The political economy of AI regulation: Towards democratic control of the means of prediction

Maximilian Kasy*

November 15, 2022

1 Introduction

This chapter discusses the regulation of artificial intelligence (AI) from the vantage point of political economy. By "political economy" I mean a perspective which emphasizes that there are different people and actors in society who have divergent interests and unequal access to resources and power. By "artificial intelligence" I mean the construction of autonomous agents that maximize some notion of reward. The construction of such agents typically draws on the tools of machine learning and optimization.

AI and machine learning are used in an ever wider array of socially consequential settings. This includes labor markets, education, criminal justice, health, banking, housing, as well as the curation of information by search engines, social networks, and recommender systems. The adoption of these technologies has potentially wide-ranging ramifications. This implies a need for public debates over desirable directions of technical innovation, the use of technologies, and constraints to be imposed on technologies.

In this chapter, I will review some frameworks to help structure such debates. The discussion in this chapter is opinionated and based on the following premises:

1. AI concerns the construction of agents which maximize a measurable objective (reward). Such agents take data as an input, and produce chosen actions as an output.

2. Maximization of a singular objective by autonomous agents is taking place in a social world where different individuals have divergent objectives. These divergent objectives might stand in conflict. Evaluated in terms of these divergent objectives,

*This chapter has benefitted from feedback by and discussions with numerous people, including Rediet Abebe, Daron Acemoglu, David Autor, Carlos Gonzalez Perez, Lukas Lehner, and the participants of the Oxford Machine Learning and Economics reading group and the MD4SG inequality group.

the actions and policies chosen by AI agents (almost) always generate winners and losers.

3. Going from individual-level assessments of gains and losses to society-level assessments requires aggregation, which trades off gains and losses across individuals. In order to normatively evaluate AI, as well as proposed regulations, we need to explicitly assess the resulting individual gains and losses, and explicitly aggregate these gains and losses across individuals.

4. The social issues raised by AI, including questions of fairness, privacy, value alignment, accountability, and automation, can only be resolved through democratic control of algorithm objectives, and of the means to obtain them - data and computational infrastructure. Democratic control requires public debate and binding collective decision-making, at many different levels of society.

My discussion will draw on concepts and references from machine learning, economics, and social choice theory. I will touch on several debates regarding the ethics and social impact of artificial intelligence, without any pretension of doing justice to the vast and growing literature on these topics; instead my goal is to give an internally coherent and principled account.

The remainder of this chapter is divided into two parts. Section 2 reviews background material. I first provide a brief overview of some subfields of AI, including supervised learning and deep neural nets, targeted treatment assignment, multiarmed bandits, and reinforcement learning. I then discuss a normative framework based on the evaluation of individual welfare and its aggregation via social welfare functions, as well as debates about agents of social change and democratic control. Section 3 then draws on this background material to consider several issues that have received attention regarding the ethics and social impact of AI, including (i) fairness, discrimination, and inequality, (ii) privacy, data property rights, and data governance, (iii) value alignment and the impending robot apocalypse, (iv) explainability and accountability for automated decision-making, and (v) automation and the impact of AI on the labor market and on wage inequality.

2 Background

2.1 AI and machine learning

AI in general, and machine learning in particular, are large and fast-growing fields, where new approaches are gaining attention continuously. Many of the underlying principles, which build on the fields of optimization and statistics, have however remained remarkably constant over time. A grasp of these principles is essential for effective regulation of AI. In the following, I give a brief high-level overview of some of the principles and approaches in machine learning.[1] Throughout, numbered equations are used to show different objective functions that are considered in different branches of AI, and in economics.

AI as automated decision-making One of the leading textbooks on AI (Russell and Norvig 2016, chapter 2) defines AI as the construction of rational agents. Agents receive information from their environment through perceptors (sensors), and act on their environment through actuators. The agent program maps sequences of percepts into actions. A rational agent is then defined as follows.

> For each possible percept sequence, a rational agent should select an action that is expected to maximize its performance measure, given the evidence provided by the percept sequence and whatever built-in knowledge the agent has.

This general definition covers a wide range of different approaches to the construction of rational agents. Of central importance for the present discussion is the performance measure or "reward" – what is it that the AI is trying to maximize? I will argue that democratic debate and decision-making needs to focus on the choice of such performance measures.

The leading approach to AI is based on machine learning, which will be the focus of the remainder of my discussion. This differs from alternative approaches that dominated in the past. One such alternative approach is the construction of "expert systems" which perform automated logical reasoning based on knowledge databases that were *hand-coded* by human experts. Machine learning, by contrast, leverages *statistical inference*, draws on (large) data-sets, and automates knowledge accumulation as much as possible.

Supervised learning Supervised learning is concerned with the prediction of outcomes Y given observable features X, for new data points (Shalev-Shwartz and Ben-David, 2014). This paradigm includes regression, classification, and more generally prediction,

[1] The review in this section draws on my lecture slides on the foundations of machine learning, available at `https://maxkasy.github.io/home/ML_Oxford_2022/`.

and has a wide range of applications, such as image recognition (predicting labels Y given images X), voice recognition (predicting text Y given sound X), automatic translation (predicting text in one language Y given text in another language X), evaluation of positions in a game (predicting the probability of winning Y given a position X), etc.

Applications of supervised learning in directly socially consequential settings include the evaluation of job applicants (predicting performance measures Y given employee characteristics X), bail setting for defendants in court (predicting future crime Y given defendant history X), evaluation of credit applicants (predicting repayment probability Y given applicant characteristics X), and evaluation of students applying to a university (predicting future test performance Y given their resume X). Supervised learning is also used by search engines and social media (predicting clicks Y for a given ad or entry and user with characteristics X), and recommender systems (predicting ratings Y for a given movie and user with viewing history X).

Supervised learning methods are evaluated based on their average out-of-sample prediction errors. It helps to introduce some notation. Let $g(X)$ be the prediction, which depends on the features X, where the function g has to be learned based on available observations (X_i, Y_i). The performance measure for supervised learning given by the average prediction loss,

$$E[l(g(X), Y)], \tag{1}$$

where l is the loss-function used to evaluate prediction errors, and $E[\cdot]$ denotes the average (expectation) over new draws ("out of sample") of (X, Y). A typical loss function for classification problems is given by the classification error loss, for which $l = 0$ if and only if $g(X) = Y$, that is, if the prediction is correct, and $l = 1$ otherwise. For prediction of continuous outcomes, a typical loss function is the squared error, $(g(X) - Y)^2$. The goal of supervised learning is to find ways of choosing g based on available observations to minimize the average prediction loss; this is called "training" or "estimation."

The choice of a prediction method typically involves a tradeoff between approximation errors ("bias") and estimation errors ("variance"). Methods which obtain small loss $l(g(X_i), Y_i)$ for all the available past observations i ("in sample") might still obtain a large average prediction loss for new data points ("out of sample"); this phenomenon is known as overfitting. Supervised learning methods need to limit overfitting. A common way of achieving this balance, trading off bias and variance, uses data splitting: The algorithm is run on one part of the data, and its average prediction loss is estimated on another part of the data. The algorithm is then fine-tuned to minimize the estimated average prediction loss.

Smaller prediction errors can be obtained with more observations i and richer predictive features X, combined with more flexible models g. Many of the advances of machine

learning in the 21st century can be traced to improvements on these dimensions, along with corresponding improvements in computer hardware.

