Article

Thomas Ferretti*

# An Institutionalist Approach to AI Ethics: Justifying the Priority of Government Regulation over Self-Regulation

https://doi.org/10.1515/mopp-2020-0056
Published online August 5, 2021

**Abstract:** This article explores the cooperation of government and the private sector to tackle the ethical dimension of artificial intelligence (AI). The argument draws on the institutionalist approach in philosophy and business ethics defending a 'division of moral labor' between governments and the private sector (Rawls 2001; Scheffler and Munoz-Dardé 2005). The goal and main contribution of this article is to explain how this approach can provide ethical guidelines to the AI industry and to highlight the limits of self-regulation. In what follows, I discuss three institutionalist claims. First, principles of AI ethics should be validated through legitimate democratic processes. Second, compliance with these principles should be secured in a stable way. Third, their implementation in practice should be as efficient as possible. If we accept these claims, there are good reasons to conclude that, in many cases, governments implementing hard regulation are in principle (if not yet in practice) the best instruments to secure an ethical development of AI systems. Where adequate regulation exists, firms should respect the law. But when regulation does not yet exist, helping governments build adequate regulation should be businesses' ethical priority, not self-regulation.

**Keywords:** artificial intelligence, business ethics, division of moral labor, political CSR, regulation

## 1 Introduction

Artificial intelligence (AI) has seen interesting developments in the last decades. To illustrate the relevance of these developments, think about Google's DeepMind

***Corresponding author: Thomas Ferretti**, Department of Philosophy Logic and Scientific Method, The London School of Economics and Political Science, London, UK, E-mail: t.ferretti@lse.ac.uk. https://orcid.org/0000-0003-4683-883X

Open Access. © 2021 Thomas Ferretti, published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.

defeating Lee Sedol, the best human player of Go, with their program AlphaGo in 2015. The latest version of the program, AlphaZero, is remarkable in that it was designed to learn how to play Go entirely by itself, with only the rules of the game, through trial and error, and playing millions of games against itself. While the program DeepBlue, which defeated Garry Kasparov at Chess in 1997, had been programmed by human beings to know which moves to make in every situation, AlphaZero used deep reinforcement learning to essentially learn how to play Go by itself from scratch. This is one example of the potential of machine learning algorithms that can also be used to solve more practical problems, such as image recognition in medical diagnostics and efficient energy management (Hassabis 2018).

Today, AI systems are increasingly used by governments and businesses to help make a range of difficult decisions, from an individual's eligibility for government benefits and prison sentences in the judicial system to content moderation on Facebook (LeCun 2020; Zimmermann, Di Rosa, and Kim 2020). They also have the potential to disrupt labor markets by fueling a new wave of automation (Stiglitz 2018). Yet, AI may not raise fundamentally new ethical challenges and some underline that the technology itself is neutral, only the way we choose to use it determines its good or bad effects (Hassabis 2018). While this may be true to a large extent, we still need to think about how to use this new technological tool ethically. What kind of ethical values should be embedded in the design of decision-making algorithms? How should individual privacy be protected, given the amount of personal data often required to train AI systems? How should machine learning algorithms be prevented from reproducing societal biases existing in their training data, which could lead them to discriminate against marginalized groups? How should the economic impact of automation be mitigated when a variety of skilled and unskilled jobs are at risk of becoming obsolete? These questions ask for an answer and responsible action from governments and businesses (Dubber, Pasquale, and Das 2020).

This article explores the cooperation of government and the private sector to tackle the ethical dimension of AI. The argument draws on the institutionalist approach in philosophy and business ethics defending a 'division of moral labor' between governments and the private sector (Rawls 2001; Scheffler and Munoz-Dardé 2005). The goal and main contribution of this article is to explain how this approach can provide ethical guidelines to the AI industry and to highlight the limits of self-regulation. In what follows, I discuss three institutionalist claims. First, principles of AI ethics should be validated through legitimate democratic processes. Second, compliance with these principles should be secured in a stable way. Third, their implementation in practice should be efficient. I argue that, if we accept these claims, there are good reasons to conclude that, in many cases, governments implementing "hard" regulation are in principle (if not yet in

practice) the best instruments to secure an ethical development of AI systems. Where adequate regulation exists, firms should respect the law. But when regulation does not yet exist, I argue that helping governments build adequate regulation should be businesses' ethical priority, not self-regulation.

