NBER WORKING PAPER SERIES

INTEGRATING ETHICAL VALUES AND ECONOMIC VALUE TO STEER PROGRESS IN ARTIFICIAL INTELLIGENCE

Anton Korinek

Working Paper 26130
http://www.nber.org/papers/w26130

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2019

This article is an expanded version of a chapter commissioned by the Oxford Handbook of Ethics of Artificial Intelligence, edited by Markus D. Dubber, Frank Pasquale, and Sunit Das, forthcoming, Oxford University Press, 2019. I am grateful for the many thoughtful comments and insightful discussions with Avital Balwit, John Basl, Karen Delio, Kinda Hachem, Daniel Harper, Paul Humphreys, Eric Leeper, Joseph Stiglitz and Andy Wicks as well as participants of the "Human and Machine Intelligence Group" at UVA and of the workshop "Toward a Handbook of Ethics of AI" at the University of Toronto. Any remaining errors are my own. Financial support from the Institute for New Economic Thinking (INET) is gratefully acknowledged. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

Integrating Ethical Values and Economic Value to Steer Progress in Artificial Intelligence
Anton Korinek
NBER Working Paper No. 26130
August 2019
JEL No. E25,J23,J38,O33

ABSTRACT

Economics and ethics both offer important perspectives on our society, but they do so from two different viewpoints – the central focus of economics is how the price system in our economy values resources; the central focus of ethics is the moral evaluation of actions in our society. The rise of Artificial Intelligence (AI) forces humanity to confront new areas in which ethical values and economic value conflict, raising the question of what direction of technological progress is ultimately desirable for society. One crucial area are the effects of AI and related forms of automation on labor markets, which may lead to substantial increases in inequality unless mitigating policy actions are taken or progress is actively steered in a direction that complements human labor. Additional areas of conflict arise when AI systems optimize narrow market value but disregard broader ethical values and thus impose externalities on society, for example when AI systems engage in bias and discrimination, hack the human brain, and increasingly reduce human autonomy. Market incentives to create ever more intelligent systems lead to the ultimate ethical question: whether we should aim to create AI systems that surpass humans in general intelligence, and how to ensure that humanity is not left behind.

Anton Korinek
Department of Economics
University of Virginia
Monroe Hall 246
248 McCormick Rd
Charlottesville, VA 22904
and NBER
anton@korinek.com

1) Introduction

As we enter the Age of Artificial Intelligence, there is perhaps no single question more important than what direction future progress in AI will take. Artificial intelligence has the potential to alter the way our economy and society are structured in even more fundamental ways than earlier general purpose technologies such as the steam engine or electricity since it aims to automate what is the defining characteristic that sets humanity apart from other species – our intelligence. Progress in artificial intelligence thus offers abundant opportunities to improve the human condition, but it will also pose significant challenges for our society, and these are likely to grow in coming decades, as AI systems may replace humans in a growing number of areas.

However, the main message of this article is that technological progress does not just happen but is driven (at least for now) by human decisions on what, where, and how to innovate. It would be misplaced to succumb to techno-fatalism and view our fate as pre-determined by blind technological forces and market forces that are beyond our control. Instead, our future is shaped jointly by the technological innovations that we humans create, by the social and economic institutions that we collectively design, and by the ethical values that guide it all. We as a society have the power to confront the challenges posed by our technological possibilities and, through individual and collective action, actively steer the path of technological progress in AI so as to shape the future that we want to live in. This article is an attempt to discuss how to meet these challenges by integrating an assessment of the economic value created by AI with the complementary perspective offered by our ethical values.

The following section starts with a tangible example where simplistic economic and ethical views conflict: the hotly debated question of job losses induced by automation, including AI. Then I will examine the broader question of how market value and ethical values differ, why the values imposed by the market frequently prevail in such conflicts, and how society can take corrective actions. Section 3 discusses the inequality dimension of technological progress and underlines that pushing technological progress that is blind to its effects on inequality misses an important ethical perspective. Section 4 analyzes a number of areas in which AI systems that are programmed to maximize economic value violate our ethical values, for example by engaging in bias and discrimination, by hacking and manipulating the human brain, or by curtailing the scope for autonomous human decision-making. In the final section 5, I speculate how economic forces driving us towards AI systems with super-human levels of intelligence in coming decades may conflict with fundamental ethical values, because this may expose humanity to existential risks.

2) Economics and Ethics – Two Conflicting Value Systems?

Economics and ethics both offer important perspectives on our society, but they do so from two different viewpoints – the central focus of economics is how the price system in our economy values resources; the central focus of ethics is the moral evaluation of actions in our society.

Economic value and ethical values may at times look contradictory but are in fact complementary, as argued forcefully e.g. by Amartya Sen (1987). In a market economy, the system of market prices reflects how economic actors – humans in their roles as consumers, producers, workers, employers etc. – value economic resources. Market prices play a central role in guiding economic decisions – including in steering technological progress. Market prices offer some hints on what the individual members of society value. However, they are by no means a full representation of our values, missing out for

example on anything that is not traded in the market, including externalities. Market prices thus need to be complemented by ethical values to guide decisions so as to make them desirable for society. Since the ethical values of different individuals differ, I will not argue from one specific set of ethical values in this article, but I will instead draw only on those ethical values on which a vast majority of members of our society agree.

2.1) An Introductory Example: Job Losses from Automation and AI

Let me start with a tangible question that the advent of artificial intelligence – like many other forms of automation – raises, and on which economics and ethics are frequently viewed as providing contradictory answers:

Question: Is it right to introduce new technologies that lead to job losses?

In posing a charged question and offering answers from an ethical and economic perspective, I run the risk of offending both ethicists and economists, but I am comforted by the fact that the vast majority of both ethicists and economists that I have met care a lot about the betterment of our society. Integrating the two perspectives offers the greatest chance of moving forward the debate and arriving at acceptable answers, even if my own answers are necessarily tentative and partial.

Arguing from a narrow efficient markets perspective that does not include other dimensions of human well-being, economists may be tempted to immediately respond yes to my question. They may observe that in a well-functioning market, wages perfectly reflect the social value of labor; if at the given level of wages, a company finds it desirable to innovate so as to save on costly labor, it frees up labor to be employed in other activities that are more useful to society.

Conversely, observing the misery created by job losses, ethicists may be tempted to immediately respond no to my question. They can see the tangible harm and suffering imposed on workers who are laid off and observe that it is unethical to impose these on workers, whereas they may not immediately appreciate the longer-term effects of economic progress on human well-being.