Deep learning One particular approach to supervised learning is so-called deep learning (Goodfellow et al., 2016), which uses artificial neural networks. This is the approach that has achieved the most spectacular successes in recent years. In a nutshell, deep learning amounts to using a particular class of functions $g(X)$, which map input features X into predictions. In deep learning, these functions are built from a combination of a very large number of more primitive functions, which are interpreted as neurons that map their arguments into an output or "activation." "Training" these neural networks amounts to finding a function g that achieves small average prediction loss. Neural nets usually represent complicated non-linear functions g, which implies that their training (finding an optimal g) involves difficult optimization problems. A lot of work has gone into the development of hardware and algorithms to solve these optimization problems, in addition to fine-tuning specific "architectures" (functional forms for g) that perform well on specific prediction problems.

Deep learning is especially successful in settings with very large data-sets, in terms of both the richness of features X and the number of observations i. Trained neural nets might be thought of as databases that encode the examples (X_i, Y_i) on which they were trained; such neural networks then interpolate between past examples to form predictions for new observations. It should be noted that for smaller datasets with less rich features, deep learning is often less successful than more traditional and simpler prediction models. This applies for instance to the datasets often found in the social sciences and in applications in labor markets or education.

Targeted treatment assignment While prediction is a rich paradigm that has many applications, prediction is usually only an intermediate step in the context of automated decision-making. In settings such as the evaluation of job applicants, bail setting for defendants in court, evaluation of credit applicants, evaluation of students applying to a university, or curation of search feeds, the goal is typically to make a decision W based on the observed features X (Kitagawa and Tetenov, 2018; Athey and Wager, 2021) – hiring, incarceration, approving credit, admission to school, displaying a particular entry, etc.

Put differently, the problem is to find a treatment assignment policy $W = h(X)$ which maximizes some objective. For binary treatment $W \in \{0, 1\}$, a typical objective is

$$E[h(X) \cdot (Y - c)]. \tag{2}$$

The ideal treatment assignment for this objective treats everyone whose outcome Y ex-

ceeds a threshold c – hire everyone with performance above the threshold, give credit to all with repayment probability above the threshold, etc. As Y is not known for new applicants, this ideal treatment assignment cannot be implemented. Automated treatment assignment might instead use a supervised learning method $g(X)$ to predict Y given observable features X. Based on such a prediction, it is then optimal to treat everyone for whom the predicted outcome exceeds the threshold, that is, $h(X) = 1$ if and only if $g(X) \geq c$.

The attentive reader might have noted that the objective in Equation (2) is in fact a special case of the objective in Equation (1), where we can set $l(h(X), Y) = -h(X) \cdot (Y - c)$. What often complicates matters, however, is that Y_i might only be observed in the data whenever $W_i = 1$. A bank, for instance, only knows re-payment probabilities for those who actually obtained credit, but not for those who were denied. This partial observability potentially creates selection biases, a topic to which much attention has been devoted in the literature on causal inference in econometrics and related fields (Imbens and Rubin, 2015).

Multiarmed bandits The methods discussed thus far are sometimes called "offline" methods. This refers to the fact that learning happens for a given, fixed dataset of (X_i, Y_i). After learning from such a dataset, a fixed prediction rule $g(X)$ or treatment assignment rule $h(X)$ is implemented.

This contrasts with "online" settings, where decisions have to be made over time, and decision-rules are updated as new data arrive. The canonical example of an online decision-problem is the "multiarmed bandit" setting (Bubeck and Cesa-Bianchi, 2012). In the multiarmed bandit setting, units i arrive over time, and the algorithm has to decide on their treatment W_i. Then an outcome Y_i is realized and observed, where Y_i might be impacted by W_i. Drawing on the past history of treatments and outcomes, the algorithm then decides how to treat the next unit, and so on. For a given time horizon T, the goal is to maximize the average over time of the outcomes Y_i,

$$\frac{1}{T} \sum_{i=1}^{T} Y_i. \tag{3}$$

Typical settings where multiarmed bandits might be used include online advertisements (display the content or layout W that generates the most clicks Y), online price setting (set the price W that generates the largest profits Y), and clinical trials (assign the treatment W that maximizes patient survival chances Y).

The choice of treatment for a given unit serves two purposes in a multiarmed bandit setting. First, the purpose of achieving good outcomes now, and second, the purpose

of learning about the effectiveness of alternative treatments, in order to achieve better outcomes in the future. That is, there is a trade-off between "exploitation" – choosing good treatments now – and "exploration" – learning for future treatment choices. Good algorithms find an (approximately) optimal balance between these two goals, and thereby learn the optimal treatment quickly, achieving good average outcomes Y.

A generalization of the basic multiarmed bandit setting is the contextual bandit setting, which combines the bandit setting with the targeted treatment assignment problem. Contextual bandits learn optimal targeting rules $h(X)$ (e.g., what ads to show to which internet users, or what treatment to assign to which kind of patient), again by trading off exploration and exploitation. The features X are called the "context" for these treatment assignment decisions.

Reinforcement learning For contextual bandits, the features X of new observations do not depend on past treatment assignment decisions. In many settings, however, past actions affect current states. This is captured in the framework of Markov Decision Problems (Sutton and Barto, 2018), where the last period action might impact the current state (but does not have indirect effects into the future). Markov Decision Problems provide a general model of goal-oriented interaction with an environment, under complete observability. The objective in Markov Decision Problems is again to maximize the stream of rewards, as in Equation (3).

One example of Markov Decision Problems are games such as Chess and Go. In such games, most actions in most states do not generate a direct reward, but instead change the position X of the board in a way that might impact the future reward Y, that is, the probability of winning the game. Denote the expected future reward (or "value function") in state X for action W by $Q(X, W)$. We can decompose this expected reward as

$$Q(X_i, W_i) = E[Y_i + Q(X_{i+1}, W_{i+1})|X_i, W_i], \tag{4}$$

where the right hand side is a conditional expectation, that is, an average among all instances with a given value of the current state X_i and current action W_i. This decomposition is know as the Bellman equation. When choosing W_i, the algorithm has to consider both the effect on the current Y_i and the effect on future rewards $Q(X_{i+1}, W_{i+1})$, which is mediated by the effect of W_i on X_{i+1}. The optimal action W_i in state X_i is the value W which maximizes $Q(X_i, W)$.

Reinforcement learning is concerned with Markov decision problems where the effect of actions on both rewards and future states are unknown and have to be learned. A leading approach directly learns the expected future rewards Y for a given state X and action W, that is, the value function $Q(X, W)$. In a game such as Chess or Go, for instance, this

approach would directly predict the probability of winning given the state of the board X and the next move W. Estimating the value function amounts to a recursive form of a supervised learning problem, where we predict $Y_i + Q(X_{i+1}, W_{i+1})$ given X_i, W_i. The value function captures for instance the probability of winning for a given configuration of the Go board. Estimation of value functions might leverage deep learning. Successful application of this "deep reinforcement learning" approach has lead to breakthroughs in beating human masters at Chess and Go.

2.2 Social welfare

Having completed our brief review of machine learning, we next turn to a framework for the normative evaluation of (automated) decision-making systems more generally. This section loosely builds on the exposition of social choice theory and the theory of justice provided by Sen (1995) and Roemer (1998).[2] I will discuss a general framework for the evaluation of social welfare, which quantifies how "good" a given society is, and some of the key questions that arise in this framework. This framework is individualist in that it first evaluates the welfare of every individual under consideration, and then aggregates to evaluate social welfare. Such a framework is consistent with many different views about social justice; some exceptions will be mentioned below.

Social welfare might be affected by the adoption of AI-based technologies, and by policies regulating AI. A technology or policy is considered desirable if it increases social welfare. I introduce this framework here to provide a systematic way of evaluating the social impact and regulation of AI. This framework emphasizes that in society there is a multiplicity of conflicting objectives across individuals, which contrasts with the focus in machine learning on optimizing a singular objective of a single agent.