The article is organized as follows. In Section 2, I introduce the institutionalist approach and the idea of a division of moral labor between governments and private agents. In Section 3, I discuss three institutionalist arguments in favor of tackling the ethical dimension of AI through hard regulation, not self-regulation. These arguments are based on the principles of legitimacy, stability, and efficiency. In Section 4, I discuss the objection of 'justice failure' in non-ideal circumstances, when governments fail at regulating technology. Under these circumstances, I argue that businesses' priority should be helping governments build adequate regulation and that self-regulation should only be a last resort.

## 2 Institutionalists and the Division of Moral Labor

A key insight from leading liberal philosophers of the late 20th century such as John Rawls is that the moral principles that should guide our personal lives are not necessarily the same as the political principles that should guide social life. The reason is that we live in pluralistic societies in which people have different moral beliefs, preferences, and interests. Therefore, while we should be free to live our personal life according to our own beliefs, institutions regulating social cooperation such as governments, the judicial system, or market regulations, should abide by 'public' standards of justice. These public standards should not derive from one single, comprehensive conception of morality but from an agreement acceptable by all, despite moral disagreements. Moreover, these public standards should not concern our personal lives but only institutions dealing with social conflicts (Porter 2009; Rawls 2001). A consequence of this separation between the 'moral' and the 'political', and of the need for political agreement, is that neither my personal moral beliefs nor yours should ever prevail in a just society. Only the principles we can agree on should.

Yet, despite pluralism and fundamental disagreements about moral values, partners involved in a given society have a shared interest in preserving social cooperation because it is mutually beneficial. This is at least one common ground on which to build compromise and to establish rules that everyone could agree on. Such rules can only foster stable agreement if they are the product of decision procedures perceived by all as legitimate and treating everyone fairly. Rawls hopes that people can reach an overlapping consensus about social rules: If each person

can find, in their own value system, a reason to agree on a set of principles of justice, then the agreement will be stable over time (Rawls 2001; Wenar 2017).

Authors in this liberal tradition such as John Rawls and Samuel Scheffler argue for a division of moral labor (or an institutional division of labor) between governments and private agents in the effort to realize a just society. In this view, governments should be in charge of validating principles of justice and implementing regulations to organize social cooperation. Once just regulations have been implemented, individuals and businesses simply have to follow the rules (e.g. vote, pay their taxes, stop at red lights, etc.) and their social obligations are fulfilled. This allows us to lead our lives as we see fit and frees us from the constant worry of doing what is right. For Scheffler, "the idea of a division of moral labour is best understood as the expression of a strategy for accommodating diverse values." He continues: "the idea of a division of moral labour embodies a strategy for resolving the tensions to which pluralism gives rise. If social institutions are designed in conformity with the principles of justice, then, it suggests, individual conduct within those institutions may legitimately be responsive to the various norms and ideals that govern our personal lives and interpersonal relationships" (Scheffler and Munoz-Dardé 2005, p. 229, p. 250; see also Porter 2009).

In this institutionalist approach to justice, governments are assumed to be the primary agents in charge of realizing a just society (Rawls 2001; Scheffler and Munoz-Dardé 2005; Scherer and Palazzo 2011). Secondary agents like individuals and businesses are expected to contribute only through proper public channels and to follow the law. For O'Neill (2001, p. 180) "there are often implicit assumptions that the primary agents of justice are states, and that all other agents or agencies are secondary agents of justice, whose main contribution to justice will be achieved by conforming to the just requirements of states." The assumption is that democratic administrations have the legitimacy, coercive power, and infrastructure required to enforce regulation and coordinate large-scale collective action (Friedman 1970; Rawls 2001; Scheffler and Munoz-Dardé 2005; Yeung, Howes, and Pogrebna 2020). Individualist approaches to justice, by contrast, do not deny the role of institutions but hold that businesses and individuals can also contribute to bringing about justice on their own (Berkey 2016; Murphy 1999; O'Neill 2001; Porter 2009; Weinberg 2009). In business ethics, this belief underlies many theories of corporate social responsibility (CSR) holding that firms should sometimes self-regulate and take on social responsibilities beyond what is required by law (Berkey 2016; Freeman 1984; Moriarty 2016). Some argue that managers should voluntarily supervise their firm's supply chain to prevent human rights violations or, in the case of AI, that they should develop safe AI systems aligned with societal values (Gabriel 2020). This is particularly relevant with new technologies because,

until regulations are updated, businesses seem to have no choice but to self-regulate.