After further deliberation, economists may appreciate that there are many other considerations that matter aside from the narrow efficient markets perspective argued above. First, markets are not complete in the real world. Workers cannot fully insure against unemployment (for example because full insurance would lower incentives to work with full effort). As a result, job losses are socially more costly than what an efficient markets view suggests. This is exacerbated by the fact that jobs are social arrangements that not only entail the exchange of labor against wages, but they also provide (or in technical language, are bundled with) other valuable experiences such as social connections, structure, personal meaning, status and a sense of belonging that the worker loses upon losing a job and that cannot be separately purchased on the market. As a result, losing a job is among the most traumatic events that people can experience during peace-time. The associated losses go far beyond what is captured by the purely economic loss of income. People who lose their jobs also experience a loss of meaning, become socially more isolated, and frequently become depressed. All this also affects their families and their communities, imposing externalities on them. When markets are so incomplete, it may be socially undesirable for markets to be the sole guides of human decision-making.

Second, the majority of economists also care about questions of income distribution. Even if markets generate resource allocations that are efficient, the distribution of incomes matters, and market

outcomes may generate a more unequal distribution that society perceives as less desirable. For example, job losses driven by automation frequently reduce labor demand and the incomes of workers and may increase the incomes of entrepreneurs and shareholders.

Third, and most fundamentally, there is no theoretical reason to believe that the free market will direct innovative efforts to the most socially desirable innovations. The first fundamental welfare theorem in economics, commonly referred to as the "invisible hand theorem," states that under certain idealized conditions, the market will generate an efficient distribution of the existing resources in the economy. However, this theorem does not apply to technological progress, and the market may thus guide innovation in the wrong direction.

Similarly, after some deliberation, ethicists may appreciate that markets do provide price signals that reflect scarcity and thus the societal value of resources, up to a point. These price signals aggregate the decisions of every single person participating in economic transactions and thus reflect many aspects of the ethical values of society. For example, if a sufficient number of consumers demand eggs from chickens that are raised in humane conditions, the market will provide such conditions to chickens. If a sufficient number of consumers demanded fast food provided by workers who earn a living wage, the market would provide fast food jobs paying a living wage. An important caveat is that prices can only reflect what consumers know about, and frequently unethical behavior can be hidden from the end consumer. Another caveat, already observed earlier, is that many things that matter cannot be traded in markets. However, although price signals omit a variety of multi-faceted ethical considerations, they still provide some useful information which, if guided correctly, can contribute to the common good.

Ethicists may also appreciate that their insights on the shortcomings and omissions of the market can sometimes best be corrected by imposing the right regulation on the market and letting the market – with proper ethical guidance – do its job. Looking specifically at our example of job losses, ethicists may appreciate that the pain that workers experience when they lose their job must be weighed against the long-term gains for society – in the long run, society overall may greatly benefit from deploying new technologies that displace some workers from existing jobs. Perhaps most fundamentally, many ethicist will appreciate that the alternatives to economic systems that assign an important role to markets are not very promising.

Taking into account the arguments from both perspectives, economists and ethicists may ultimately agree on a number of points: They may agree that it is desirable to ensure that workers who lose their jobs are cared for – not only in monetary terms (e.g. via unemployment insurance) but also in terms of the broader value that society assigns to their losses. After all, these workers were sacrificed for the sake of economic efficiency: when their jobs were displaced, they were the collateral damage to enable the economy to adopt more efficient production processes that will ultimately make society overall more prosperous, and so society owes them. Furthermore, they may concur that an unfettered market's decisions on what, where and how to innovate are not always in society's best interests. Likewise, the decisions by private enterprises on making workers redundant are not always in society's best interest. However, they may also agree that it is nonetheless important to carefully take into account the price signals provided by the market when making decisions about economic resources, since price signals do contain useful information.

2.2) Professional Biases

Even if economists and ethicists agree on the general points discussed, individual members of either profession may still reasonably disagree on the extent to which it is desirable for other institutions, including for government, to interfere in the described processes. This is a question of both political preferences and beliefs, for example beliefs in the effectiveness of such alternative institutions.

If we compare economists to non-economists, including ethicists, they probably tend to believe more strongly in the power of markets versus other institutions such as governments. Similarly, ethicists perhaps tend to believe more strongly in the relevance of careful ethical deliberation than non-ethicists, including economists. It is probably true of all scientific fields that people working in the field believe on average more strongly in the relevance of their subject of inquiry than people outside of the field. The reasons include both selection – people are more likely to specialize in a field that they believe is relevant – and cognitive biases that make researchers feel that what they know more about and what they have invested more time in is more important. However, the effectiveness of institutions is an empirical question, and both economists and ethicists can learn from evidence.

The gap is thus by no means unbridgeable. In fact, our society is most likely to benefit from the work of both economists and ethicists when they bridge their differences and integrate their insights. As the example above on job losses illustrates, an ethical perspective is useful for economists because it serves as a reminder that society values aspects of our human experience that are not appropriately captured and valued by the market. An economic perspective is useful for ethicists because it serves as a reminder that the market is a powerful force that shapes our world in significant ways – no matter if we want it to or not.

2.3) Conceptual Differences between Ethical Values and Market Value

Nonetheless, our systems of market prices and of ethical values differ in very significant conceptual ways:

Market prices are generally objective, single-dimensional and unambiguous. They put a well-defined dollar value on anything that is traded in the market. One of the reasons is that markets were created by humans specifically for the purpose of efficiently exchanging resources.

Each person's ethical values, by contrast, are subjective, multi-faceted and at times implicit, making them more ambiguous and difficult to compare. One of the reasons for this is that the ultimate arbiters of our ethical values are neural networks: our ethical values have been encoded in the deep neural networks that constitute our brains by the processes of nature and nurture, i.e. by biological and cultural evolution, and by our experiences and decisions that have shaped our lives. It is famously difficult to capture in general rules how complex deep neural networks arrive at decisions, yet in describing our ethical values we need to do precisely that – we need to describe in general rules how our brains decide what is ethical. Combining the ethical values of different individuals to guide decisions for society as a whole adds yet another layer of complexity.

Some of the differences between market value and ethical values thus boil down to the difference between systems that are purposefully designed versus systems created by evolution: the former type of systems are generally more efficient in a single-dimensional way – at accomplishing the specific purpose for which they were designed; the latter type of systems are generally more robust and adaptable – they are better at adjusting to changing environments and exhibit more "common sense."

2.4) Why Economic Value All Too Often Prevails Over Ethical Values

If we care about integrating ethical values in economic decisions, it is concerning that economic forces frequently seem to prevail over ethical values in today's world, and it is important to understand why. Without providing an exhaustive list, let me describe several factors that tilt the playing field towards economic value.

