limited ability to communicate), the individual can either share political news or gossip, but not both.

Specifically, suppose that the state of the world is 0 or 1 (Left or Right), and all individuals start with prior $\mu_0 = \frac{1}{2}$ about the underlying state being $s = 0$. Suppose that individual 0 receives a piece of news that shows that the underlying state is in fact $s = 0$. If she shares this information with her neighbor (individual 1) and her neighbor believes it, then her neighbor's belief will also shift to $\mu = 1$. However, each individual is also concerned that some agents in society may have ulterior motives and try to convince them that the state is the opposite of the true state $s$. Suppose the probability that each individual attaches to their neighbor being of this extremist type is $q$. Then the posterior of individual 1 that the state is $s = 0$ after receiving this news will be

$$\mu_1 = \frac{\frac{1}{2}}{\frac{1}{2} + \frac{1}{2}q} = \frac{1}{1+q} > \frac{1}{2}.$$

Finally, I assume that individuals receive additional utility from shifting their neighbor's beliefs towards the truth (or her own belief), so the overall utility of individual $i$ is

$$v^G x_i^G + v^N \left| \mu_{i+1} - \mu_i \right|,$$

where $x_i^G = 1$ denotes whether this individual gossips, $\mu_i$ is her belief, and $\mu_{i+1}$ is her neighbor's belief (given the line network).

Let us assume that $v^N > 2v^G$, so that if an individual is convinced that her information will be believed and can thus shift her neighbor's belief from 1/2 to her views, she prefers to share the political information rather than gossip. Therefore, there exists $\bar{q}$ such that if $q < \bar{q}$, the individual will share the political news. Let us suppose that in the in-person social network this inequality is satisfied.

In this scenario, individual 0 will start sharing the political news. This will convince individual 1, who will then attach a sufficiently high probability to the underlying state being $s = 0$. With the same argument, she would also like to convince her neighbor to the right, individual 2. If $q$ is sufficiently small, individual 2, even when she is worried about the possibility that either individual 0 or individual 1 being an extremist, would still believe this information. In general, there will exist some $n(q)$, such that the political news will be communicated up to

individual $n(q)$, and then after that there will be pure gossip on the network. For $q$ sufficiently small, the entire network might share the political news.

Next suppose that we go to online communication. This leads to a larger network and less personal contact. As a result, it is plausible to assume that the probability that each agent attaches to the event that the person communicating with them is an extremist is now higher, say $q' > q$. If $q' > \bar{q}$, then in online communication, there will be no news exchange and all communication will be gossip.

The situation may be worse when we take into account that online interactions typically take the form of broadcast (rather than bilateral) communication. This would exacerbate the situation I have outlined here for two reasons. First, there may be many agents who may want to broadcast their views, and the same "channel constraints" would imply that only one of them can do so, and broadcasting may be particularly attractive to extremists. This would endogenously increase the assessment of all the agents that political news is coming from extremists. Second, if there is heterogeneity in the utility of gossip across agents, those who value gossiping most may be the ones monopolizing the channels. In both cases, online communication becomes less effective as a way of sharing politically or socially relevant information. If political communication and news sharing in social network is an important aspect of democratic politics, then the forces identified in the subsection also create new challenges for democracy, again rooted in the use of AI technologies.

The general lessons from this brief analysis are simple as well:

1. Bilateral, off-line communication, especially when the subject matter is political or social, relies on trust between parties. Naturally-existing trust in in-person social networks may enable this type of communication.

2. When communication is taking place online and in multi-lateral settings, such as in modern social media platforms powered by AI technologies, this type of trust-based communication becomes harder. This may favor non-political messages, such as gossip, which then drive out political communication, with potentially deleterious effects for political discourse and democracy.

3. This potential barrier to online communication is exacerbated when there is competition for attention, which is encouraged by the broadcast or multi-lateral nature of online communication.

## 4.3 Big Brother Effects

The previous two subsections focused on how AI-powered social media and online platforms change the nature of communication, with potentially negative effects on the sharing of political information, which is the bedrock of democratic participation by citizens. In this subsection, I suggest that the other crucial pillar of democratic institutions, citizen protests, is likely to be hampered by AI technologies.

I have argued in Acemoglu and Robinson (2000, 2006) that protests, riots and uprisings are critical for the emergence of democratic regimes (because the threats that they pose for power-holders in nondemocratic regimes induces democratization). This argument is relevant for democratization in currently authoritarian governments, such as China, Russia or Iran. In particular, if AI-based monitoring of communication and political activity extinguishes political dissent and makes it impossible for opposition groups to organize, the longevity of nondemocratic regimes may increase significantly.

The problem is not confined to these nondemocratic nations, and applies to the US and other countries with democratic institutions as well. A similar argument suggests that protests and civil disobedience are often critical for the functioning of democratic regimes as well. The civil rights movement in the US illustrates this vividly. Even though the US was democratic at the federal level, the Jim Crow South routinely violated the political, social and economic rights of Black Americans. Democratic institutions in the North and, to the extent that that they existed in the South, did not create a natural impetus for these discriminations to cease. The turning point came with civil disobedience organized by various Black (and later multi-ethnic) civil society groups, such as the NAACP. Vitally, even federal politicians opposed to Jim Crow were not in favor of these protests initially, viewing them as disruptive and a political drawback for them, especially given that any federal action against Jim Crow practices would trigger backlash from Southern politicians (see the discussion in Acemoglu and Robinson, 2019). Without civil disobedience, protest and other sources of bottom-up pressures, it is likely that reform of voting, civil, education and discrimination laws in the US South would have been further delayed.