The goal of this article is to draw on the existing institutionalist approach and the idea of the division of moral labor to explain how this approach can provide ethical guidelines to private actors in the AI industry and to give a response to the individualist approach in AI ethics which focuses on industry self-regulation. More specifically, I will explain why the institutionalist approach prioritizes "hard" regulation by governments over "soft" or "self" regulation, in many cases, and apply these arguments to some cases in AI ethics. The institutionalist argument on which this article will draw can be summarised as follows:

> Premise 1: Realizing a just society as quickly as possible should be everyone's ethical priority.
>
> Premise 2: Governments, in many cases, are the best means to realize a just society because they are in principle more legitimate, stable, and efficient than alternatives.
>
> Premise 3: Using suboptimal means delays justice at best and sustains injustice at worse.
>
> Premise 4: Therefore, in many cases, helping governments build adequate regulation should be the ethical priority of all private agents, including businesses.

The first premise 1 is an assumption that I do not discuss in this article. In a liberal framework, while everyone is free to lead their lives according to their own moral beliefs, everyone also has a duty to help establish and maintain just institutions because failing at doing so would perpetuate injustices (Rawls 1971, p. 115; Rawls 2001, p. 201). I will explain in detail the next two premises to provide guidance to private actors in the AI industry. Premise 2 claims that governments, in many cases, are in principle the best instruments to realize a just society. This disputes the idea underlying some individualist conceptions that "private governance systems might ultimately challenge existing state-centered authority and public policy-making processes" (Cashore 2002, p. 503). This premise 2 is supported by three institutionalist arguments based on the principles of legitimacy, stability, and efficiency. I will illustrate how these arguments apply to important issues in AI ethics and why, in these cases, governments are "in principle" better than private agents at realizing justice. I respond to the objection of "justice failure" in non-ideal circumstances in the last section. The conclusion remains that the best option to realize justice is often to implement adequate public regulation and that governments should not be supplanted by private governance systems.

Indeed, as premise 3 underlines, using suboptimal means to realize a just society delays justice at best and sustains injustice at worse. This premise disputes

the received idea that public regulation and individual initiative are always complementary and that individuals can choose whatever means they prefer. For example, Liam Murphy objects to premise 3: "it is obviously true that, as a practical matter, it is overwhelmingly preferable that justice be promoted through institutional reform rather than through the uncoordinated efforts of individuals" (Murphy 1999, p. 252) but for him, it cannot be right that "justice requires [a person] to promote just institutions even if she is sure that the aim of the just institutions she is promoting would be better served if she herself pursued that aim directly" (Murphy 1999, p. 281; see also Berkey 2016; Porter 2009). Murphy is right only if justice is indeed better served by individual action in some cases. I argue that this is rarely true because of opportunity costs: If individuals chose suboptimal means to contribute to justice, the right measures can be delayed and precious resources can be wasted, which is ultimately detrimental to justice. Therefore, premise 4, helping governments build adequate regulation should be the ethical priority of private agents, including AI businesses.

The main contribution of this article is not to justify the institutionalist approach (this is achieved in Rawls 2001; Scheffler and Munoz-Dardé 2005) but to explain how it can provide guidelines in important cases of AI ethics. I am not claiming that the narrow arguments in AI ethics cases establish the general truth of the institutionalist approach in all ethical cases nor that the general truth of the institutionalist approach is the reason why its arguments are convincing in all cases. I am only explaining the institutionalist arguments of legitimacy, stability, and efficiency and why, at least in some important AI ethics cases, this approach seems convincing and action-guiding. This can already be of interest for practitioners in the AI industry seeking ethical guidance on the best way to develop AI responsibly. Another contribution is to illustrate the appeal of the institutionalist approach through some revealing cases in AI ethics. While the institutionalist approach may not be relevant in all business ethics cases, the specific cases of AI ethics discussed in this article illustrate why this approach can be relevant in these and similar cases. This can be of interest for researchers wondering about the scope of the institutionalist approach.