First, the conflict between market value and ethical values typically reflects the broader tradeoff between personal benefit versus societal benefits. Humans are pro-social, but only up to a point – our pro-social instincts have evolved mainly to benefit the small tribe of people around us, not humanity at large. For example, people who hesitate to pollute their neighbor's backyard frequently have fewer hesitations to contribute to global warming that hurts humanity as a whole – they apply lower ethical standards to externalities that affect larger groups and instead listen more to market signals. As a result, the trade-offs between personal and societal benefits that humans have evolved to make instinctively, may not be a good guide for ethical decisions that have broader societal repercussions. This is a significant problem in the context of new technologies that affect humanity as a whole.

Second, the subjectivity of ethical values leads to different views among different people: the person with the smallest conflict is most likely to perform an action that the market values, even though others may find it ethically questionable. Cynical economists may say that differences in ethical values create gains from trade based on comparative advantage in immorality. The result may be a race to the bottom in ethical values so that those with the fewest moral restraints in a given area will take up business opportunities that generate value in the market.

Moreover, market prices are so clear and visible in comparison to ethical values. Partly due to cognitive biases and partly due to ambiguity aversion, our brains favor single-dimensional and clear decision factors over multi-faceted and more complex decision factors. The clarity of the market system is then allowed to drive economic decisions towards utmost economic efficiency – but efficiency in a single-dimensional sense that ignores other ethical considerations.

2.5) Dealing with Discrepancies of Value

The vast majority of ethicists, of economists, and of society at large agree that the market should not win out when market values and ethical values conflict. Within economics, for example, an entire subfield called welfare economics describes policy tools that can be used to deal with situations when the market does not value things the same way as society. Economists frequently use the term *externalities* for discrepancies between social values and market value. Classic examples of such externalities include pollution or congestion, when the market does not correctly value the cost to society of limited resources like nature or road space. Examples of positive externalities include spillovers from technological progress, when the market does not correctly internalize that one person's ideas and inventions also benefit others who indirectly benefit from the ideas. If individuals behave in a purely self-interested fashion and do not account for the externalities that they create (as *homo oeconomicus* is postulated to do in most economic models), then there will be too many activities generating negative externalities and too few generating positive externalities. The problem would be resolved if individuals simply followed society's ethical values instead of the value assigned by the market.

However, welfare economics offers an alternative solution: economists have a rich and well-developed toolkit for how to regulate externalities. Such regulations can take a variety of forms: they can restrict the permissible quantity of harmful externalities or encourage a certain quantity of positive externalities; they can tax harmful externalities or subsidize positive externalities; and they can assign property rights (permits) on the creation of externalities and allow people to trade such permits in a new market, as for example in cap-and-trade schemes for pollution externalities. These solutions require the political choice of what ethical values to assign to externalities.

Generally speaking, the realm of such externalities – of issues in which there is a discrepancy between the value assigned by the market and our ethical values – is large. And even if we narrow our focus on the realm that is relevant for artificial intelligence, there is still a lot of ground to cover.

In the remainder of this article, I will attempt to analyze three categories of such discrepancies of value. First, I will focus on how markets are oblivious to questions of inequality and discuss why AI may lead to large increases in inequality. Next, I will discuss a range of other externalities generated by progress in AI that result from unidimensional market incentives conflicting with a multi-dimensional ethical perspective. Lastly, I will speculate on the conflict between market incentives and ethical values in the race towards superintelligence.

3) Progress in AI and Inequality

This section focuses on the effects of progress in AI on economic inequality – a question on which economics has many insights to offer. Several of the lessons of this section are more general and apply to any form of automation, but they are particularly relevant in the context of AI (see e.g. Acemoglu and Restrepo, 2019; Korinek and Stiglitz, 2019).

Technological progress is generally understood as a process that expands how much output the economy can produce for a given amount of inputs – it expands our production possibilities. Put this way, progress may sound almost uncontroversial – if we can produce more, technological progress carries the potential to make everybody in society better off from a material perspective. Arguing along those lines, it would seem almost unethical to oppose it!

However, there are two important caveats to our description of technological progress. First, technological progress *could* make everybody better off but is not guaranteed to do so. Secondly, *better off* refers to a strictly material perspective. These two caveats imply that technological progress frequently goes counter to the promise of improving everybody's livelihood.

3.1) From the Industrial Revolution to the Future

Looking at the broader context of technological progress since the Industrial Revolution serves as a reminder of how fundamental technological and economic forces have been in shaping the fate of mankind over the centuries. This underlines how important it is to actively steer future progress in AI with an ethical perspective in mind.

Prior to the Industrial Revolution that started in 18th century England, the vast majority of humanity lived at subsistence levels – in other words, most humans barely had enough material resources to survive and regularly went to sleep hungry. Like our fellow animals inhabiting planet Earth, humans were caught in a Malthusian trap: any time there was technological progress, it enabled population

growth, and the additional population ate up the additional output produced so that human living standards stubbornly remained at subsistence levels.

Over the centuries since the Industrial Revolution, by contrast, economic growth has outpaced population growth by so much that average material living standards for humans in advanced countries have increased by more than a factor of ten. No wonder that many contemporary economists believed, at least until recently, that it was a fundamental principle of technological progress that everybody benefitted, or that "a rising tide lifts all boats."

Focusing on the past four decades, however, the picture has been considerably more mixed: even though overall economic growth continued to pace ahead, the distribution of economic gains was more and more unequal. In the US, the bottom half of the population, consisting mainly of low-skilled workers, has barely experienced any income gains when adjusted for inflation. Large parts of the population, for example unskilled white Americans, have even experienced declines in life expectancy because of so-called "deaths of despair" from drug and alcohol abuse and suicides. Over the same period, the real incomes of the top 1% have doubled, those of the top 0.1% have tripled, and those of the top .01% have quadrupled.

These income statistics reflect economic forces that determine what our economic system values: the Industrial Revolution revolved around machines that replaced hard physical labor but badly needed human workers to operate them, and over time, these machines greatly increased the productivity and value of human labor. Market forces did their job, and the greater productivity of human labor was soon reflected in higher wages – across the board.

More recent waves of automation, by contrast, have almost banished humans from factory floors and from routine information processing tasks (see e.g. Acemoglu and Autor, 2011). This has made certain categories of human labor, especially unskilled labor, less and less useful to the economy.[2] Economists like to emphasize that lower demand for labor, e.g. for unskilled labor, translates primarily into lower wages, not unemployment. The reason is that labor supply is fairly inelastic, i.e. most humans continue to search for employment opportunities when technological forces displace them from their jobs, and the market responds to the resulting demand-supply imbalance by lowering wages.