Here I briefly outline a simple model capturing some of these ideas. Consider a society consisting of $\lambda < 1/2$ elites and measure 1 of regular citizens. All citizens and all elites have the same economic preferences, but citizens are heterogeneous in terms of their cost of participating in protest activities, denoted by $c_i$ for individual $i$. I assume that $c$ is distributed uniformly over $[0, 1]$ in the population.

The political system is an imperfect democracy or an autocracy in which political choices are biased in favor of the preferences of the elite. In particular, suppose that there is a unique, one-dimensional policy, and the preferred policy choice of the citizens is 0, while the most preferred policy of the elite is $p^E > 0$. Consider the following reduced-form political game. The elite decide a policy $p$, and then protests take place. If a fraction $q$ of the citizens protest and engage in civil disobedience, then there is probability $\pi(q)$ that the policy will switch from $p$ to 0. With the complementary probability, the policy stays at $p$. This political structure ignores the influence of the citizens via democratic institutions, which is for simplicity. If this is incorporated, for example, as in Acemoglu and Robinson (2008), this would not affect the main message of the model presented here.

I also assume that the government can impose a punishment on those engaged in protests. Suppose, in particular, that the state has the capacity to detect at most a measure $\psi$ of protesters. If the total amount of protest, $q$, is less than $\psi$, then all protesters are detected and can be punished. I assume that the punishment imposed on protesters is a constant, $\Gamma$, independent of the number of protesters.

The two key economic decisions are therefore policy choice by elites and protests by citizens. Let me first describe the utility of the citizens. Suppose that when the policy choice of elites is $p$ and there are $q$ protesters in total, individual $i$ has the following utility as a function of her protest decision $x_i \in \{0,1\}$:

$$U_i^C(p,q,x_i) = \left[ \left( v^C |p| - c_i \right) - \min \left\{ \frac{\psi}{q}, 1 \right\} \Gamma \right] x_i,$$

where, for centricity, I have ignored components of the utility that depend on policy choice but are independent of the individual's process decision. Intuitively, the utility from protesting is increasing in the distance between the actual policy and the bliss point of citizens, as captured by $v^C |p|$ (recall that their bliss point is at zero). In addition, the individual incurs the cost of participating in protests, given by $c_i$. The second term in square brackets captures the expected punishment from protesting, taking into account that when $q \leq \psi$, protesters will be punished with probability 1.

Clearly, there exists a threshold value $\bar{c}$, such that only individuals with $c_i \leq \bar{c}$ will participate, and thus $q = \bar{c}$, since the distribution of $c$ is uniform between 0 and 1.

Let us next turn to the elite's utility. Suppose that this is given by

$$\begin{aligned} U^E(p,q) &= -\mathbb{E}\left[ \left| \hat{p} - p^E \right| \right] \\ &= -(1-\pi(q)) v^E \left| \hat{p} - p^E \right| - \pi(q) v^E \left| p^E \right|, \end{aligned}$$

where $\hat{p}$ denotes the realized policy and the expectation is over the uncertainty concerning whether protests will force the elites to change policy. As usual, the subgame perfect equilibrium can be solved by backward induction. In the second stage, $\bar{c}$ is determined such that given the policy choice of elites, $p$, we have:

$$(v^C |p| - \bar{c}(p)) - \min\left\{\frac{\psi}{\bar{c}(p)}, 1\right\}\Gamma = 0. \tag{8}$$

In general, there can be multiple equilibria in this stage, because this equation might have multiple solutions for $\bar{c}(p)$. Note in particular that its left-hand side may be non-monotonic. In what follows, I focus on the case in which it is monotonically decreasing, which ensures a unique equilibrium (if there are multiple equilibria, we could pick the one with the highest amount of protest, which will necessarily be one where the left-hand side is decreasing, yielding the same results). It is then straightforward to see that $\bar{c}(p)$ is increasing in $p$, meaning that a more pro-elite policy induces more protests.

Now turning to the elite's maximization, we first rewrite the elite's utility function taking into account the reaction of the citizens to their policy choice:

$$U^E(p, \bar{c}(p)) = -(1 - \pi(\bar{c}(p)))\left|p - p^E\right| - \pi(\bar{c}(p))\left|p^E\right|,$$

which simply substitutes $q = \bar{c}(p)$. We can therefore maximize elite utility by choosing the initial policy $p$. Taking into account that $p < p^E$, this maximization problem yields a standard first-order condition:

$$(1 - \pi(\bar{c}(p))) - \pi'(\bar{c}(p))\bar{c}'(p)(2p^E - p) = 0, \tag{9}$$

and under suitable assumptions, we can ensure that the second-order condition for maximization is satisfied. In this first-order condition, $\bar{c}'(p)$ is given from (8), and under the assumption that $\psi < \bar{c}(p)$, it can be written as

$$\bar{c}'(p) = \frac{v^C}{1 + \psi\Gamma/\bar{c}(p)^2}.$$

Consider next the introduction of AI, modeled as an increase in $\psi$ to some $\psi'$. From the previous expression, this increase will reduce $\bar{c}'(p)$, and from (9), this reduction in $\bar{c}'(p)$ will increase $p$ away from the citizens' bliss point and towards the elite's preferences. Intuitively, AI-induced government monitoring of protests weakens citizens' ability to force the elites to make concessions, and the elite respond to the deployment of the AI technology by withdrawing concessions. As a result, AI makes policies less responsive to citizens' wishes, and to the

extent that these policies impact the distribution of resources, it will also tend to favor the elite's economic interests and increase inequality.