Individual welfare and social welfare We evaluate social welfare based on the welfare of a set of individuals $i = 1, \ldots, n$. This already raises the first set of difficult questions: Who is to be included in this set of individuals? Many discussions implicitly assume we are considering a given "society." But does that mean everybody of a certain citizenship, or everybody living in a certain territory? Why not all living human beings; should humans of another country count for nothing? And what about future generations? What about animals?

Given the set of individuals, we next need to decide how to measure their welfare. The goal is to assign a number V_i to each individual i, where V_i measures how well they are doing. Depending on the setting, a stronger or weaker interpretation might be given

[2]This section also draws on Chapter 2 of my open online textbook Kasy (2016), available at `http://inequalityresearch.net`, where a more detailed discussion of the issues considered here can be found.

to this number. An ordinal interpretation would only consider whether V_i is smaller or larger. A cardinal interpretation would care about the actual magnitude of V_i. An intermediate interpretation would consider the magnitude of changes of V_i. And the V_i might or might not be comparable across persons. If they are, then it makes sense to say that V_i is bigger than V_j, that is, i is doing better than j.

How to measure individual welfare V_i again raises a whole set of difficult questions. A minimal notion would only consider the formal legal rights enjoyed by individuals. A broader notion might also take into account various resources that allow individuals to achieve their objectives, such as education, income, and health. A comprehensive notion of opportunities might aim to take into account all factors that influence individuals' options, and evaluate the options effectively available to them. And we might finally consider the outcomes actually achieved by individuals, evaluated either by some common criteria, or by their individual preferences. Utilitarianism, the most common perspective in welfare economics, evaluates individual welfare by the outcomes actually achieved, as evaluated by individual preferences.

Given the set of individuals i, and given evaluations V_i of their welfare, we finally ask how well society as a whole is doing. Formally, we consider a "social welfare function,"

$$F(V_1, \ldots, V_n). \tag{5}$$

The function F determines how much we care about different individuals. Note that everybody still has a "name" at this point; the function F tells us how much weight we assign to the welfare of i relative to the welfare of j. The function might treat different people similarly, and not care about names. In that case it would still tell us how much we care about an additional dollar for a poor person versus a rich person, for a sick person versus a healthy person, since these would affect the levels V_i. And the function F might not depend on names, but on some characteristics of individuals. The function might for instance incorporate the belief that race, gender, or parental status should not determine individual welfare.

This individualist framework, where we evaluate social welfare F as a function of individuals' welfare V_i for $i = 1, \ldots, n$, is very general. It is compatible with various perspectives, whether radical or conservative ones. It does impose some restrictions, however, excluding in particular both fascist and libertarian normative approaches. Fascist approaches emphasize the greatness of a nation as their objective, no matter what the cost to the individuals involved. Relatedly, perfectionist perspectives take greatness in cultural production, science, etc. as its objective, again independently of the welfare of individuals. Libertarian approaches, on the other hand, consider outcomes to be just as long as they are the consequence of private property, contracts, and voluntary exchange

on markets, no matter what the consequences for individuals' welfare or inequality are. I will return to this contrast between libertarian and welfare-based approaches when we discuss algorithmic fairness in Section 3.1 below.

How to measure individual welfare We will discuss two alternative ways of measuring individual welfare. The first one equates welfare to "utility," and assumes that utility is what individuals maximize in their choices. The second one measures welfare in terms of the resources ("primary goods") that individuals have in order to achieve their objectives, whatever these might be, or in terms of their "capabilities," which take into account all constraints that individuals might face in pursuing their objectives.

Evaluating individual welfare in terms of utility is by far the most common approach in economics. There are two main ingredients to this approach. First, a choice set C_i containing elements c. This choice set describes all options among which an individual can effectively choose, taking all constraints that she faces into account. Second, individual preferences, expressed in terms of a utility function $u_i(c)$. The assumption is that individuals choose the element $c \in C_i$ which yields the greatest utility. The corresponding level of utility is taken to be a measure of individual welfare, so that

$$V_i = \max_{c \in C_i} u_i(c). \tag{6}$$

Note that utility plays two roles in this framework. Utility is (i) what individuals maximize when they have a choice, and (ii) what is deemed desirable for them from a social welfare point of view. It is generally assumed that preferences as encapsulated in the utility functions u_i are given independently of policy or technology, while choice sets C_i might be affected by policy or technology. Different policies can be compared in terms of the welfare V_i that individuals can achieve given these policies.

How can we give empirical content to the idea of utility? Utility is not observable, but choices are. In this framework, it is assumed that choices are maximizing utility. Leveraging this assumption, it is possible to define the concept of "equivalent variation." Equivalent variation measures utility in monetary terms. It corresponds to the change in unearned income that results in the same change of utility V_i as any given change of the choice set. To give a simple example, would you rather get 3 Euro, or two apples? If you are indifferent between the two, then the equivalent variation of getting two apples equals 3 Euro. More generally, the equivalent variation of income, for more interesting policy changes, can be inferred from observed choices; cf. Chapter 3 of Mas-Colell et al. (1995). This approach allows us to calculate welfare effects on individuals in monetary terms. Once we have done so, what remains to decide is how much we care about additional Euros (or Dollars, etc.) for different people, in order to make statements about social

welfare.

The utilitarian approach allows to evaluate what is better or worse for a given individual, based on what they would choose if they could. It implies, in particular, that having more options (i.e., a larger choice set C_i) is generally better. It does not, however, allow to compare utility across people, at least based on revealed preference. Being able to do such a comparison is key for talking about inequality from a normative point of view.

What we can observe and compare across people are the resources that they have at their disposition, and that affect the choices they could effectively make. More resources generally means more options and thus higher utility. Considering these resources is one of the key ideas in in Rawls (1973), who calls them primary goods. He argues that health, civil and political rights, income and wealth, and social bases of self-respect should be among the primary goods by which individual welfare is to be measured. This approach allows to say who is doing better or worse in terms of resources, and thus whether equality is achieved or not.

A similar notion, following up on Rawls, was developed by Sen (1995). He argues that the key measure of welfare are "capabilities to function." Rather than being based on a fixed list of resources or primary goods, capabilities are comprehensive measures of the options effectively at the disposition of individuals, taking into account all legal, social, economic, and cultural constraints. These capabilities are quite similar to the sets C_i introduced above in our discussion of utilitarian welfare.

How to trade off welfare across people Whichever way individual welfare is measured, in order to make normative statements we must aggregate individual welfare to social welfare, using some function F which takes care of this aggregation.

Suppose we are interested in small changes of some policy, which results in corresponding small changes dV_i of welfare V_i for each individual. It is then useful to think about aggregation in terms of welfare weights ω_i: Recall that social welfare is equal to $SWF = F(V_1, \ldots, V_n)$. Assume that F is differentiable and let $\omega_i := \frac{\partial}{\partial V_i} F(V_1, \ldots, V_n)$. We can calculate the change dF of social welfare by taking derivatives,

$$dF = \sum_i \omega_i \cdot dV_i. \tag{7}$$

where dV_i is the effect of the policy change on the welfare of individual i, and ω_i measures the weight attached to the welfare of individual i when calculating social welfare, cf. Saez and Stantcheva (2016).

When welfare is evaluated in terms of individual utilities, dV_i is often expressed in monetary terms (as an amount of Euro, for instance). As discussed above, the variable dV_i is then called equivalent variation. If dV_i is measured in monetary terms, then the

ratio ω_i/ω_j measures how much we care about an additional Euro for person i relative to an additional Euro for person j. An egalitarian position assigns a large weight ω to those who are worst off in a society. More generally, we can think of the ratio between welfare weights for poor versus rich people as measuring how egalitarian some aggregation is. Utilitarianism, as originally conceived, assumes that the marginal utility of income is declining with income, and thus puts a higher weight on those worse off. In contrast to such an aggregation, some models used in the field of law and economics are just summing up Euros across people, thus assigning the same weight to additional income for everyone (Liscow, 2018). An additional Euro for a billionaire is considered equivalent to an additional Euro for a starving person. Such an aggregation is very anti-egalitarian.