By contrast, automation has greatly benefitted high-skilled workers, who have become more useful to our economic system and have, accordingly, experienced large increases in payoffs. New technologies have allowed high-skilled humans to generate vastly larger amounts of output by being the ones who oversee the more efficient production processes. The incomes of high-skilled workers have thus consistently outpaced those of low-skilled workers, as exemplified by an almost-doubling in the college wage premium, i.e. in the extra earnings that college graduates make compared to high school-only graduates. In short, value creation and payoffs in our economy have increasingly shifted from low-skilled workers to high-skilled workers and machines, exacerbating inequality.

[2] A full account of the increase in inequality in the US in recent decades also includes additional institutional forces. These involve changes in public policy and tax systems, trade with low-wage countries such as China, as well as declines in unionization and increases in the market power of corporations. Some of these forces are themselves likely driven in part by technological changes.

3.2) Technological Progress and Redistribution

An important take-away from our discussion of history is that technological progress frequently generates large redistributions of income across the economy. The main driving force behind this is that any technological change affects the prices of inputs and outputs to the production process (see e.g. Korinek and Stiglitz, 2019). The more significant an innovation is for the economy, the larger these redistributions usually are.

Consumers usually benefit from innovation, at least from a material perspective, since innovation leads to lower consumer prices, higher quality, and new products.

On the side of factor inputs to production (which include all the different forms of labor and capital that go into the process), things are much more ambiguous: there are no general economic laws as to whether specific factor inputs will benefit or be hurt by progress – it all depends on the specific nature of technological progress. When some of the factors inputs are hurt by an innovation, important ethical questions arise.

Labor is one of the key inputs to the production process, and the effects of technological progress on different types of labor may differ markedly: In our earlier example in which a new technology, say an AI system, leads to job losses, the wages of the affected workers typically fall as a result of the innovation; by contrast, those who have the skills to program and maintain the new system are likely to experience income gains. The owners of other production factors, such as capital and land, will experience changes in returns that could make them either better off or worse off. For example, if the new AI system requires less capital than what was used before, capitalists may be worse off; if the capital needs are larger, capitalists may be better off. In short, in a market economy, technological change generates not only winners, but regularly leads to significant redistributions.

The economic winners and losers of technological progress, no matter if they are workers or other factor owners, were never asked for their consent – they were "innocent bystanders" of technological progress and thus of the decisions and actions of individual innovators. Economists call it an externality when there are effects of economic actions on innocent bystanders. Since the effects occur via changes in prices and wages, they call them pecuniary externalities. We observed that the net effect of technological progress since the Industrial Revolution has probably been highly positive for humanity, implying that progress has led to large positive pecuniary externalities. However, the fates of workers with low and high skills have diverged in recent decades – so low-skilled workers have experienced stark negative pecuniary externalities over that period. In summary, even though the increase in output enabled by technological innovation makes it possible for everybody to be better off, the associated pecuniary externalities may in general create losers.

Redistribution and Utilitarianism

Some economist argue that we shouldn't care about pecuniary externalities since they constitute "mere" redistributions of income – they do not reduce overall income in the economy and could therefore, at least in principle, be undone by economic policy. Furthermore, they argue that economists can offer guidance on how to allocate resources efficiently, but what distribution of income is desirable and how society should respond to the redistributions generated by technological progress is outside of their subject area and is for the political process to decide. However, this perspective rightly opens

economists to the accusation of being biased. Except in two very specific cases, the question of how to efficiently allocate resources in the economy *cannot* be separated from redistributive questions.

The first case is if economic policy successfully manages to implement a desirable income distribution (for example by compensating the losers of progress) without introducing any distortions into the economy – economists call this idealized form of redistribution a lump sum transfer. However, this does not reflect the way that the economy works in practice – redistribution generally does create distortions of its own, as economists are quick to point out. Whenever we impose taxes on an economic resource or activity so as to raise funds for redistribution, we lower incentives to employ the resource or to engage in the activity, and we create incentives to circumvent the taxes. Furthermore, we also frequently distort the behavior of the potential recipients of such payments. The theoretical benchmark of lump sum transfers that do not distort does not exist in reality. Therefore the first case does not apply in practice.

The second case is if we only care about the overall level of income generated not the distribution of income. In line with the tradition within economics, let me call this ethical benchmark strict utilitarianism (although this does not do full justice to most varieties of utilitarianism discussed in ethics, for example what Bentham described as "the greatest amount of good for the greatest number" – economists' version of strict utilitarianism adds up the resources consumed by different individuals linearly and would find an innovation desirable even if it imposes income losses on all but one person in society so long as it increases the income of that remaining person by an amount slightly greater than the sum of the individual losses). Most people view this type of value system as at least borderline unethical – outside of economics, strict utilitarianism is a fringe perspective.

If economists strive to inform the choices facing society, then it is their job to employ society's values, not to impose their own. It is at best biased, and at worst manipulative, to evaluate social choices by imposing a value system such as strict utilitarianism that society does not share. If economists repeatedly offer advice based on a system of values that does not reflect societal values, society will become skeptical of economists. This would be a pity since economics does have many useful perspectives to offer on how to compensate the losers of technological progress.

Once we leave aside the two – unrealistic – special cases that we just spelled out, questions of efficiency and distribution cannot be separated, and it is indispensable to consider the distributional impact of technological innovations when evaluating their social desirability.

Ethical Perspectives on Economic Redistribution

Let me discuss two complementary ethical perspectives on how to think about the income redistributions generated by technological progress. For economists, the traditional way of looking at the distributive effects of technological progress is purely consequentialist, focusing solely on the desirability of the outcome in the form of the income distribution that is realized after progress has taken place. An alternative focus is on changes from the status quo, focusing on the desirability of innovation redistributing incomes across the economy and generating winners and losers.

In recent decades, much of the technological progress has hurt low-skilled workers and has increased inequality. If we care about equality, both the process of redistribution and the realized outcome seem undesirable in this example. However, the two perspectives differ and sometimes point in opposite directions: innovation may redistribute incomes, but it is a separate question whether this increases or decreases inequality. Progress in AI is about to hurt some high-skilled workers and may, in fact, reduce

inequality in some instances. For example, if radiologists are displaced by technology, inequality may in fact go down since radiologists were rather highly-paid. Advances in AI may make such examples more common in the future. In such cases, whether we view the effects of an innovation as desirable or not depends on whether our point of reference is the status quo or whether we are interested solely in the realized outcome.