## 3 Three Arguments for Public Regulation: Cases in AI Ethics

In this section, I present institutionalist arguments supporting premise 2, according to which governments, in many cases, are in principle the best means to realize a just society. This premise relies on the claims that the principles of justice,

and of AI ethics in particular, should be validated and scrutinized through legitimate democratic processes, that compliance to these principles should be secured in a stable way, and that their implementation in practice should be as efficient as possible. In what follows, I explain each of these claims and why they lead to the conclusion that, in many cases, helping governments build adequate regulation should be businesses' ethical priority. In each case, I argue that these arguments are relevant to thinking about some important cases in AI ethics.

## 3.1 Legitimacy and AI Value-Sensitive Design

There is a growing call for a "value-sensitive design" of AI systems. Indeed, when they are used to automate decisions involving value judgments, moral values must be embedded in AI systems. This can happen either when choosing which goal the system must optimize or, in a supervised learning process, when engineers tell the algorithm what right or wrong decisions are (Yeung, Howes, and Pogrebna 2020; Zimmermann, Di Rosa, and Kim 2020). Therefore, some reasonably argue that AI systems should be designed with values that "align" with societal values (Gabriel 2020).

**Case**: The challenge of such "value alignment" can be illustrated in the case of content moderation on social media. Facebook uses AI systems and machine learning to improve content moderation. Every single post or picture published on the platform is filtered through hierarchical neural networks to recognize the content of the post or picture and decide whether to show it or take it down (often, in ambiguous cases, with the help of human content moderators). For example, AI systems are trained to detect and take down terrorist propaganda, false accounts used to disseminate fake news, online harassment, and hate speech. There are obvious advantages to automating this process since it can improve the scale, speed, and accuracy of content moderation while reducing the psychological burden on human moderators. Yet, in doing so, AI systems make controversial moral judgments regarding what is "acceptable" or "unacceptable" speech (Gorwa, Binns, and Katzenbach 2020; LeCun 2020).

The problem is that "removing content which is not universally agreed to be harmful can … undermine users' freedom of expression" (Cambridge Consultants 2019, p. 5; Gorwa, Binns, and Katzenbach 2020). Given the power of social media platforms today and their critical role in amplifying or censoring speech at scale, some argue that platforms should respect free speech principles that align with societal values to secure a fair marketplace of ideas (Lambrecht 2020; Llansó et al. 2020; Sander 2020). While filtering out terrorist propaganda and direct incitation to violence, against which legislation already exists, may be morally uncontroversial,

other cases are more controversial. Examples include filtering out political judgments about a politician's fitness for office, satirical cartoons that some judge blasphemous or offensive, or depictions of nudity or violence in art or photojournalism (Cambridge Consultants 2019).

In these cases, Facebook's CEO Mark Zuckerberg claimed until very recently that "Facebook shouldn't be the arbiter of truth of everything that people say online" (McCarthy 2020). However, its current content policy seems controversial: "Facebook has been criticised for removing an image of a statue of Neptune in Bologna, Italy for being sexually explicit and the iconic photograph of a young girl fleeing a napalm bombing during the Vietnam War for showing child nudity" (Cambridge Consultants 2019, p. 5). Nevertheless, Facebook's Chief AI scientist Yann LeCun gives weight to the idea that the company should not be the arbiter of truth when it comes to controversial moral judgments and political opinions. He stresses that defining "hate speech", "valid" interpretations of facts, or the boundary between forms of expression that are "acceptable" and ones that "should be suppressed" is not a technological question with any technical fix. These are ethical questions that should be debated by a diverse and independent press and by civil society. Therefore, in a polarized political landscape in which Facebook's non-neutrality is a widespread worry on all political sides, the fact that Facebook has the power to impose your moral values on everyone and to shut down your political opponent does not mean that it should (LeCun 2020). The worry, of course, is that Facebook could use such power, not for the public good, but for misguided or self-interested purposes.