Overall, the general lessons from this simple model are:

1. AI technologies can be used for improving government monitoring against protest activities.

2. Since the threat of protests has a disciplining role on nondemocratic governments, and even on some democratic governments, the shift of power away from civil society towards governments will weaken democracy and aggravate policy distortions.

## 4.4 Automation, Social Power and Democracy

The previous subsection explained how the use of AI as a tool for controlling society and political dissent can have harmful effects on democratic politics. In this subsection, I argue that AI-powered automation can further weaken democracy and undermine social cohesion.

To develop this argument, let us first go back to the framework in Acemoglu and Robinson (2006), which emphasizes two types of conditions for the emergence and survival of democratic institutions. First, there must be enough discontent with nondemocratic regimes to generate a demand for democracy. Second, democracy should not be too costly for the elite, who would otherwise prefer repression or other means to avoid sharing political power with the broader population. One aspect of this problem, which could be important but was not analyzed in Acemoglu and Robinson (2006), is cooperation from workers. For example, if workers become disenchanted with the current regime or decide that they need to take action against current (economic or political) power-holders, they may choose not to cooperate with capital in their workplaces. When labor is essential for production, this withdrawal of cooperation could be very costly for capital. When capital owners are influential in the political system, they may then push for democratization or redistribution in order to placate labor.

The main argument in this subsection is that automation may make workers less indispensable in workplaces, and as such, it will tend to reduce their political power.

Let us return to the model in Section 3.1 and for simplicity, suppose that $N = 1$ and $\sigma = 1$ in the production function (3) and also assume that labor markets are competitive (there is no wage floor). This implies that the equilibrium level of production as a function of capital and labor can be written as

$$Y(K, L) = K^I L^{1-I},$$

and thus with competitive labor markets, the labor share is $s^L = 1 - I$, and

$$w = (1 - I)\frac{Y(K, L)}{L}.$$

Note also that in this case the impact of automation on output can be written as

$$\frac{\partial \ln Y(K, L)}{\partial I} = \ln K - \ln L.$$

Therefore, automation is output-increasing, when $\ln K > \ln L$ (or $K > L$, which is also equivalent to the competitive rental rate of capital being less than the wage, ensuring that automation is cost-saving). Conversely, low-productivity("so-so") automation now corresponds to the case in which $K \approx L$.

Consider a political system as in Acemoglu and Robinson (2006), where all capital owners (capitalists) are elites and all workers are non-elites, with no within-group heterogeneity. To start with, let us consider a nondemocratic regime in which the capitalists hold power, or alternatively a democratic regime in which they have disproportionate power. For brevity, I am going to ignore any threat of revolution or protests along the lines of the models considered in Acemoglu and Robinson (2006) and will also abstract from the considerations discussed in the previous subsection.

Suppose in addition that there is a lump-sum tax on capitalists, which can be redistributed to workers, and let us denote the per-worker transfer on the basis of this by $\tau$. Suppose that the workers have an aspiration for a level of net income given by $w^A$, and if $w + \tau < w^A$, they withdraw their cooperation, and as a result, the effective productivity of labor declines from 1 to $\delta < 1$. In this reduced-form model, I interpret the transfer $\tau$ from the elite both as a measure of redistributive politics and also as a general concession to democratic politics (for example, allowing workers to have more voice or making less use of lobbying and other activities in order to distort democratic politics).

The key question is whether the elite will make the necessary transfers to convince workers to continue cooperating in workplaces. This boils down to the comparison of the following two options for the elite: redistribute via $\tau$ so that the aspirations of the workers are met, or make do with lower labor productivity due to lack of cooperation. Let us write the payoffs to the elite from these two strategies. Suppose there is no within-elite heterogeneity, so it is sufficient to look at the overall income of capital. When the capitalists choose to meet the aspiration constraint of workers, their payoff is

$$U_1^K = K^I L^{1-I} - \max\left\{(1 - I)K^I L^{1-I}, w^A L\right\},$$

where the first term in the max operator applies when the market wage is already greater than $w^A$, while the second term is when they have to make transfers in order to bring workers to this level. Clearly, the relevant case for the discussion here is the latter, so I assume that $(1-I)K^I L^{1-I} < w^A L$, and thus

$$U_1^K = K^I L^{1-I} - w^A L.$$

The alternative is not to make the necessary transfer, in which case the elite simply receive the market return to capital when the productivity of labor has been reduced to $\delta$, i.e.,

$$U_2^K = K^I (\delta L)^{1-I}.$$

What is the effect of automation on the comparison of these two strategies? Differentiating the difference in their payoffs, $U_1^K - U_2^K$, with respect to $I$ yields

$$\frac{\partial (U_1^K - U_2^K)}{\partial I} < 0 \text{ if and only if } \ln K - \ln L + \frac{\ln \delta}{1-\delta^{1-I}} < 0.$$