If we focus solely on the realized outcome, then the fundamental ethical question is how much inequality we as a society wish to permit. Frequently, this choice is subject to the familiar trade-off between equality and efficiency. Most societies engage in some forms of income redistribution. Modest forms of redistribution do not interfere significantly with economic efficiency. However, as the extent of redistribution rises, the economic distortions that it generates rise more than proportionally and reduce economic efficiency. As a result, society needs to ask how much efficiency we are willing to give up to achieve greater equality. This is a political choice that our society has to make.

However, at times, there is not even a trade-off between equality and efficiency – it is possible to increase both equality and the total amount of wealth produced for our society. For example, if large corporations abuse their market power to extract monopoly rents, both equality and efficiency could be increased by regulating them or breaking them up. The same holds in all situations in which economic players extract rents from the economy (although, of course, the beneficiaries of the rents would lose).

Looking at changes from the status quo and at how innovation leads to redistribution, brings up an additional set of ethical questions. Should innovators in fact have the right to hurt others? Isn't it odd that we have criminal laws against theft, but that our society celebrates the entrepreneurs who take away the livelihoods of countless workers that they replace by automation? Put this way, many people tend to view the redistributive effects of innovation as unethical.

A more ethical perspective would be to evaluate the benefits and costs of an innovation by looking at all the members of society that are affected. We would then view an innovation as desirable if our ethical evaluation of the losses imposed on the losers of progress, after any compensation that they receive, is less than our ethical evaluation of the gains for the winners of technological progress. Critically, the ethical value assigned to a one dollar loss for a minimum wage worker is likely higher than the ethical value we assign to a one dollar gain for a billionaire. In practice, however, it is all too common that the losers of technological progress are left to fend for themselves or receive only minimal support from the winners or from society to make up for their losses.

3.3) Inequality and Steering Progress in AI

In short, some forms of technological progress may not desirable for society, even from a purely material perspective. Our society faces the choice of whether to let the free market or other decentralized forces determine which innovations take place, without regard for the common good, or whether we in fact want to steer the course of technological progress in a direction that helps workers. Although it is difficult to be certain of the overall economic effects of any given innovation, we generally have a sense whether an innovation will complement workers or substitute for them. We may well want to pass on innovations that increase output if a side effect is that a large number of people are actually worse off and if there is no realistic scope for compensating them. Conversely, society may want to actively work on innovations that do not strictly pass the market test but that offer large benefits to a large number of people.

Steering the course of progress could be done in a variety of ways:

Firstly, it would be possible to raise more awareness of the redistributive consequences of their work among ethically conscious entrepreneurs and researchers, and this could make a significant difference. Many entrepreneurs in the technology sector are quite public-spirited, exemplified by Google's former motto "don't be evil." However, in determining what is good or evil for society, what matters are not only the direct effects of new AI systems. The redistributive implications of new AI systems, especially the implications in labor markets, may also have significant effects on the welfare of individuals. As AI is affecting more and more sectors of the economy, entrepreneurs and researchers in the field of AI must be aware that their actions will increasingly shape the fate of workers and the overall income distribution across the economy. To provide a tangible example for how they could make a difference, I am currently participating in a research project to develop Intelligent Assistants (IAs) that aid unskilled human workers and enable them to do higher value tasks so as to enhance the market value of their labor. If creative entrepreneurs put their minds to it, I am sure that there are numerous examples of innovations that would both create jobs for unskilled workers and be economically profitable.

Secondly, governments have traditionally played an important role in shaping technological progress and could focus their efforts on promoting technologies that maintain or increase labor demand. A prime area for this is government-sponsored research, which could be guided more intentionally toward technologies that enhance the economic prospects of workers rather than replacing them. Furthermore, governments are large employers, both directly and indirectly via government procurement. By steering both their own automation decisions and those of their suppliers, they can have large effects on labor markets.

Thirdly, our society could steer progress in the private sector via taxes or subsidies that depend on whether an innovation replaces workers or enhances the role of workers. This would provide explicit incentives to innovators that reflect the likely labor market impact of an innovation. One caveat is that judging that impact is at times difficult to ascertain before an innovation is developed. A complementary approach is to directly target the market price of human labor. At present, our tax system inflates the cost of labor because labor is the most highly-taxed factor in the economy, providing extra incentives to develop technologies that save on labor. As we enter the Age of AI, our society would be better served by shifting the burden of taxation to other factors in the economy and provide subsidies to labor (for example by expanding programs such as the Earned Income Tax Credit). This would dis-incentivize investments into automating labor and steer progress in other directions.

So far our discussion has focused entirely on economic inequality. This is useful because we saw that there are significant ethical problems in this area, but inequality represents only one dimension of the many potential ethical dilemmas that innovation generates for our society.

4) Progress in AI Creating Externalities

This section moves beyond questions of income distribution and focuses on other dimensions in which market value does not adequately reflect the ethical values of our society. It is useful to distinguish two categories of such externalities arising from progress in AI: first, discrepancies in value that are newly introduced by AI; and secondly, existing market imperfections and externalities that are inherent in any economic system but that are exacerbated by the economic disruptions generated by AI.

4.1) Novel Ethical Problems and Externalities Introduced by AI

The rise of artificial intelligence opens up many new areas in which conflicts between market value and ethical values arise so that, along the way, new externalities are introduced. A number of the resulting ethical dilemmas are the subjects of individual chapters in the *Oxford Handbook on the Ethics of AI* for which this article was prepared (see Das et al, 2019).

A common theme in many of these dilemmas is that the technological innovations involved look like they create value in terms of economic profits, but they actually drain our broader societal values and do damage from an ethical perspective. In some instances, they are even doing more social harm than the private value that they create. A tangible example (from the days before AI) would be a factory that produces a valuable output but that pollutes so much that the social cost of pollution exceeds the market value of its output.

In the following, let me discuss a few specific examples of externalities generated by AI systems that are driven by economic efficiency at the expense of ethical values, and which can be addressed by integrating economic and ethical perspectives.

AI Discrimination, Biases and Fairness

Since AI algorithms make a growing number of decisions about our lives, one particularly concerning problem is that AI may either perpetuate biases or introduce new biases into how different people are treated. Consider for example an AI system that screens candidates for jobs, school admissions, or loans.