**Institutionalist argument**: The important point from an institutionalist perspective is that the principles of justice and, in our case, the principles of content moderation that Facebook should implement should be selected through legitimate decision procedures that everyone can agree on. Governments are, in many cases, the best agents to realize this end. The argument is that they are in principle better than private agents (even when well-intentioned) at building democratic deliberation mechanisms that can lead to legitimate compromises between people or groups with conflicting moral views.

Indeed, in pluralistic societies, we must expect moral disagreements about the right content moderation policy and reaching a workable compromise between citizens requires consensus-building strategies. The first and easiest strategy is to let everyone freely pursue their own beliefs, at least when their behavior does not affect anyone else. This is one reason justifying the liberal presumption of freedom and "against legal restrictions" (Rawls 2001, p. 44). A second consensus-building strategy when personal behavior can affect others is to seek an agreement between the people affected. The problem is that, in pluralistic societies, we cannot expect to base this agreement on the "right" moral answer, nor should we impose the

values of one group on everyone because this would lead to conflicts. A solution is to reach an agreement on a decision procedure that everyone accepts as fair. This is an argument for constitutional democracy. Because constitutional democracies combine constitutional rights, democratic representation, and other checks and balances, they can prevent both the tyranny of one and the tyranny of the majority. This helps provide public justifications for political decisions because such procedures are perceived as fair (Rawls 2001, p. 26). This is why governments are the best means to accommodate pluralism and reach legitimate agreements. Without such a legitimacy test, no one can claim to act on behalf of "justice" or "the public good", they can only claim to act on behalf of their own subjective belief. In Rawls' words: "a liberal political conception … is not reasonable in the first place unless it generates its own support in a suitable way by addressing each citizen's reason … Only so is it an account of political legitimacy as opposed to an account of how those who hold political power can satisfy themselves in the light of their own convictions that they are acting properly" (Rawls 2001, p. 186).

**Objection**: An objection consists in arguing that expert philosophers or business ethicists could anticipate what conception of justice or free speech is likely to generate a legitimate agreement before it goes through democratic processes. Experts could therefore tell us what are morally adequate restrictions of speech on social media. Once they have spoken, everyone can self-regulate following the experts' moral judgment. This is why Weinberg claims that we should "separate this picture of the just society from the means (the agency) by which it is achieved. This will leave the theory open to a range of possible agents of justice that may contribute to the realization of that theory's picture of the just society" (Weinberg 2009, p. 322).

A response to this objection relies on a crucial condition for the legitimacy of public decisions: the principle of publicity (Gosseries 2017). To be legitimate, public decisions, especially about basic rights such as free speech, should be open to everyone's scrutiny to make sure that agreed principles are adequately interpreted and implemented in practice and to hold decision-makers accountable when they are not. Democratic administrations are, in many cases, the most legitimate means to realize a just society because democratic processes and checks and balances can offer guarantees of publicity and freedom of information, as well as a wealth of experience and judicial decisions to adjudicate conflicts between rights in a way demonstrably consistent with agreed principles embodied in the law (Yeung, Howes, and Pogrebna 2020).

In the specific case of AI, a first worry with self-regulation in the AI industry has to do with what is often called "AI transparency". AI systems are often opaque because machine learning algorithms make decisions according to principles that are sometimes obscure even to programmers and certainly to

external stakeholders (Cambridge Consultants 2019; Gorwa, Binns, and Katzenbach 2020; Sander 2020). But this problem could eventually be solved by technological progress in "explainable" artificial intelligence (XAI) to secure better transparency in algorithmic decision-making (Andrus, Bhatt, and Xiang 2020; Diakopoulos 2020; Yeung, Howes, and Pogrebna 2020).

A second worry concerns the transparency of businesses themselves. To make sure that businesses like Facebook adequately implement agreed principles of justice, such as free speech in content moderation, and to hold them accountable when they do not, they must subject their use of AI systems to public scrutiny and independent audit (Gorwa, Binns, and Katzenbach 2020; Kroll 2020; Yeung, Howes, and Pogrebna 2020). Making sure that we hold decision-makers accountable is important to comfort everyone in the knowledge that no one can impose their personal beliefs or interests to the detriment of others (Baumol 1974; Rawls 2001, p. 186). To this end, governments must impose transparency standards and independent oversight on all private agents (Kroll 2020; Sander 2020).