This expression has two immediate implications. First, if we have low-productivity automation, so that $\ln K \approx \ln L$, then because $\ln \delta < 0$, automation always makes the second strategy of stopping redistribution and foregoing labor's cooperation more attractive. Intuitively, automation makes labor less central for production, and thus losing its cooperation becomes less costly for capital. The economic force going against this calculation is that when automation increases productivity (when $\ln K > \ln L$), the output loss due to lack of cooperation from labor becomes more costly. Second, however, we can also see that, for any $K$ and $L$, there exists a threshold level $I^*$ such that once $I > I^*$, the second strategy is preferred, even taking into account the productivity gains from automation. Therefore, automation makes cooperation from workers less important, and to the extent that securing this cooperation was an important part of the motivation for redistribution and democratic politics, automation may make elites turn their back on or even become hostile to democracy.

In summary, this subsection has established:

1. Automation can also generate an indirect negative impact on democracy and redistributive politics when ensuring cooperation from labor in workplaces is an important motivation for elites to make concessions to labor.

2. When automation brings only small productivity gains, it encourages the elite to reduce redistribution and make fewer democratic concessions. This will make policies less responsive to the majority's wishes and may further raise inequality.

3. Productivity benefits of automation may soften this effect, because an automation-driven increase in output raises the opportunity cost of losing labor's cooperation. Nevertheless, there exists a sufficiently high level of automation such that once we reach this level, labor becomes sufficiently irrelevant for production that the withdrawal of their cooperation ceases to be very costly. After this threshold, the elite may prefer to abandon democratic institutions and withhold any concessions, and proceeds without labor's cooperation, once again with harmful effects on democracy, redistribution and social cohesion.

# 5 Other Potential Costs

In this section, I briefly list a few other areas that may be important, but without providing as much detail or any formal analysis.

*The threat of AI and bargaining:* Even when the actual labor market effects of AI discussed in Section 3 are not realized, the threat of adopting AI technologies may influence wages and inequality. Specifically, if there is bargaining and rent-sharing, employers may use the threat of AI-based automation as a way of increasing their bargaining power, which could have some of the same effects as actual automation and AI-powered monitoring.

*Discrimination:* Bias in AI has already received considerable attention. As AI gains greater importance in our social and economic life, ensuring that popular algorithms are fair and unbiased has become vital. Existing studies show that simple AI algorithms can improve important public decisions, such as bail or sentencing, without increasing discrimination (e.g., Kleinberg et al., 2018). However, in most such applications, algorithms use data generated from biased agents and potentially discriminatory practices (e.g., Thompson, 2019). For example, both the police and the legal system in the US are generally thought to be biased against certain groups, such as Black Americans. In such situations, there is a danger that these biases will become a fundamental part of AI algorithms. This may not only promote persistent bias and discrimination, but may in fact cement these biases more deeply in society via a process similar to the "signaling role of laws" (e.g., Posner, 2002). Indeed, if society starts trusting AI algorithms, their discriminatory choices may come to be accepted as more justifiable than when they were made by individual decision-makers.

*Technocracy versus democracy:* Advances in AI may create the temptation to delegate more and more public and even political decisions to algorithms or to technocrats designing and using

these algorithms. Although this may be justifiable for certain decisions, excessive reliance on technocracy, without citizen input, may also start encroaching into political decisions — such as the extent of redistributive taxation or how much we should protect disadvantaged groups (Sandel, 2020). In this case, reliance on AI may further undermine democracy, amplifying the concerns highlighted in the previous section.

*AI-powered weapons:* AI technologies have already started being incorporated into weapons and are advancing towards autonomous weapon systems. These new technologies will cause a host of ethical and social dilemmas, and may need to be regulated before prototypes are deployed or even fully developed. In addition to these ethical and social issues, AI-powered weapons may further strengthen governments against civil society, protesters and even some opposition groups, adding to the concerns we discussed in the previous section.

*The alignment problem:* The potential downside of AI technologies that has received most attention is the "alignment problem": the problem of ensuring that intelligent machines have objectives that are aligned with those of humanity. Many researchers and public intellectuals are concerned about machines reaching super-human capabilities and then implicitly or explicitly turning against humans (e.g., Bostrom, 2014, Russell 2019, Christian, 2019). Although my own view is that these concerns are somewhat overblown and often distract from shorter-term problems created by AI technologies (on which this essay has focused), they deserve careful consideration, monitoring and preparation.

*The international dimension:* The current development of AI technologies is intertwined with international competition, especially between the US and China (Lee, 2018). A discussion of regulation of AI has to take into account this international dimension. For example, it may not be sufficient for the US and Europe to start regulating the use of data or excessive automation, when they remain almost completely unregulated in China. This suggests that the regulation of AI needs to have a fully-fledged international dimension and we may need to build new international organizations to coordinate and monitor deregulation of AI across the globe.