We can also think about welfare weights "in reverse:" Given the policy choices actually made in a society, there is a corresponding set of welfare weights which justifies these policy choices. Such weights can be thought of as measuring effective social power – whose interests are represented by the powers that be.

2.3 Agents of change and democratic governance

In the previous section we discussed the normative evaluation of technologies and policies in a society of individuals with multiple conflicting objectives. Such normative evaluations need to be complemented by a strategic vision of how to affect social change in order to be consequential. But who are the potential agents of social change? The question of agents of change has been much discussed in political theory over the last century (e.g. Wright, 2019, chapter 6). Some versions of this question are: Which agents have the interests, the values, and the capacity, to move technology and policy? Which agent can move technology and policy in a direction that improves social welfare, according to some notion of justice, for instance by improving the wellbeing of those who are worse off? Lastly, in a fragmented world, where no single actor or organization is strong enough to affect change, how can alliances and coalitions be built around hegemonic actors, where actors in the coalition share interests and values and develop a shared sense of identity (Gramsci, 2010)?

Closely related to this instrumental or strategic perspective, regarding the power to affect change, is the normative question of democracy and control over technologies; see for instance Wright (2019), chapter 4 on economic democracy. Control over the objective functions of AI needs to be broadly shared in order to align the objectives of AI with social welfare, in particular if agents act in a self-interested manner.

Profit maximization, agents of change, and threat points Who are potential agents of change who can direct the development, deployment, and regulation of AI in a

socially beneficial direction?[3] Much of the discourse on the social impact of AI focuses on voluntary ethical behavior by companies. The implicit assumption is that corporate managers or engineers will act as agents of change toward a more equitable society.

This approach contrasts with a presumption shared by mainstream economics, left-wing political theory, and neoliberal thinking: Private companies are, in the first place, profit maximizing entities. If this presumption is correct, then the focus on the ethics of corporate decisionmakers necessarily circumscribes the scope of possible change quite narrowly. Voluntary corporate action toward socially beneficial regulation of AI requires consistency of these actions with profit maximization. If private companies are primarily profit maximizing, then it follows that ethical considerations either remain cosmetic, or play only an instrumental role. Such an instrumental role could reflect the (elusive) "business case for diversity," or actions to avert antidiscrimination lawsuits, union organizing, bad press, consumer boycotts, or government regulation.

This points us to potential avenues for change in a world where technological decisions are made in private corporations: External pressures need to be such that profit-maximization aligns better with social welfare. In order for such pressures to play a meaningful role in corporate calculations, we need external regulation, advocacy, and oversight. That is, we need actors outside these corporations who are aware of the problems new technologies might be causing, who understand how they impact all members of society, and who can influence norms, change incentives, and take direct action to alter the course of development. There are many potential actors, and many forms this action might take, falling in both the categories of "exit" and "voice," as discussed by Hirschman (1986). There are organizations and unions of workers who have the potential leverage of strikes. There are civil society actors, nongovernmental organizations, and journalists who have the potential leverage of public attention and consumer boycotts. And there are government policymakers, the judiciary, regulatory agencies, and politicians who have the leverage of taxation, legislation and litigation. All these actors have an essential role to play in any strategy leading to a more socially beneficial future for AI.

Democratic governance and the means of prediction The argument just made is a strategic argument for addressing and empowering actors whose interests and values align with a broad base of society, so that the objectives of AI are directed to align with social welfare. Actors who are bound by an objective of profit maximization (or other narrow interests) are structurally incapable to make choices about technology that broadly improve welfare.

This strategic or instrumental argument is related to a more directly normative ar-

[3]The following draws on an argument previously made in Abebe and Kasy (2021), `https://www.bostonreview.net/forum_response/the-means-of-prediction/`.

gument: Self-determination – the ability to determine our own fate – is a goal in itself, and self-determination needs to go beyond the individual level in a society of complex interdependencies, in order to be effective. Put differently, democracy on various levels of society is a goal that we should work towards. For arguably historically contingent reasons, modern societies tend to separate the sphere of politics from the sphere of the economy. Democracy in the sphere of politics is a goal broadly supported across the political spectrum. In the sphere of economics, on the other hand, and especially in the English speaking part of the world, plutocracy is broadly accepted. In the part of our waking time that we spend in employment relationships, the decision-making power of those who own the capital remains largely unquestioned. It has been argued (e.g. Wright, 2019, chapter 4) that the project of democratization thus remains unfinished, and that a good society needs to expand democracy beyond the sphere of politics, narrowly construed, and into other spheres that are consequential for our lives.

In terms of AI, what this perspective implies is that we need democratic debate and control over the objectives of automated decision-making systems. Such control should be exerted in particular by those who are affected in important ways by the decisions of such systems. It should be exercised not only at the level of nation states, in the context of regular elections, but at various levels of society. Democratic control of the objectives for automated decisionmaking, as discussed in Section 2.1, in turn requires democratic control over the "means of prediction," including data and computational infrastructure, as well as a basic understanding of the workings of automated decision systems. Appealing to the ethical impulses of agents constrained by profit maximization is not enough. Neither are minimal regulatory constraints geared only toward preventing the worst harms. And neither are individual-level property rights over data, a proposal that we will discuss below. The remainder of this chapter elaborates on this perspective in the context of various debates about the ethics and social impact of AI.

3 The ethics and social impact of AI

Having concluded our review of conceptual frameworks in machine learning and social choice theory, we now turn to debates and concerns about the ethics and social impact of AI. These debates concern (i) fairness, discrimination, and inequality, (ii) privacy, data property rights, and data governance, (iii) value alignment and the impending robot apocalypse, (iv) explainability and accountability for automated decision-making, and (v) automation and the impact of AI on the labor market and on wage inequality. Rather than providing a comprehensive review of these debates, I will provide a brief account of each of them in terms of the frameworks set up in Section 2. I will then connect each of these debates back to our discussion of objectives for AI, social welfare, and democratic governance.

3.1 Fairness and inequality

One of the concerns raised about automated decisionmaking is whether the resulting decisions are fair, or reversely, whether a decisionmaking system is discriminatory.[4] The canonical setting for discussions about algorithmic fairness is the setting of targeted treatment assignment, as discussed in Section 2.1. To recall, in this setting the algorithm produces a treatment assignment policy $W = h(X)$, assigning treatment W as a function of observable features X. A typical objective for the case of binary treatments is to maximize the sum of outcomes Y net of the cost of treatment c among the treated, $E[h(X) \cdot (Y - c)]$. In the context of hiring, for instance, W is the decision to hire a worker, Y is some measure of their productivity (contribution to firm profits), and c is their wage.

Fairness What does it mean for a policy $h(X)$ to be fair, or discriminatory? Many different definitions have been proposed in the literature, and some of these are mathematically inconsistent. Many of these definitions are, however, conceptually closely related, and are similar to notions of taste-based discrimination in economics. In particular, typical definitions are based on some notion of merit Y, which corresponds to the decisionmaker's objective, such as profits, and some salient identity groups A, such as race or gender. An assignment policy is designated as unfair if there are inequalities in the probability of being treated across groups, where these inequalities are not justified by differences in "merit" between the groups. One such definition of fairness is "predictive parity," which requires that the difference

$$E[Y|W = 1, A = 1] - E[Y|W = 1, A = 0]$$

[4] The discussion in this section draws on Kasy and Abebe (2021).

equals 0. In words, average merit among the treated should be the same across the groups $A = 1$ and $A = 0$. This condition is also known as the "hit rate test for taste-based discrimination" in empirical economics, cf. Knowles et al. (2001). Other definitions which are closely related, and which share the same limitations, are "equality of true positives," "equality of false positives," and "balance for the negative class." Also related are so-called "individual fairness" notions; see for instance Pessach and Shmueli (2020) for a comprehensive review of fairness definitions.