From a narrow economic perspective, the goal of an AI system that performs such screening is to identify the highest value candidates for businesses, schools, or lenders. Taking a typical data set to train the AI system, certain individual characteristics are correlated with higher value whereas others are correlated with lower value. An AI system identifies these correlations in far more intricate ways than the human brain and, in that sense, may be able to make more efficient screening decisions. Greater screening efficiency would translate into greater economic value. However, one of the ethical values of our society is that it is undesirable to discriminate against individuals based on personal characteristics that are outside of their control, in particular characteristics such as race, gender, or age.

Nonetheless, there are two scenarios in which AI systems may engage in precisely such discrimination. The first scenario is that the algorithm or the training data themselves are biased. This may be the case either because they are based on biased human decisions or because they are unrepresentative and thus generate less efficient decisions for underrepresented groups, which result in fewer positive screening decisions and bias. In this scenario, the bias is undesirable from both economic and ethical perspectives so the desirable path forward is clear.

The second scenario is that, even if the training data is fully representative and unbiased, many of us view it as unfair to base decisions solely on past observed correlations because doing so perpetuates the discrimination that has occurred in the past. Say, for example, that members of an ethnic group have historically defaulted on loans at higher rates; most of us would view it as morally wrong to charge members of that group a higher interest rate just because of their ethnicity. Even if AI systems are not explicitly fed data on protected individual characteristics such as ethnicity, they can still infer such characteristics from other data with a growing degree of accuracy and employ them in making decisions that look unbiased in the statistical sense and highly efficient from an economic perspective.

In the past, human decision makers that acted upon moral values of non-discrimination would attempt to evaluate candidates for jobs, school admissions, or loans impartially – by intentionally disregarding data that they knew are highly correlated with protected attributes, for example what they infer from looks, names, addresses, etc.[3] If we replace the human decision maker by an AI system that is focused solely on efficiency, then the AI system extracts greater economic value by lowering the ethical value it contributes to society.

AI systems can be programmed to explicitly follow principles of non-discrimination. However, in doing so they put a numerical value on non-discrimination practices and, whether explicitly or implicitly, highlight the discrepancy between the narrow economic cost and the broader ethical value of non-discrimination. For reasons discussed earlier, seeing the dollar value of discrimination may tempt decision makers to put greater weight on the clearly measurable economic dimension of a business decision compared to the ethical dimensions.

Hacking the Human Brain

Another example of hollowing out the human experience to earn extra profits is when AI systems are employed to hack the human brain. In computer science, hacking refers to situations when somebody intrudes into a system to either steal information or manipulate the behavior of the system. By AI algorithms hacking the human brain I refer to situations when algorithms tap into our simple human drives in order to manipulate us into behaviors or decisions that ultimately do not deliver the fulfillment that our drives were meant to deliver. The human brain constantly makes trade-offs between conflicting objectives, for example between primal instincts and rational thoughts. AI systems understand better and better how to tip the balance between the two, exploit our instincts, influence our thoughts, and manipulate us into whatever best achieves their objectives.

For example, AI-based advertising systems may manipulate us far more efficiently and in a much more personal way than traditional advertising to buy goods or services. Targeted links to sensational news stories tempt us into clicking and keeping reading, but may ultimately offer little informational value. Auto-play functions start the next video without asking after a user watches one video, and may keep us watching longer than we intended. Social networks promise to connect us in more efficient ways and automate many of our social interactions, keeping users engaged with constant friend updates. However, ultimately they may not generate the face-to-face human connection that is necessary to provide us with true fulfillment. The outcome in all these cases may be similar to a mild form of drug addiction in that our simple drives are exploited to the detriment of our long-term goals.

Conversely, AI systems could also hack our brain with the opposite objective in mind – to assist us in the pursuit of our long-term goals by regularly providing beneficial nudges, as for example fitness apps or dieting apps do, and to ultimately make us better off.

Curtailing Human Autonomy

The increasing use of AI to automate human decisions also runs the risk of reducing the human experience by curtailing our human autonomy. Many people assign significant value to human autonomy, i.e. to the ability to make independent decisions (in a direct contradiction to the

[3] There are also examples when human decision makers intentionally employed such characteristics: for example, banks engaged in “red-lining” whereby they denied credit to applicants based solely on their addresses being correlated with lower repayment rates. We as a society decided to ban such practices for ethical reasons.

consequentialist view that only the outcomes of decisions matter). For example, many car owners report that they value the ability to make their own decisions on how to drive, even if an autonomous vehicle could drive better along all objective dimensions. As AI systems in a given area get better and better, it becomes ever more tempting to impose the superior decisions of AI systems on human users, but doing so incurs the cost of reducing our autonomy.

Most humans will experience further limits to their autonomy as a result of increasing economic inequality. In Section 3, we discussed that AI systems that displace workers frequently increase income inequality. Over time, inequality in income leads to inequality in wealth and in the ownership of economic resources. Since ownership confers control, AI systems that earn an increasing fraction of the output of our economy will, over time, strip away control over resources from the workers who experience income losses and conversely, confer increasing levels of control over resources to their owners.[4]

The three discussed cases – bias, hacking the brain, reducing autonomy – are just a few examples in which technological innovations may generate economic value but diminish and hollow out other dimensions of our rich multi-faceted human experience.

4.2) Existing Market Failures Exacerbated by AI Disruption

In the real world, the market economy never quite works as efficiently as in the economist's textbook description. Furthermore, when the economy experiences significant disruptions, these frequently magnify existing tendencies towards inefficiency and lead to market allocations that exhibit far greater inefficiencies than in normal times.

One example of such inefficiencies are the missing markets for social connections, personal meaning, status and sense of belonging that we discussed in our introductory example on job losses. When the economy is near full employment and laid-off workers are quickly rehired, these market failures matter little; however, when large technological disruptions generate high unemployment and laid-off workers are unemployed for prolonged periods of time, the resulting welfare losses are significant. If, at some point in the future, progress in AI renders a growing fraction of the population unemployable, the laid-off workers will suffer tremendous losses in addition to the income that is lost. These losses represent externalities of technological innovation.

Furthermore, technological disruptions frequently generate significant aggregate demand effects. In a well-functioning market economy, aggregate demand is close to what the economy can produce. However, if many workers lose their jobs, aggregate demand declines. Government policy (including monetary policy) may not always be able to restore demand to an efficient level, leading to further job losses and wasted economic potential. By implication, significant technological disruptions may generate aggregate demand externalities and excessive recessions or even depressions. As an example from economic history, Delli Gatti et al. (2012) argue that the Great Depression of 1929 – 1933 was in part driven by a technological disruption: the mechanization of the agricultural sector and the resulting job losses.

[4] Whereas Bostrom (2014) articulated an AI control problem that was about the risk of humans losing control over super-intelligent AI systems, the described mechanism is a version of an economic AI control problem that may arise long before superintelligence.