To conclude, following legitimate principles when designing AI systems to automate content moderation on social media is important to protect free speech. Legitimacy is also important in other areas of value-sensitive design. Examples include defining acceptable ways to use algorithmic profiling and targeting in political and commercial advertising (O'Neil 2016; Sander 2020), identifying demeaning gender and racial stereotypes to avoid reproducing or amplifying them through search engines and recommendation algorithms (Llansó et al. 2020) or selecting the relevant definition of fairness when tackling algorithmic bias and discrimination. As Zimmermann, Di Rosa, and Kim (2020) sum up, "we need greater democratic oversight of AI not just from developers and designers, but from all members of society."

## 3.2 Stable Compliance and AI Safety

AI safety is another important area of AI ethics. Indeed, the safe development of AI systems requires a variety of safeguards from privacy protection and fairness guarantees in algorithmic decision-making, to meaningful human control of automated weapon systems, to making sure that learning algorithms that could continue to evolve in unpredictable ways after delivery remain safe for use over time (Andrus, Spitzer, and Xiang 2020; Hassabis 2018).

**Case**: The challenge of AI safety can be illustrated in the case of privacy protection. Learning algorithms often have to be trained on extensive amounts of data, which fuels AI firms' thirst for personal data. For example, in 2017 the UK Information Commissioner's Office ruled that the transfer of the personal data of

1.6 million patients from a London hospital to DeepMind failed to comply with the UK's Data Protection Act (Hern 2017). One particularly interesting piece of privacy regulation is the European General Data Protection Regulation (GDPR) because of the importance it gives to individual consent. Article 6 allows processing personal data for any one of six reasons, including "free, informed, and unambiguous consent" by individuals (European Union 2020). One positive aspect of that regulation is that Article 5 specifies general principles applying even when data subjects give their consent that happen to include conditions akin to Helen Nissenbaum's famous principle of "contextual integrity", which requires respecting the context in which data subjects have consented to disclose personal information (Nissenbaum 1998). Even when data subjects give their consent, the transparency and purpose limitation principles specify that data must be processed by businesses in a transparent way and only for the legitimate purposes specified explicitly to the data subject when they collected it, the data minimization principle specifies that they should collect and process only as much data as necessary for the purposes specified, and other principles of storage limitation, integrity and confidentiality prevent businesses from shifting the data to third parties (voluntarily or not), notably by mandating efforts to prevent data breaches and accidental loss (European Union 2020).

Yet, the GDPR allows businesses to process personal data as soon as there is free, informed and unambiguous consent. This is a problem for two main reasons. First, because AI systems also increase the capacity to analyze data already disclosed by consenting individuals to uncover new information that they did not intend to disclose, not only about themselves but also about un-consenting people "like them". For example, governments or businesses can use facial recognition systems and existing social media pictures to predict every individual's sexual orientation, they can use criminality and poverty data in a given zip code to guess an individual's recidivism or credit default risk, or even use shopping habits to find out whether their customers are pregnant (Duhigg 2012; O'Neil 2016; Zimmermann, Di Rosa, and Kim 2020). This is why Fairfield and Engel claim that privacy is a public good and that "your privacy is not yours alone". Today, being cautious about disclosing your own data is not enough to secure your privacy, you can also be vulnerable merely because others have been careless with their data. As a result, "protection requires group coordination" to secure the "optimal level of privacy" in society as a whole (Fairfield and Engel 2015, p. 387, p. 424).

Second, the emphasis on consent is also a problem because there are other reasons to protect privacy beyond the obligation to respect individual choices. As Solove notes, when governments collect and process large amounts of personal information in unaccountable ways, "it creates a power imbalance between individuals and the government …This issue is not about whether the information

gathered is something people want to hide, but rather about the power and the structure of government" (Solove 2007, p. 767). Both governments and businesses can acquire disproportionate power in this way. Governments can use facial recognition to spot and target leaders of democratic protests and engage in widespread surveillance with chilling effects on public discourse (Fairfield and Engel 2015; Powers and Ganascia 2020). Political parties can team up with data analytics firms like Cambridge Analytica to engage in digital tracking, profiling, and targeting which are powerful tools to influence electoral outcomes and shape public policy (Manheim and Kaplan 2019; Yeung, Howes, and Pogrebna 2020). Businesses can uncover information about employees to target union leaders, track "labor organizing threats", or discriminate against protected groups, and they can use advances in behavioral science to manipulate customers (Ajunwa and Schlund 2020; Duhigg 2012; Franceschi-Bicchierai 2020). The resulting power imbalances threaten the protection of principles of justice and democracy such as political equality and non-domination. Therefore, determining the optimal level of privacy is a controversial issue and could imply having to collectively decide what level is necessary to limit power imbalances between powerful governments or businesses and powerless citizens.