As argued in Kasy and Abebe (2021), definitions of this form have a number of normative limitations:

1. They legitimize and perpetuate inequalities justified by "merit" both within and between groups. The focus on "merit" – a measure promoting the decision-maker's objective – reinforces, rather than questions, the legitimacy of the status quo.

2. They are narrowly-bracketed. Fairness only requires equal treatment within the context of the algorithm at hand, and does not consider the impact of the algorithm on inequality of welfare in the wider population. Unequal treatment that compensates pre-existing inequalities might reduce inequality of welfare.

3. They focus on categories (protected groups) and ignore within-group inequalities as emphasized in particular by intersectional critiques (Crenshaw, 1990). Equal treatment across groups can be consistent with great inequality within groups.

Closely related to these limitations is the fact that definitions of fairness such as predictive parity essentially equate discrimination to a deviation from profit maximization: When outcomes Y are perfectly predictable, then fairness holds automatically for the profit-maximizing policy $h(X)$, no matter how unequal the resulting allocation is in terms of welfare, both between and within the groups A. When Y is binary and perfectly predictable given X, for instance, then $W = h(X) = Y$ is profit maximizing and satisfies predictive parity, as well as balance for the negative class, equality of true and false positives, etc.

This is no coincidence. At the origin of definitions of taste-based discrimination lies the work of Becker (1957). This work equated normative desirability to the decisions of profit maximizing firms in competitive markets – independently of consequences for the welfare of those affected by such decisions. Such notions of fairness and discrimination are based on an implicitly libertarian normative framework. This contrasts with consequentialist notions of social welfare, as discussed in Section 2.2 above. Fairness, in Becker's sense, amounts to choosing a treatment assignment policy maximizing the profit objective $E[h(X) \cdot (Y - c)]$, rather than a policy maximizing some notion of social welfare $F(V_1, \ldots, V_n)$.

Equality and social welfare From the perspective of social welfare, we care about the impact of a policy $h(X)$ on the welfare V_i of individuals, and on the aggregate welfare $F(V_1, \ldots, V_n)$. To discuss this impact formally, consider the potential outcomes framework of causal inference (Imbens and Rubin, 2015). Consider binary treatments W_i and assume there are no spillovers across individuals, in the determination of their welfare V_i. Then we can assume that individual welfare is determined by the potential outcome equation

$$V_i = W_i \cdot V_i^1 + (1 - W_i) \cdot V_i^0. \tag{8}$$

In words, if individual i is treated, their welfare equals V_i^1, otherwise their welfare equals V_i^0. Any treatment assignment policy h which maps X into the probability of being treated $h(X)$ then implies a realized distribution of welfare, joint with features X, of

$$p_{V,X}(v, x) = \left[p_{V^0|X}(v, x) + h(x) \cdot \left(p_{V^1|X}(v, x) - p_{V^0|X}(v, x)\right)\right] \cdot p_X(x).$$

Typical social welfare functions F can then be written as functions of this distribution.

This is a generic description for evaluating the impact of a treatment assignment policy on social welfare, as in Section 2.2. Let us emphasize again how this differs from the fairness notions discussed above: Fairness is about *treating* people of the same "*merit*" independently of their *group* membership. Social welfare is about the (counterfactual / causal) *consequences* of an algorithm for the distribution of *welfare* of different *people*. There are many situations where these notions deliver conflicting normartive evaluations. One example concerns the consequences of increased surveillance, or better prediction algorithms. Typically, more accurate prediction will lead to treatments more aligned with "merit," which is good for fairness. They will also lead to more unequal treatment, which is often bad for equality and social welfare. Another example concerns all forms of affirmative action, redistribution, or compensatory interventions for pre-existing inequalities. Such interventions are always bad for fairness, as defined above, but good for equality and social welfare.

The main takeaway from this discussion for the regulation of artificial intelligence is that evaluations of automated decision systems should focus less on fairness, which typically requires alignment with profit maximization, and more on the consequences for individual and social welfare. Mechanisms for auditing and control by unions, civil society organizations, and public policymakers, need to draw on tools for the evaluation of consequences for welfare. These tools include causal inference and the measurement of welfare, as developed in empirical economics, among other fields.

3.2 Privacy and data governance

In order to learn, machine learning needs an objective function, a learning algorithm, and data. In the notation of Section 2.1 for supervised learning, the data often take the form of observations (X_i, Y_i), where i indexes different individuals. As machine learning has entered many socially consequential domains, privacy concerns become paramount: Who gets access and control to what kind of data about individuals? To what extent can an individual effectively veto such access? How can harms to individuals that result from data-collection be prevented?

The leading formalization of privacy in computer science is differential privacy. I next review this formalization and discuss its advantages and limitations. Differential privacy essentially guarantees that individuals are indifferent about collection of their data, no matter what happens later. Correspondingly, they do not have reason to object to collection of their data if differential privacy is guaranteed, even if they could do so based on individual data property rights.

I then discuss the limitations of this approach. While differential privacy effectively ensures that no harm results for *individual* i from collection of their *own data*, it does not ensure that no harm results for *groups* of people from collection of their data. The reason are "data externalities." Machine learning is about learning patterns, rather than individual instances. Differential privacy, and individual property rights, do not prevent the learning of such patterns, and thus do not prevent any resulting downstream harms. Correspondingly, harms from data collection can only be prevented by means of democratic data governance.

Differential privacy Consider a database $\mathcal{X}$, with entries of the form (X_i, Y_i). How can we reveal useful information about such a database $\mathcal{X}$ to some third party, without revealing the identity of those represented in the database? Differential privacy, which provides an answer to this question, is a well-studied and coherent approach to privacy in computer science (Dwork and Roth, 2014). This approach can be understood as a response to the limitations of earlier attempts at maintaining privacy while releasing some information. Naive approaches to privacy are typically vulnerable to auxiliary information. One such approach is the removal of "identifying information," such as names, addresses, social security numbers, etc. The problem with this approach is that even a small number of seemingly innocent variables are often sufficient to uniquely identify individuals. For example, it is likely that a user is uniquely identified by the list of the most recent movies they might have watched on a streaming platform. Another approach commonly employed is aggregation, where only averages of variables across groups of people are reported. Again, it is surprisingly easy to identify the individuals in such an

aggregate, once the averages for multiple variables are reported. A striking example are medical studies that report frequencies of genetic variants ("single nucleotide polymorphisms") among patients with a certain disease. From such aggregate frequencies, for a large enough number of genes, it is possible that the identity of all patients in the sample can be recovered, thus revealing the fact that they have a certain disease.