If innovations generate significant societal disruptions that the invisible hand of the market is not good at managing because of market failures, then innovators have a moral duty to internalize the disruptive effects that they generate.

4.3) Externalities and Steering Progress in AI

Whenever market-provided price signals differ from broader ethical values, there is scope for integrating the two in order to steer technological progress. The two critical steps required are (i) to identify and understand the discrepancies in value (externalities) and (ii) to act upon that understanding.

When AI introduces novel externalities, such as in the examples given of bias & discrimination, hacking humans, and reducing human autonomy, the first step is to realize what is going on. The ideal course of action would be to anticipate potential ethical problems that are generated by new AI technologies and steer away from them. Some have suggested that innovators be required to conduct Technological Impact Assessments before making significant investments in new technologies, modeled on Environmental Impact Assessments, which attempt to evaluate the likely impact of innovations on society (see García and Janis, 2019). In practice, awareness of ethical problems frequently only arises after an innovation is introduced, and identifying novel ethical problems requires the collaboration of civil society, nonprofits, universities, governments and, above all, of course, the entrepreneurs or corporations who introduce the innovations in question. Once there is sufficient societal awareness, ethically-minded entrepreneurs may even leverage the potential positive externalities of progress, as in our example of fitness and dieting apps.

Given the tendency of the market to sponsor a race to the bottom when it comes to monetizing ethical transgressions, it may also be necessary to pass regulation to compel innovators to take into account their adverse effects on society. In practice, it is impossible to perfectly account for all externalities via regulation – and the political process as well as regulators typically lag behind the new externalities generated by novel technologies, leaving considerable moral responsibility with the innovators who introduce new technologies that potentially generate new externalities.

When existing market failures are exacerbated by progress in AI and disrupt the functioning of the economy, as in our example of aggregate demand problems, governments play an important role in both measuring the scope of the problem and in facilitating the adjustment process through macroeconomic and structural adjustment policies. However, their instruments are limited and they are typically not able to fully undo the negative effects. This leaves the innovators who roll out new technologies with considerable moral responsibility to take into account the broader effects of their actions. Large economic disruptions typically result from thousands of small steps to introduce new technologies such as AI throughout the different sectors of the economy. Each innovator who plays a role in one of the small steps should bear in mind their contribution to the disruption. If each of them makes an effort to steer their innovation in a direction that mitigates the disruptive impact, the overall positive effect will be quite significant.

5) The Race towards Superintelligence

Progress in artificial intelligence is continuing unabatedly, driven by the complementary forces of human curiosity and market incentives. Many of our brightest minds are working hard on improving the hardware and software required for AI, driving both exponential growth in computing power and continued advances in our ability to understand and write the software behind AI. Market incentives are

doing their part by generously rewarding the growing capabilities of existing AI systems and by pouring hundreds of billions of dollars into the development of new ones. In doing so, they have elevated the status of AI experts from geeks to rock stars.

The continuing exponential progress raises the question whether AI will, at some point, surpass human intelligence. Present-day AI systems exhibit *narrow* artificial intelligence – they have great (and frequently super-human) capabilities in narrowly-defined domains such as playing chess or Go, targeting ads, or reading x-rays. By contrast, humans possess general intelligence – the ability to act intelligently across a wide number of domains and integrate them all. This capacity enables us humans to employ the powers of AI in the service of our human goals. However, with each passing year, the capabilities of narrow AI systems are growing broader, and the advantage of narrow AI over humans in each specific domain is expanding. Unless progress in AI comes to a standstill, it seems to be largely a question of time when machines will reach human levels – and ultimately super-human levels – of general intelligence. Although this may sound like science fiction, Bostrom (2014) reports in a survey that several AI researchers predict that artificial general intelligence (AGI) will be achieved as early as next decade, and a majority of AI researchers expect AGI by the second half of the 21st century. There are also more skeptical voices who predict that AI will never be able to replicate human general intelligence and that the economic effects of AI will be far less significant than the general purpose technologies that were introduced in the 20th century (see e.g. Gordon, 2016). However, given the vast potential implications of AGI for mankind, it seems prudent to seriously think about the ramifications for our society, even if the advent of AGI is just one of several possible scenarios for the future.

What would artificial general intelligence and superintelligence imply for humanity if this scenario materializes? Our intelligence has been the defining characteristic that set humanity apart from other species of animals and that has allowed us to rule over Planet Earth, including over all the less intelligent co-inhabitants of our planet. Would superintelligent AI treat humanity the way that humans have treated other animals, domesticating and exploiting us when useful and terminating us when a nuisance? What other roles would there be left for humans? Or could we perhaps instill our goals and ethical values into super-intelligent machines so that they help us improve human well-being in ways that are presently unthinkable for modest human minds? These questions have been much discussed by philosophers of AI, and a full treatment is beyond the scope of this article. Let me refer to e.g. Bostrom (2014) for a comprehensive analysis.

What I want to focus on in the remainder of this article are two areas in which discrepancies between economic value and ethical values may play a significant role if the scenario of AGI materializes. I will attempt to shed light on these areas by integrating the perspectives from ethics and economics.

5.1) Superintelligence, Inequality, and the Economic Viability of Humans

One of the central dilemmas created by ever-more intelligent AI is that the agents that are morally relevant may become increasingly economically irrelevant, whereas the agents that are economically relevant may not be morally relevant.[5]

[5] Whether we should, will, and/or will have to attribute moral agency to artificially intelligent machines at some point is a far more difficult question that is beyond this article.

From an economic perspective, superintelligence might be the most productive and most profitable human invention ever. The market would greatly value the vast potential returns that human-level artificial intelligence or superintelligence could generate.

Human labor, by contrast, may become economically redundant in the event that superintelligence is achieved. Superintelligent machines may use their superior problem-solving capacity to figure out how to perform economically relevant tasks ever more efficiently. If, at some point, they could perform all formerly human tasks more cheaply than what it costs to keep humans alive (i.e. at a cost below human subsistence wages), then there would be no more economic justification to employ human labor, and humans would become technologically obsolete. In that scenario, human labor would be a redundant factor of production and a dominated technology – just as we no longer use oxen to plough fields because the cost of maintaining the oxen is not worth the economic value that they produce, it would no longer be economically worth it to pay humans what they need to survive. This superintelligence scenario would thus condemn the vast majority of humanity to technological unemployment.