**Institutionalist argument**: The important point from an institutionalist perspective is that the protection of principles of justice and, in this case, an optimal level of privacy should be a stable feature of society. Governments are, in many cases, the best agents to realize this end. The argument is that they are often better than private agents (even when well-intentioned) at building safe coercive mechanisms to secure stable compliance. To understand why, we need first to understand why stability matters and, second, why governments are able to build safer coercive mechanisms.

First, the stability of justice requires guarantees of stable compliance. As Rawls (2001) argues, society is truly just only when public institutions guaranteeing justice and, in our example, privacy are stable over time. These institutions should not be stable merely because the ones in power are able to impose their will on others, they should be stable because they create the conditions for their own support and citizens living under them "acquire a reasoned and informed allegiance to those institutions sufficient to render them stable" (Rawls 2001, p. 185). One condition to create the right kind of stability is the legitimacy of decision procedures, for instance the legitimacy of the decision regarding the optimal level of privacy to protect in society (Rawls 2001, p. 186). Another condition is that institutions must generate the expectation that they will always treat people equally, that justice will be served predictably, and that they will not suddenly subject citizens to arbitrary power, domination, and injustice, for example, that they will not suddenly lose their privacy protection. This disposes citizens to

further support and trust social institutions (Rawls 2001, p. 196). In order to create the right kind of stability, therefore, fair treatment or the protection of privacy cannot merely be punctual properties of social institutions that can disappear overnight, they must be a permanent and predictable feature. This is why institutions making sure that everyone, including public officials and business leaders, reliably complies with collective rules are necessary to create a truly just society (Rawls 2001, p. 180–198). By contrast, making justice or the protection of privacy rely on the benevolence of people in power would fail to prevent the risk of domination and abuse of power. Indeed, even if a "kind" and "responsible" slave owner does not use their power to harm their slaves at a given point in time but, as slaves, they remain powerless and under the risk of harm, so they remain unjustly dominated. Therefore, society cannot be just without establishing guarantees against abuses of power, both in government and in business, that secure legal recourse to prevent the backsliding of protections over time (Anderson 2017; Rawls 1993, p. 269; Rawls 2001, p. 53).

Second, public institutions are often better than private actors at securing stable compliance. While the stability of just institutions and the respect for privacy regulation over time cannot merely rely on coercion, there are nevertheless legitimate justifications for introducing coercive mechanisms and sanctions to prevent injustice and non-compliance. Joseph Heath underlines the importance of self-binding rules. Individuals engaged in a cooperative effort, even when they have the best intentions, can nevertheless be subjected to dynamic preference inconsistency: We may have a general preference for doing what is right, and yet we are sometimes tempted to act in self-serving ways. Therefore, willingly accepting to constrain our own future choices is sometimes rational if we want to reliably attain our collective goals (Heath 2006, p. 324). Yet, if we are to allow some powerful organizations to impose large-scale self-binding rules, we have to do it safely by instituting constraints on power and checks and balances.[1] With this in mind, government should be preferred over private agents when it comes to imposing self-binding rules, not only because they can effectively coerce everyone (Simon 2000) but also because they can often do so more safely. They can create legal safeguards and checks and balances to prevent potential abuses of power and arbitrary interference by the people in charge (Friedman 1970).