"Differential privacy" provides a coherent and robust definition that avoids these pitfalls. Differential privacy considers randomized algorithms $\mathcal{M}$ that map databases $\mathcal{X}$ into random reports $\mathcal{M}(\mathcal{X})$. A report here is any description or summary of the data $\mathcal{X}$; for instance a summary produced by a machine learning algorithm. Consider two databases $\mathcal{X}$ and $\mathcal{X}'$ which are the same, except that the data (X_i, Y_i) for one individual i have been added or removed between $\mathcal{X}$ and $\mathcal{X}'$. We say that a randomized algorithm $\mathcal{M}$ is ϵ-differentially private if for all sets of possible reports $\mathcal{S}$, and for all $\mathcal{X}$ and $\mathcal{X}'$ that only differ by one entry, we have that

$$\frac{P(\mathcal{M}(\mathcal{X}) \in \mathcal{S})}{P(\mathcal{M}(\mathcal{X}') \in \mathcal{S})} \leq \exp(\epsilon). \tag{9}$$

If ϵ is small, this means that the probability distribution of $\mathcal{M}(\mathcal{X})$ and of $\mathcal{M}(\mathcal{X}')$ is almost the same, for any pair of databases that differ by one entry. It makes no discernible difference whether any given individual is included in the data or not. Any differentially private algorithm needs to employ randomness. Without randomness, there is always a pair of databases $\mathcal{X}$ and $\mathcal{X}'$ which differ by only one entry such that $\mathcal{M}(\mathcal{X}) \neq \mathcal{M}(\mathcal{X}')$, except in trivial cases.

This definition or privacy has a number of desirable properties. Most importantly, it is immune to post-processing. No matter what additional data someone might possess, and what calculations they might apply to the reported $\mathcal{M}(\mathcal{X})$, the resulting report remains ϵ-differentially private. If multiple reports are made based on a database $\mathcal{X}$, say reports that are ϵ_1 and ϵ_2 differentially private, then the bound on the privacy loss adds up – the result is $\epsilon_1 + \epsilon_2$ differentially private. This raises the issue how a given "privacy budget" might be spent. This issue came up, for instance, in recent debates about the US census, where different social scientists wished to learn different facts about the US population, subject to a global privacy budget constraint.

A large literature has developed randomized algorithms that reveal relevant information while maintaining differential privacy. Of particular importance for the present chapter are algorithms for differentially private machine learning. Recall from Section 2.1 the objectives of different branches of machine learning: Predicting outcomes Y given features X (for supervised learning), choosing treatments W given features X to maximize some outcome Y (for targeted treatment assignment), or choosing treatments W adaptively to maximize the stream of outcomes Y (for multi-armed bandits). Crucially,

all of these objectives require learning some population relationships between the X, W, and Y; none of these objectives require the identification of data for specific individuals in a dataset. As it turns out, and quite remarkably, differential privacy does impose very little cost in terms of these objectives, cf. Dwork and Roth (2014), chapter 11. Put differently, differential privacy can be preservered without affecting machine learning or any of its downstream consequences.

Property rights, data externalities, and democratic data governance Differential privacy implies that it makes (almost) no observable difference whether any given individual is included in the data or not. From this it follows that it makes (almost) no difference to an individual who might be impacted by third party actions based on $\mathcal{M}(\mathcal{X})$, whether they are included in the data or not. This holds regardless of which actions a third party might take based on the report $\mathcal{M}(\mathcal{X})$; cf. Dwork and Roth (2014), chapter 10. Notably, this holds no matter who gets to see the queries, what other information they possess, or what actions they might take based on the queries.

This insight connects the notion of differential privacy to a closely related debate about individual property rights in data. One proposal to address issues of privacy is to give individuals the legal right of control over data that concern themselves, and an associated right of selling these data to third parties, or exchanging them against services; this idea has been promoted by Jaron Lanier, among others. Based on the argument above, individuals who hold property rights for their data have no incentive to not reveal their data to a differentially private mechanism.

It might appear plausible that a legal arrangement with individual property rights over data would protect individuals from harms arising from data collection and processing. This is not the case, however, as shown by the following argument. If differential privacy is guaranteed, then individual data property rights will not reduce the collection of data. Since machine learning is, furthermore, essentially unimpeded by differential privacy, it follows that any harms or benefits arising from the use of machine-learning based technologies are unaffected by differential privacy or individual property rights. Economists have described this fact as "data externalities," cf. Acemoglu et al. (2022). When an individual reveals their data, even to a differentially private mechanism, this can have consequences for other individuals, and vice versa. Such externalities might lead to variants of the "prisoner's dilemma:" No individual has an incentive to withhold their data, but collectively everyone may be harmed by data collection.

To provide a concrete example, consider a health insurance company collecting data on individual predictive features (risk factors) X for a certain disease Y, where the predictive features X are in the public domain, but the disease data are subject to privacy restrictions. If patients sell their data to the company, subject to differential privacy, they

gain individually, based on any payment they may receive. As a consequence, however, premia for all the high-risk patients go up, making them worse off, and the insurance arguably less equitable.

Such harms can, by construction, not be prevented through the enactment of individual-level privacy rights. Instead, as argued by Viljoen (2021), they can only be addressed through some form of collective democratic data governance. Only with collective processes of deliberation and decisionmaking regarding the collection of data, as discussed in Section 2.3, can we prevent harms from surveillance by private or public entities, while also reaping the potential benefits of AI and machine learning. Those who might be affected by automated decision-making, which is based on patterns learned from individual data, need to have a say in the collection of these data. An important task for the regulation of AI is to create legal environments and organizational structures that allow for such collective deliberation and control over the means of prediction.

3.3 Value alignment and conflicts of interest

As discussed in Section 2.1, AI is concerned with the construction of rational agents which maximize some stream of rewards. One approach covered by this general definition is Reinforcement learning, cf. Sutton and Barto (2018); François-Lavet et al. (2018). In Reinforcement learning, the algorithm has to solve a Markov Decision Problem, where there are endogenous states of the environment which evolve over time. The algorithm has to learn the probability of going from any given state to the next, and the distribution of rewards for any state and action, in order to maximize a stream of rewards. An issue that arises in this context, as well as in other branches of AI, is the issue of value alignment (Russell, 2019):

> "*[...] we may suffer from a failure of value alignment—we may, perhaps inadvertently, imbue machines with objectives that are imperfectly aligned with our own.*"

A famous thought-experiment illustrating this issue was formulated by Bostrom (2003), who considered a hypothetical AI which has been endowed with the goal to produce as many paper clips as possible. Since humans don't want infinitely many paper-clips, the AI will eventually need to eliminate the possibility of humans turning it off, by eliminating humans.[5]

While such scenarios might not be our most immediate concern, variants of this danger in more mundane settings are very much relevant at present. One example is the auto-

[5] Discussions of value alignment are often combined with predictions of a "singularity" in the development of AI. The idea of a singularity is that once AI achieves "human level intelligence," it will be able to improve itself, leading to an exponential explosion of AI capabilities. Such a presumed singularity is emphatically *not* part of my discussion here.

mated curation of news feeds and search results by social networks and search engines to maximize user engagement and, ultimately, ad clicks. The success of these algorithms has arguably contributed to filter bubbles and political polarization, the rise of click bait and conspiracy theories, and a rise of mental health problems among teenagers. Another example is the use of machine learning to match unemployed workers to jobs. Such algorithms are considered by employment agencies in order to maximize the rate of proposed matches which result in employment. This objective, combined with penalties for job seekers who do not accept proposed matches, might lead to systematically placing workers in jobs for which they are over-qualified, resulting in de-skilling, wage declines, and reductions in worker welfare.

Inverse reinforcement learning The value alignment problem arises when the objective function (reward) of an algorithm is not exactly the same as the objective function of humans. One possible solution to the value alignment problem could be to more carefully encode human objectives in the reward maximized by the algorithm. The problem with this approach is that it is very hard to fully anticipate and capture the complexity of relevant human objectives, for all possible circumstances, in a single numerical measure.