# 6 The Role of Technology Choice and Regulation

In the preceding sections, I went through a number of theoretical arguments suggesting that the deployment of AI technologies may generate economic, political and social costs. In this

section, I highlight that in all of these cases the problems are not inherent to AI technologies per se. Rather, the harms I have emphasized are caused by corporate and societal choices on how these technologies are deployed. Even though these costs are far-ranging, taking place in product markets, in labor markets and in the realm of politics, they exhibit a number of commonalities, which I explore in this section. I also discuss some possible remedies. The general emphasis in this section will be on three main ideas:

1. The importance of choices, both on how existing AI technologies are used and on the direction of AI research. The costs I have modeled are not in the nature of AI technologies, but depend on how this new technological platform is being developed to empower corporations and governments against citizens and workers.

2. The inadequacy of market solutions that mainly rely on increasing competition.

3. The need for regulation.

Let me start with the mechanisms discussed in Section 2. All three of these potential costs of AI turn on how AI technologies enable the use and control of data. In each one of these cases, a different way of distributing control rights over data would ameliorate or prevent most of the costs (Posner and Weyl, 2019). Let us start with the model of data markets in Section 2.1. The source of inefficiency in this case is the ability of platforms to find out information about others from the data that an individual shares. This then opens the way to potential misuses of data, for example in order to reduce the surplus of consumers or by violating their privacy in other ways. Effective regulation in this case could take one of two forms. First, as suggested in Acemoglu et al. (2021), it may be possible to strip away parts of the data of an individual in order to prevent or minimize information about others being leaked (though the details matter here, and just anonymizing data would not be useful). Second, more systematic regulations on how platforms can utilize the information they acquire would lessen the harmful effects working through privacy.

In contrast, increasing competition may not be sufficient, and not even useful, in this case. My analysis in Section 2.1 focused on a monopoly platform. Acemoglu et al. (2021) show that if there are two platforms that are competing to attract users, this may exacerbate the pernicious effects of data externalities.

Let me then turn to the model considered in Section 2.3. The source of inefficiency in this case is the ability of platforms to use the data individuals reveal about themselves in order to

manipulate their weaknesses. If this misuse of data can be prevented or if consumers can be made more aware of how data are being used, some of these costs could be prevented. Suppose, for example, that consumers are informed frequently that platforms know more about their preferences and sometimes market products that are bad for them. There is no guarantee that such informational warnings will work for all consumers, but if they are displayed saliently and are specific (for example, calibrated according to the group of individuals and relevant class of products), they may prevent some of the harms identified by the analysis above. In this case, too, increasing competition would not be an effective solution. If two platforms are competing for consumers, but consumers continue to be semi-behavioral and fail to recognize the increase in platform capabilities, both platforms may try to exploit their ability to offer products that have short-term apparent benefits and long-term costs.

The issues are similar when we turn to the economic forces studied in Section 2.2, but now the implications of competition are more nuanced. In this case, effective regulation would prevent one of the firms from using the additional information it acquires for capturing all of the consumer surplus. Price controls and limits on price discrimination might be some of the methods for achieving this, though clearly such regulation is far from straightforward. What happens if we can increase competition in this case? Greater competition that results from firm 0 also using AI methods to estimate its own past consumers' preferences and customize its services accordingly would not be necessarily useful. Now both firms become local monopolists, capturing all of the consumer surplus. However, if each firm can also acquire information about the other's customers and collusion can be prevented, then they can be induced to compete fiercely, from which consumers might benefit. This case thus highlights that in some scenarios fostering competition might have benefits, though even in this instance, it can only do so to a limited extent and only when some case-specific conditions are satisfied.

I would also like to emphasize the implications for the direction of AI research in this case. Suppose that AI researchers can devote their time in order to develop alternative applications of this broad technological platform. For example, some of them may be able to use AI to create tools that empower citizens or consumers, or develop new technologies for preserving privacy. All the same, if any one of the mechanisms related to the control and misuse of information are relevant, then this will also produce a powerful demand for technologies that enable corporations to acquire and better exploit this type of information. This is more so when the ability of consumers to pay for alternative technologies is limited relative to the resources in the hands of corporations. In such scenarios, the demand for "misuse of AI" will

be transmitted to AI researchers, who may then respond by devoting their time to developing the AI technologies that corporations need and by moving further away from technologies that may have greater social value or empower consumers and citizens. This is a general point, which applies whether the harmful effects of AI are on the control of information, in labor markets or in the context of politics. It is for this reason that innovation, when it is itself unregulated, is unlikely to produce self-correcting dynamics. On the contrary, the demand for misuse of AI will typically distort the allocation of research across different applications, amplifying its social and economic costs.

The same considerations apply even more evidently in the models I discussed in Section 3. If automation is excessive, increasing competition in the labor market would not be particularly useful. On the other hand, the demand for automation technologies from firms will tend to be strong, encouraging researchers to double down on using AI for developing automation technologies. Regulatory solutions are again feasible, but may be more difficult to design and implement in this case. In theory, when automation is excessive and AI research is not being directed to creating new tasks, welfare-promoting regulation should discourage automation at the margin and encourage the creation of new labor-intensive tasks.