If our decisions were solely guided by economic value, then it would be logical to phase out humanity once humans were to become economically redundant. The arc of our material progress would then have come full circle: before the Industrial Revolution, humanity started out in a Malthusian world in which our population numbers were held back by lack of material resources and starvation; after the advent of superintelligence, human labor would become redundant, and the fate of all but the wealthiest would end up being driven by Malthusian forces yet again, ultimately leading to starvation and declines in the human population. Whenever humans and machines compete over scarce resources in the economy, it would be economically more valuable to use them as inputs for machines rather than for humans in this scenario. Malthus's disciple Darwin would call the result of this competition over scarce resources between humans and machines survival of the fittest.

In economic terms, introducing an AI innovation that reduces human wages below subsistence levels would entail a particularly strong version of the pecuniary externalities that we discussed earlier in Section 3. However, given that the magnitude of these externalities would put the survival of most humans at risk, what would be at stake is not merely inequality but sheer human survival.

If, in this scenario, humans were no longer able to earn income from their labor, but it is ethically desirable to keep humanity alive (which I personally advocate in strong terms), then an alternative mechanism to provide for the material needs of those humans who have no other source of income is required. In principle, the vast potential growth that may be generated by superintelligence could make it comparatively easy to provide some resources to the technologically unemployed. The basic economic devices for how to conduct such redistribution have already been discussed in section 3, although the stakes would be far higher should superintelligence be developed.

5.2) Superintelligence, Externalities and Existential Risk

Although superintelligence carries enormous promise to improve the condition of mankind, it also poses unfathomable risks, which may not be correctly reflected in the economic incentives of its potential creators. Intelligence is commonly defined as the ability to accomplish complex goals (see e.g. Tegmark, 2017). A superintelligence is then almost by definition more effective at accomplishing its goals than humans. If its goals conflict with human goals, it is most likely that superintelligent AI will win over humans.

Given the likely complexity of superintelligent AI systems, conflicts with human goals may in fact arise quite easily – especially as unintended consequences. To make the existential risks inherent in superintelligent AI tangible, Bostrom (2014) offers a thought experiment of a system that is programmed to pursue a single narrow (and rather trivial) objective: to produce as many paperclips as possible. He argues that it could not be entirely ruled out that such a system, once superintelligent, may decide to kill off humanity in pursuit of its programmed objective, for example to use the iron in our bodies for paperclips, or to preempt the threat of being turned off, which would prevent it from maximizing its objective. Given that the system had not been programmed to pursue broader goals such as human wellbeing, it would simply not have cared about the demise of humanity. AI safety researchers have articulated dozens of additional scenarios in which superintelligent AI may endanger humanity.

More broadly, the incentives of AI researchers and of society as a whole may not be aligned when it comes to weighing the potential benefits of superintelligent AI against its existential risks. A researcher who has a tangible shot at creating and being in charge of the most powerful AI system ever built would have a huge potential upside in terms of scientific fame, power and material rewards. She may also be somewhat overconfident in her abilities to control such a system. But humanity as a whole would pay the price if things go as wrong as in Bostrom's example of existential risk. Given the asymmetry of who would obtain most of the benefits and who would bear most of the costs, the researcher may well be tempted to proceed and impose a small risk of existential catastrophe on humanity. And the sum total of risk exposure for humanity would keep rising if hundreds of research teams worked on advancing AI and each imposed a small existential risk on humanity. In short, individual incentives may not properly reflect the benefits and costs of incurring such existential risks.

The asymmetry of risks for individual AI researchers and for society as a whole may be exacerbated by competitive dynamics among multiple teams working on advanced AI. Given that there may be a first-mover advantage to whoever would first develop human-level artificial general intelligence, there might be strong incentives to disregard ethical values and program such systems at the highest speed possible without sufficient regard for the risks involved. Even more ominously, the race may also be sponsored by an additional set of powerful actors: the world's militaries who may see vast strategic advantage in achieving AI supremacy. Furthermore, there may be strong incentives for work on superintelligence to happen in secrecy, making it more difficult for broader society to weigh in with ethical concerns.

5.3) Superintelligence and Steering Progress in AI

Steering technological progress towards superintelligence may be the ultimate challenge for human society. However, although the stakes may be vastly higher, the challenges would be similar to the ones that we are currently facing with narrow AI – to ensure that AI systems carry out our economic interests while their behavior is guided by our ethical values, avoiding negative externalities.

Given the existential risks and the potential for economic irrelevance facing humanity in the event that superintelligence comes to fruition, we should not view progress towards superintelligent AI systems as primarily an economic project or primarily a research project – the ethical challenges and the stakes for humanity are too high to be determined by the commercial interests of any single corporation or by the research interests of any single research team. A large and concerted public effort to integrate the perspectives of all stakeholders of society could ensure that we develop AI in a direction that is both economically beneficial and ethically attractive.

References

Acemoglu, Daron and David H. Autor (2011), "Skills, Tasks and Technologies: Implications for Employment and Earnings," in Ashenfelter, Orley and David E. Card (eds.), *Handbook of Labor Economics* Vol. 4, pp. 1043-1171.

Acemoglu, Daron and Pascual Restrepo (2019), "The Wrong Kind of AI? Artificial Intelligence and the Future of Labor Demand," NBER Working Paper w25682.

Bostrom, Nick (2014), *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press.

Das, Sunit, Markus Dubber and Frank Pasquale (2019), *Oxford Handbook on the Ethics of AI*, Oxford University Press.

Delli Gatti, Domenico, Mauro Gallegati, Bruce C. Greenwald, Alberto Russo, and Joseph E. Stiglitz (2012), "Mobility constraints, productivity trends, and extended crises," *Journal of Economic Behavior & Organization* 83(3): 375-393.

García, José and Madeline Janis (2019), "How to keep the robots from taking jobs," *Politico*, 5/1/2019.

Gordon, Robert (2016), *The Rise and Fall of American Growth: The U.S. Standard of Living since the Civil War. Princeton*, NJ: Princeton University Press.

Korinek, Anton (2019), "The Rise of Artificially Intelligent Agents," Working Paper, University of Virginia.

Korinek, Anton and Joseph Stiglitz (2019), "Artificial Intelligence and Its Implications for Income Distribution and Unemployment," in Agrawal et al.: *The Economics of Artificial Intelligence*, NBER and University of Chicago Press.

Naidu, Suresh, Dani Rodrik and Gabriel Zucman (2019), *Economics for Inclusive Prosperity: An Introduction*, Economists for Inclusive Prosperity. http://www.econfip.org

Sen, Amartya (1987), *On Ethics and Economics*, Blackwell Publishing.

Tegmark, Max (2017), *Life 3.0: Being Human in the Age of Artificial Intelligence*, New York: Knopf.