Inverse reinforcement learning has been proposed as an alternative to this "hardcoding" approach, cf. Ng and Russell (2000). The idea of inverse reinforcement learning is to observe human behavior, and to infer the reward function which would make the observed human behavior optimal. This inverts the problem of reinforcement learning, where rewards are observed and optimal behavior has to be learned. The inverse reinforcement learning approach builds on a long tradition in economics, aiming to infer preferences from observed behavior, cf. Train (2009). This approach is sometimes marketed as "provably safe AI."

Multi-tasking, surrogate outcomes, and fundamental limits An alternative perspective on value alignment is suggested by economic theory, in particular by the frameworks of mechanism design and contract theory: There are fundamental limits to automated decision-making when not all relevant outcomes are measurable. These limits cannot be overcome through better engineering choices.

The designer of an AI system faces similar issues to the designer of incentive pay systems (contracts) which are based on quantitative measures of performance, where the goal is to guide the efforts of human agents. Models of multitasking, as introduced by Holmstrom and Milgrom (1991), provide an explanation why high-powered incentives based on observable outcomes can gravely distort agent behavior, and are rarely observed in actual employment contracts: If agent effort has multiple dimensions, and some of these are unobservable, then incentives based on the observable dimensions will distort effort away

from the unobserved ones. Teacher incentives based on student test performance, and the resulting distortion of education towards "teaching to the test," is a much-discussed example. The literature on surrogate outcomes in bio-statistics (Athey et al., 2019) provides another perspective on the same problem. This literature considers settings where outcomes of interest, such as patient mortality, are not observable, for instance because they take too long to realize. This literature discusses conditions under which causal effects can still be inferred based on intermediate (surrogate) outcomes, for instance blood pressure or some other biomarkers.

What unites these frameworks is the emphasis on observability. Not all relevant consequences of the actions (which might be chosen by an algorithm) are available in the form of observable quantitative measures. Any consequence that is not measurable is then not taken into account by an optimizing algorithm. This problem becomes more severe, the *better* the algorithm is at optimizing the measured objectives. Crucially, approaches such as inverse reinforcement learning cannot provide a solution to this issue of observability. Inverse reinforcement learning might adjust the weights on measurable components of human objectives, but it cannot learn components that are not measurable. Just as it is often not a good idea to have high-powered monetary incentives for human agents, there might be many settings where we need to refrain from deploying high-powered algorithms which maximize measurable objectives.

Conflicts of interest and democratic control Our discussion of value alignment thus far focused on the difference between the objective maximized by an algorithm and the objective of "humans." As emphasized throughout the discussion in this chapter, however, different humans have very different objectives and conflicting interests between themselves. The alignment problem in many cases is not so much about about a difference in objectives between algorithms and humans, but rather about a difference in objectives between those who define the objective of an algorithm, and the rest of society.

Our motivating example illustrates these points: A company operating a social network, using algorithms to maximize ad-clicks and user engagement, might well be acting optimally for their profit objective. If ad-click maximization implies that democracy or teenage mental health are harmed, then that is unfortunate collateral damage in the quest for profits, from the perspective of the company, rather than reflecting a failure of alignment between the algorithm and the company operating it. Solving issues of alignment between algorithms operated by specific individuals or organizations, and society at large, requires democratic control of the means of prediction. Only public debate and collective decisionmaking about the objectives to be maximized can ensure that the pursuit of specific measurable objectives by automated decisionmaking systems is socially beneficial, and does not lead to broad harms.

3.4 Explainability and accountability

Automated decision-making raises questions regarding accountability: Who is responsible for decisions made automatically? Can these decisions be appealed? Can these decisions be explained? Can the *rules* leading to these decisions be appealed? Why was a particular decision made, or reversely, what would the decision have been for different input features X? In this context, Vredenburgh (2022) argued for a right to explanation, as a precondition for informed self-advocacy.

Explaining decisions For our purposes, it is useful to distinguish between explaining individual decisions, and explaining decision rules. Someone affected by an individual decisions might rightfully wish to appeal that decision. In particular, she might ask why that decision was made, and what she could have done differently to yield a different decision. This is especially salient in the context of targeted treatment assignment, as discussed in Section 2.1; consider for instance of the case of hiring, or of university admissions.

As the framework of causal inference teaches us, questions of "why" a decision was made are somewhat ill-defined, even if we have the full decision-rule and its derivation at our disposition. This point was elegantly expressed by Tolstoy (1869):

> *Why does an apple fall when it is ripe? Is it brought down by the force of gravity? Is it because its stalk withers? Because it is dried by the sun, because it grows too heavy, or because the boy standing under the tree wants to eat it?*

Questions of the form "what if," by contrast, can be well-posed. They form the core of the interventionist framework for causality in the empirical social sciences (Imbens and Rubin, 2015). In the context of targeted treatment assignment rules of the form $W = h(X)$, we might ask what the decision W would have been if X had taken on a different value. To the extent that X is the product of individual decisions, such knowledge allows individuals to adjust their behavior to achieve a desired treatment, such as being hired, or admitted to a university.

Notably, the answer to such "what if" questions is more readily available, in principle, for automated decision-making systems than it is for human decision-makers. It is easier to have a computer evaluate the function $h(X)$ for counterfactual values of X than it is to get human decision-makers to credibly reveal counterfactual decisions. Consider a hiring decision by a corporate manager, or a university admissions decision by an admissions officer. In these settings, we don't know how the hiring manager or admissions officer would have decided for counterfactual applications.

What complicates the explanation of decision-functions $h(X)$, in terms of "what if" queries, is the possibility of high-dimensional features X and complicated, non-linear mappings h. This complication is especially salient for algorithms based on deep neural

networks. This is an aspect discussed in the field of algorithm explainability; one proposal, for instance, is to report local linear approximations to decision functions, to make them more human-readable.

Explaining decision rules Going beyond individual instances where a decision rule was applied, we might wish to have public debate and democratic accountability for decision rules. This aspect of explainability is closely connected to the remainder of my discussion in this chapter. Explanation of rules is a necessary pre-condition for democratic control of rules. In order to ensure socially beneficial development and deployment of AI, we cannot leave decisions about development and deployment to a narrow class of experts and technocrats, such as corporate engineers. Rarely will the interests and viewpoints of these experts align with the interests and viewpoints of the wider public, and, in particular, of vulnerable groups affected by new technologies. A key role, then, for experts interested in advancing social welfare is to explain AI to the wider public.

Some might object that the explanation of algorithmic rules to the public is impossible, especially when they are represented by very complicated functions, and derived in an automated, blackbox manner, using complicated algorithms, as is the case for deep neural networks, for instance. Such an objection is misguided, however. AI and machine learning are surprisingly simple, in the ways that matter, and can be very much made open to public debate. As emphasized throughout this chapter, AI concerns the construction of algorithms maximizing some performance measure, based on data. What is central to explanations of rules for the purpose of democratic debate, is the definition of the objective (performance measure), the space of actions considered, and possibly the data used. Furthermore, how well a given algorithm performs in terms of its performance measure is also easily communicated. Relative to these, the specific implementation and mechanics of an optimizing algorithm are secondary.

3.5 Automation and the labor market

Technical change in general, and AI in particular, affect individual and social welfare, and inequality, by shifting demand in the labor market. This is one of the most consequential mechanisms by which technology impacts social welfare. This mechanism has received a lot of attention in labor economics; see e.g. Acemoglu and Autor (2011).

Production functions The effect of shifting technology on labor demand is typically modeled in terms of (aggregate) production functions, cf. Mas-Colell et al. (1995), Chapter 5. The assumption is that total output Y is determined by a relationship of the

form

$$Y = y(N_1, \ldots, N_J, K_1, \ldots, K_M, A). \tag{10}$$

Here N_j is the number of workers of type j employed in production, and K_m is the amount of capital good of type m used. The argument A indicates the dependence of the production function on technology. As technology progresses, A changes, and the function y shifts, typically upward. If aggregate production can be represented in this way, and if wages are determined by competitive labor markets,[6] then the wages of each type of worker are equal to their marginal productivity (contribution to profits), i.e.,

$$w_j = \partial y / \partial N_j.$$

The effect of technical change on wages in this production function framework is theoretically ambiguous. It is possible for the *slope* $\partial y/\partial N_j$ to go either up or down, as the *level* y increases, due to a shift in A.