However, distinguishing marginal (low-productivity) and infra-marginal (higher-productivity) automation is difficult. Even more challengingly, regulators might have a hard time separating AI used for creating new tasks from AI being used for automating low-skill tasks while empowering higher-skilled or managerial workers. But it is also possible to view these problems not as absolute barriers but as measurement challenges. More research might shed light on how to distinguish different uses of AI in the labor market and might reveal new regulatory approaches for influencing the direction of AI research.

Finally, there are similar lessons from the models we discussed in Section 4, though there are also some new challenges related to the fact that the effects are now on political and democratic outcomes. For one, increasing competition is unlikely to be a very effective way of dealing with misaligned platform incentives. For example, if there are multiple social media platforms trying to maximize engagement, each may have incentives to create filter bubbles. Pro-competitive solutions may also be less effective when there are systemic issues, such as the widespread malfunctioning of democratic institutions.

The effects of these new technologies for democratic politics raises new conceptual issues as well. Most importantly, if the incorrect deployment of AI technologies is weakening democratic politics, developing after-the-fact regulatory solutions might become harder, since democratic

scrutiny of those who benefit from the distortionary use of AI technologies would also becomes more difficult. These considerations then suggest a "precautionary regulatory principle" — ex ante regulation slowing down the use of AI technologies, especially in domains where redressing the costs of AI become politically and socially more difficult after large-scale implementation. AI technologies impacting political discourse and democratic politics may be prime candidates for the application of such a precautionary regulatory principle.

# 7 Conclusion

In this essay, I explored several potential economic, political and social costs of the current path of AI technologies. I suggested that if AI continues to be deployed along its current trajectory and remains unregulated, then it can harm competition, consumer privacy and consumer choice, it may excessively automate work, fuel inequality, inefficiently push down wages, and fail to improve productivity. It may also make political discourse increasingly distorted, cutting one of the lifelines of democracy. I also mentioned several other potential social costs from the current path of AI research

I should emphasize again that all of these potential harms are theoretical. Although there is much evidence indicating that not all is well with the deployment of AI technologies and the problems of increasing market power, disappearance of work, inequality, low wages, and meaningful challenges to democratic discourse and practice are all real, we do not have sufficient evidence to be sure that AI has been a serious contributor to these troubling trends.

Nevertheless, precisely because AI is a promising technological platform, aiming to transform every sector of the economy and every aspect of our social lives, it is imperative for us to study what its downsides are, especially on its current trajectory. It is in this spirit that I discussed the potential costs of AI this paper.

My own belief is that several of these costs are real and we may see them multiply in the years to come. Empirical work exploring these issues is therefore greatly needed.

Beyond empirical work, we also need to understand the nature and sources of these potential costs and how they can be prevented. It is for this reason that I suggested various policy responses, in each case emphasizing that the costs are rooted in the way that corporations and governments are choosing to develop and use these technologies. Therefore, my conclusion is that the best way of preventing these costs is to regulate AI and redirect AI research, away from these harmful endeavors and towards areas where AI can create new tasks that increase

human productivity and new products and algorithms that can empower workers and citizens. Of course, I realize that such a redirection is unlikely, and regulation of AI is probably more difficult than the regulation of many other technologies — both because of its fast-changing, pervasive nature and because of the international dimension. We also have to be careful, since history is replete with instances in which governments and powerful interest groups opposed new technologies, with disastrous consequences for economic growth (see, e.g., Acemoglu and Robinson, 2012). Nevertheless, the importance of these potential harms justifies the need to start having such conversations.

# References


**Abramowitz, Alan I. (2010)** *The Disappearing Center: Engaged Citizens, Polarization, and American Democracy.* New Haven, CT: Yale University Press.

**Acemoglu, Daron (2002)** "Technical Change, Inequality, and The Labor Market," *Journal of Economic Literature*, 40(1): 7–72.

**Acemoglu, Daron (2019)** "It's Good Jobs Stupid," *Economics for Inclusive Prosperity*, Research Brief, June 2019. https://econfip.org/wp-content/uploads/2019/06/Its-Good-Jobs-Stupid.pdf

**Acemoglu, Daron (2021)** "AI's Future Doesn't Have to Be Dystopian," *Boston Review*, May 20, 2021. http://bostonreview.net/forum/science-nature/daron-acemoglu-redesigning-ai

**Acemoglu, Daron and David H. Autor (2011)** "Skills, Tasks and Technologies: Implications for Employment and Earnings," in Ashenfelter, Orley and David Card, eds., *Handbook of Labor Economics, Volume 4*: Amsterdam, Holland: El Sevier.

**Acemoglu, Daron, David H. Autor, Jonathon Hazell and Pascual Restrepo (2021)** "AI and Jobs: Evidence from Online Vacancies," NBER Working Paper No. 28257.

**Acemoglu, Daron and Simon Johnson (2022)** *Men of Genius: How Hubris Ruined Technology and Prosperity.* book manuscript.

**Acemoglu, Daron, Jon Kleinberg and Sendhil Mullainathan (2022)** "Problems of Online Communication," work in progress.

**Acemoglu, Daron, Ali Makhdoumi, Azarakhsh Malekian and Asu Ozdaglar (2021)** "Too Much Data: Prices and Inefficiencies in Data Markets," forthcoming *American Economic Journal: Micro.*

**Acemoglu, Daron, Ali Makhdoumi, Azarakhsh Malekian and Asu Ozdaglar (2022)** "A Model of Behavioral Manipulation," work in progress.

**Acemoglu, Daron, Andrea Manera and Pascual Restrepo (2020)** "Does the US Tax Code Favor Automation?" *Brookings Papers on Economic Activity*, 2020(1): 231–285.

**Acemoglu, Daron and Andrew F. Newman (2002)** "The Labor Market and Corporate Structure," *European Economic Review* 46(10): 1733-1756.

**Acemoglu, Daron, Asu Ozdaglar and James Siderius (2021)** "Misinformation: Strategic Sharing, Homophily and Endogenous Echo Chambers," NBER Working Paper No. 28884.

**Acemoglu, Daron and Jörn-Steffen Pischke (1998)** "The Structure of Wages and

Investment in Gen. Training," *Journal of Political Economy,* 107(3): 539-572.

**Acemoglu, Daron and James A. Robinson (2006)** "Why Did the West Extend the Franchise? Democracy, Inequality and Growth in Historical Perspective," *Quarterly Journal of Economics,* 115(4): 1167-1199.

**Acemoglu, Daron and James A. Robinson (2006)** *Economic Origins of Dictatorship and Democracy.* New York, NY: Cambridge University Press.

**Acemoglu, Daron and James A. Robinson (2008)** "Persistence of Power, Elites and Institutions," *American Economic Review*, 98(1): 267-93.

**Acemoglu, Daron and James A. Robinson (2012)** *Why Nations Fail: The Origins of Power, Prosperity, and Poverty.* New York, NY: Crown Business.

**Acemoglu, Daron and James A. Robinson (2019)** *The Narrow Corridor: States, Societies, and the Fate of Liberty.* New York, NY: Penguin Press.

**Acemoglu, Daron and Pascual Restrepo (2018)** "The Race Between Man and Machine: Implications of Technology for Growth, Factor Shares and Employment," *American Economic Review*, 108(6): 1488–1542.

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

**Acemoglu, Daron and Pascual Restrepo (2021)** "Tasks, Automation and the Rise in US Wage Inequality," NBER Working Paper No. 28920.

**Agarwal, Ajay, Joshua S. Gans and Avi Goldfarb (2018)** *Prediction Machines: The Simple Economics of Artificial Intelligence.* Cambridge, MA: Harvard Business Review.

**Allcott, Hunt and Matthew Gentzkow (2017)** "Social Media and Fake News in the 2016 Election," *Journal of Economic Perspectives*, 31: 211–36.

**Applebaum, Ann (2020)** *Twilight of Democracy: The Seductive Lure of Authoritarianism.* New York, NY: Signal.

**Athey, Susan, C. Catalini, and Catherine Tucker (2017)** "The Digital Privac Digital: Small Money, Small Costs, Small Talk," NBER Working Paper No. 23488.

**Autor, David H. (2014)** "Skills, Education and the Rise of Earnings Inequality among the Other 99 Percent," *Science* 344(6186): 843-851.

**Autor, David H., Frank Levy and Richard J. Murnane (2003)** "The Skill Content of Recent Technological Change: An Empirical Exploration," *Quarterly Journal of Economics,* 118(4): 1279–1333.

**Autor, David H., David Dorn, and Gordon H. Hanson (2013)** "The China Syn-

drome: Local Labor Market Effects of Import Competition in the United States," *American Economic Review* 103(6): 2121–68

**Bergemann, Dirk, Alessandro Bonatti, and Tan Gan (2021)** "Markets for Information," Yale mimeo.

**Bostrom, Nick (2017)** *Superintelligence.* New York, NY: Danod.

**Brynjolfsson, Erik and Andrew McAfee (2014)** *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies.* New York, NY: W. W. Norton & Company.

**Budd, Christopher, Christopher Harris, and John Vickers (1993)** "A Model of the Evolution of Duopoly: Does the Asymmetry between Firms Tend to Increase or Decrease?" *Review of Economic Studies* 60(3): 543-573.

**Choi, J.P., D.-S. Jeon, and B.-C. Kim (2019)** "Privacy and Personal Data Collection with Information Externalities," *Journal of Public Economics*, 173: 113–124.

**Christian, Brian (2019)** *The Alignment Problem: Machine Learning and Human Values.* New York, NY: WW Norton & Company.

**Dreyfus, Hubert L. (1992)** *What Computers Still Can't Do: A Critique of Artificial Reason.* Cambridge, MA: MIT Press.

**Farboodi, Maryam, R. Mihet, Thomas Philippon, and Laura Veldkamp (2019)** "Big Data and Firm Dynamics," *American Economic Review: Papers and Proceedings*,109, 38–42.

**Ford, Martin (2015)** *Rise of Robots.* New York, NY: Basic Books.

**Forester, Tom (1985)** *The Information Technology Revolution.* Cambridge, MA: MIT Press.

**Guriev, Sergei, Emeric Henry, and Ekaterina Zhuravskaya (2020)** "Checking and Sharing Alt-Facts." mimeo.

**Hanson Jon D. and Douglas A. Kysar (1999)** "Taking Behavioralism Seriously: Some Evidence of Market Manipulation," *New York University Law Review*, 74: 630.