One possible effect of technological change on the labor market is so-called "skill-biased technical change" (Goldin and Katz, 2009). The empirical literature on skill-biased technical change has focused on the US labor market, in particular. It has been argued that there is a secular trend, due to technological progress, which continually increases the demand for workers with higher levels of education, relative to the demand for workers with lower levels of education. This shift in relative demand was not met by a corresponding shift in relative supply. The growth in the share of workers with advanced degrees has stalled since the 1980s, in the US. The shift in relative demand, combined with roughly constant relative supply, so the argument goes, led to an increase in wage inequality between education groups.

This pattern seems not to hold since the turn of the 21st century, however. Correspondingly, there was a shift in the emphasis of the literature towards so-called "polarization" of the (US) labor market, cf. Autor and Dorn (2013). The polarization argument is that technological advances and automation, including through AI-based technologies, led to an erosion of demand for tasks performed by workers in the middle of the income distribution. Typical examples are clerical and office jobs, affected by the advent of PCs and the internet, or blue collar jobs in industry, affected by the increased use of robots. Tasks at either end of the income distribution, on the other hand, are harder to automate. The consequence of such polarization, if it takes place, is an erosion of the middle class.

Objectives, and control of the means of prediction An insight that immediately emerges from the production function framework is that the effect of technical change on

[6] There are many reasons why this might not be the case, including monopsony, efficiency wages, and search frictions.

economic inequality is not pre-ordained. This opens the door for explanations that relate technical change to social choices made by those in power, subject to the institutional and ideological constraints that they are facing. An example of such an argument is provided in Chapter 7 of Acemoglu and Johnson (2023): Consider a setting where union power and sectoral bargaining imply that firms cannot unilaterally cut wages or fire workers. As a consequence, the potential returns to automating jobs in order to raise profits are limited. Rather than pursuing automation, profits are then maximized by investing in technologies that increase worker marginal productivity, while providing the corresponding training to workers.

The production function framework is very useful for structuring debates about the impact of AI on economic inequality. It also has some limitations, however. First, the production function framework is very flexible. Because of this, it can explain many different possible developments of inequality, after the fact, in terms of technical change. It does not yield sharp testable predictions, however. Second, this framework applies to any technology; there is nothing specific to AI relative to other technologies, in terms of its impact on the production function. And third, the technological channel through which choices impact production functions is left open in this framework.

The frameworks I have emphasized in this chapter can provide additional and complementary structure to the production function framework, as it relates to AI. Artificial intelligence is the construction of automated decision-making systems that maximize some measurable objective. The choice of this objective is made by those who control the means of prediction, in particular data and computational infrastructure. How, then, can the deployment of technologies based on machine learning be steered in a direction that favors increasing the marginal productivity of workers, especially for those at the bottom of the wage distribution? Once again, the answer is democratic control. Workers need to have a say over (i) what kinds of data are collected within their company, and (ii) what objectives automated decision-making should maximize. To the extent that these decisions are also consequential for workers who are not currently working for a given company, these decisions are furthermore best made through democratic governance at a sectoral level, rather than at individual establishments. The regulation of AI in the workplace needs to create the conditions for such democratic decision-making, to ensure that AI does not contribute to a further explosion of inequality and dehumanization of workplaces.

References

Abebe, R. and Kasy, M. (2021). The means of prediction. *Boston Review*.

Acemoglu, D. and Autor, D. (2011). Skills, tasks and technologies: Implications for employment and earnings. *Handbook of labor economics*, 4:1043–1171.

Acemoglu, D. and Johnson, S. (2023). *Power and Progress: Our Thousand-Year Struggle Over Technology and Prosperity*. PublicAffairs.

Acemoglu, D., Makhdoumi, A., Malekian, A., and Ozdaglar, A. (2022). Too much data: Prices and inefficiencies in data markets. *American Economic Journal: Microeconomics*, forthcoming.

Athey, S., Chetty, R., Imbens, G. W., and Kang, H. (2019). The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely.

Athey, S. and Wager, S. (2021). Policy learning with observational data. *Econometrica*, 89(1):133–161.

Autor, D. H. and Dorn, D. (2013). The growth of low-skill service jobs and the polarization of the us labor market. *American Economic Review*, 103(5):1553–97.

Becker, G. S. (1957). *The economics of discrimination*.

Bostrom, N. (2003). Ethical issues in advanced artificial intelligence. *Science fiction and philosophy: from time travel to superintelligence*, pages 277–284.

Bubeck, S. and Cesa-Bianchi, N. (2012). Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122.

Crenshaw, K. (1990). Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review*, 43:1241.

Dwork, C. and Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407.

François-Lavet, V., Henderson, P., Islam, R., Bellemare, M. G., and Pineau, J. (2018). An introduction to deep reinforcement learning. *Foundations and Trends® in Machine Learning*, 11(3-4):219–354.

Goldin, C. D. and Katz, L. F. (2009). *The race between education and technology*. Harvard University Press.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT Press.

Gramsci, A. (2010). *Prison notebooks*. Columbia University Press.

Hirschman, A. O. (1986). Exit and voice: an expanding sphere of influence. *Rival views of market society and other recent essays*, pages 77–104.

Holmstrom, B. and Milgrom, P. (1991). Multitask principal–agent analyses: Incentive contracts, asset ownership, and job design. *The Journal of Law, Economics, and Organization*, 7(special_issue):24–52.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Kasy, M. (2016). Empirical research on economic inequality.

Kasy, M. and Abebe, R. (2021). Fairness, equality, and power in algorithmic decision making. *ACM Conference on Fairness, Accountability, and Transparency*.

Kitagawa, T. and Tetenov, A. (2018). Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616.

Knowles, J., Persico, N., and Todd, P. (2001). Racial bias in motor vehicle searches: Theory and evidence. *Journal of Political Economy*, 109(1):203–229.

Liscow, Z. (2018). Is efficiency biased? *The University of Chicago Law Review*, 85(7):1649–1718.

Mas-Colell, A., Whinston, M. D., and Green, J. R. (1995). *Microeconomic theory*. Oxford University Press.

Ng, A. Y. and Russell, S. (2000). Algorithms for inverse reinforcement learning. In *ICML*, volume 1, page 2.

Pessach, D. and Shmueli, E. (2020). Algorithmic fairness. *arXiv preprint arXiv:2001.09784*.

Rawls, J. (1973). *A theory of justice*. Harvard University Press, Cambridge.

Roemer, J. E. (1998). *Theories of distributive justice*. Harvard University Press, Cambridge.

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Penguin.

Russell, S. J. and Norvig, P. (2016). *Artificial intelligence: a modern approach*. Pearson Education Limited.

Saez, E. and Stantcheva, S. (2016). Generalized social welfare weights for optimal tax theory. *American Economic Review*, 106(1):24–45.

Sen, A. (1995). *Inequality reexamined*. Oxford University Press, Oxford.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Tolstoy, L. (1869). *War and peace*.

Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge University Press.

Viljoen, S. (2021). A relational theory of data governance. *Yale Law Journal*, 131:573.

Vredenburgh, K. (2022). The right to explanation. *Journal of Political Philosophy*, 30(2):209–229.

Wright, E. O. (2019). *How to be an anticapitalist in the twenty-first century*. Verso Books.