**Huntington, Samuel P. (1991)** *The Third Wave: Democratization in the Late Twentieth Century.* Norman, OK: University of Oklahoma Press.

**Jones, Charles I. and Christopher Tonetti (2020)** "Non-rivalry in the Economics of Data," *American Economic Review,* 110(9): 2819-58.

**Judis, John B. (2016)** *The Populist Explosion: How the Great Recession Transformed American and European Politics.* New York, NY: Columbia Global Reports.

**Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018)** "Human Decisions and Machine Predictions," *Quarterly Journal of Economics*, 133(1): 237-293.

**Klemperer, Paul (1995)** "Competition When Consumers Have Switching Costs: An Overview with Applications to Industrial Organization, Macroeconomics and International Trade," *Review of Economic Studies*, 62(4): 515-539.

**Lanier, Jaron (2018)** *Ten Arguments for Deleting Your Social Media Accounts Right Now.* New York, NY: Hoffmann and Co.

**Lee, David S. (2009)** "Wage Inequality in the United States During the 1980s: Rising This Version or Falling Minimum Wage?" *Quarterly Journal of Economics,* 114(3): 977-1023.

**Lee, Kai-Fu (2018)** *AI superpowers: China, Silicon Valley, and the New World Order.* New York, NY: Houghton Mifflin Harcourt.

**Levy, Roee (2021)** "Social Media, News Consumption, and Polarization: Evidence from a Field Experiment." *American Economic Review.*

**MacCarthy, M. (2016)** "New Directions and Privacy: Disclosure, Unfairness and Externalities" mimeo. https://ssrn.com/abstract=3093301.

**Marantz, Andrew (2020)** *Antisocial: Online Extremists, Techno-utopians and the Hijacking of the American Conversation.* Penguin, New York, NY

**Mishra, P. (2017)** *Age of Anger: A History of the Present.* Macmillan, New York, NY.

**Markoff, John (1996)** *Waves of Democracy: Social Movements and Political Change.* Thousands Oaks: Pine Forge Press.

**Mosleh, Mohsen, Cameron Martel, Dean Eckles, and David G. Rand (2021)** "Shared Partisanship Dramatically Increases the Social Tie Formation in a Twitter Field Experiment," *Proceedings of the National Academy of Sciences*, 118(7): 1-3.

**Neapolitan, Richard E. and Xia Jiang (2018)** *Artificial Intelligence: With an Introduction to Machine Learning, 2nd ed.* London, UK: Chapman and Hall/CRC.

**Nilsson, Nils J. (2009)** *The Quest for Artificial Intelligence: A History of Ideas and Achievements.* New York, NY: Cambridge University Press.

**Pasquale, Frank (2015)** *The Black Box Society: The Secret Algorithms that Control Money and Information.* Cambridge, MA: Harvard University Press.

**Posner, Eric A. (2002)** *Law and Social Norms.* Cambridge, MA: Harvard University Press.

**Posner, Eric A. and E. Glen Weyl (2019)** *Radical Markets.* Princeton, NJ: Princeton

University Press.

**Russell, Stuart J. and Peter Norvig (2009)** *Artifical Intelligence: A Modern Approach, 3rd ed.* Hoboken, NJ: Prentice Hall.

**Russell, Stuart J. (2019)** *Human Compatible: Artificial Intelligence and the Problem of Control.* New York, NY: Penguin Press.

**Sandel, Michael J. (2020)** *The Tyranny of Merit: What's Become of the Common Good?* New York, NY: Penguin Press.

**Shapiro, Carl and Joseph E. Stiglitz (1984)** "Equilibrium Unemployment As Worker Discipline Device" *American Economic Review* 74(3): 433-444.

**Simonite Tom (2020)** Algorithms Were Supposed to Fix the Bail System. They Haven't." *Wired*. https://www.wired.com/story/algorithms-supposed-fix-bail-system-they-havent/

**Snyder, Timothy (2017)** *On Tyranny: Twenty Lessons from the Twentieth Century.* Toronton, Ontario: Tim Duggan Books.

**Sunstein, Cass (2001)** *Republic.com*, Princeton, NJ: Princeton University Press.

**Tirole, Jean (1989)** *Industrial Organization.* Cambridge MA: MIT Press.

**Tirole, Jean (1921)** "Digital Dystopia," *American Economic Review*, 111(6): 2007-48.

**Thompson, Derek (2019)** "Should We Be Afraid of AI in the Criminal-Justice System?" *The Atlantic.* https://www.theatlantic.com/ideas/archive/2019/06/should-we-be-afraid-of-ai-in-the-criminal-justice-system/592084/

**Varian, Hal R. (2009)** "Economic Aspects of Personal Privacy," *Internet Policy and Economics*, 101–109. Springer.

**Vosoughi, Soroush, Deb Roy, and Sinan Aral (2018)** "The Spread of True and False News Online," *Science*, 359: 1146–1151.

**West, Darrell M. (2018)** *The Future of Work: Robots, AI and Automation.* , Washington, DC: Brookings Institution Press.

**Zuboff, Shoshana (2019)** *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. London, UK: Profile Books.