**Objection**: A potential objection consists in questioning such a pessimistic conception of human nature. We should perhaps trust that most public officials

1 As James Madison once wrote: "Ambition must be made to counteract ambition… It may be a reflection on human nature, that such devices should be necessary to control the abuses of government. But what is government itself, but the greatest of all reflections on human nature? If men were angels, no government would be necessary" (Federalist 51, cited in Waldron 2012).

and business leaders have good intentions. Perhaps the tech giants mean it when they claim to be concerned about their users' privacy. A response is that failure to comply with our moral or legal obligations does not always derive from evil intentions but from conflicting moral obligations, situations where there is no good outcome, or weakness of the will. Therefore, even well-intentioned people can be subjected to dynamic preference inconsistency and fail to comply with their obligations. Among factors explaining non-compliance, there is value pluralism itself and the prevalence of the belief that we have a greater moral responsibility for what we do ourselves than for what we merely fail to prevent. Another is that our motivation tends to dwindle when doing the right thing demands a substantial personal sacrifice. Some could also doubt the effectiveness of such personal sacrifice, especially if they doubt that others will comply with their obligations. Finally, if free-riding is the norm, people tend to follow the crowd even when compliance would be mutually beneficial (Heath 2014, p. 294–321; Scheffler and Munoz-Dardé 2005, p. 230–232). For example, despite all their talk about beneficial, inclusive, and responsible AI, many large tech companies engage in tax avoidance schemes and fail to comply with the minimum requirement of paying their fair share of taxes (Dietsch 2011, 2015; Reuters 2019).

In the case of businesses, in particular, their capacity to reliably do the right thing (e.g. to protect privacy) is limited because of structural constraints, like competition, that often force them to focus on short-term profits and imitate their competitors' harmful behavior. In fact, there are good normative reasons to put firms in competitive environments: Well-designed competitive markets allow for lower prices, provide incentives to invest in innovation, and allocate resources efficiently by decentralizing economic decisions through the price system, which produces economic benefits (Heath 2014, p. 25–36). But as a result of competition, firms have a limited capacity to be agents of justice and are incentivized to engage in regulatory arbitrage and to exploit loopholes in regulation (Baumol 1974; Slee 2020; Yeung, Howes, and Pogrebna 2020). While many business leaders and ethicists still believe that CSR is always good for business, David Vogel (2005) argues that no systematic link between "good behavior" and "profit" can be empirically established. If Google, Apple, Facebook, and other visible firms well-known by customers, can sometimes use ethical initiative in marketing, small firms may not have this option. Even when businesses can strategically profit from ethical behavior, incentives are structured in a way that pushes them to pursue ethical goals only up to the point necessary to reach their strategic goal, but not further. Finally, while Vogel himself believes that there is a "market for virtue", he underlines that bad practices still pay. This means that, sooner or later, many firms have to choose between ethics and profit, and structural economic constraints

often lead them to choose profit. This is why he insists that "governments remain essential to improving corporate behavior" (Vogel 2005, p. 170).

To conclude, with the rise of AI and data analytics, it is important to secure stable compliance with privacy regulations to secure an optimal level of privacy and avoid the power imbalances that would result from a sub-optimal level of privacy protection. Stability is also important in other areas of AI safety. An example is AI fairness and bias in algorithmic decision-making. There is empirical evidence that machine learning algorithms fed with biased data can replicate historical injustice. Even accurate algorithms can perpetuate existing injustices if they operate in societies with severe background injustices because they more accurately target marginalized people and exclude them from social advantages. For example, algorithms can help decide an individual's eligibility to entitlements and benefits, including housing, social security, and bank loans based on their risk profile, so accurate algorithms could actually make social exclusion even worse. The result is that AI systems can help maintain inequalities of wealth and power in society (Gebru 2020; O'Neil 2016; Yeung, Howes, and Pogrebna 2020; Zimmermann, Di Rosa, and Kim 2020). Securing stable compliance with anti-discrimination rules is important to avoid the backsliding of protection against domination and resulting power imbalances. This cannot rely on industry self-regulation but must rely on adequate regulation and reliable enforcement.

## 3.3 Efficiency and AI's Distributive Impact

AI systems are also likely to impact the future of work and distributive inequalities in society. While there is some unwarranted alarmism in this area, even industry actors acknowledge that "AI is poised to reshape the global economy and change the makeup of the skills required to succeed in it. The burden of adjusting to these changes is placed by default on those who have the fewest resources to bear it" (Klinova 2020). Therefore, we should investigate how to make sure that AI systems are implemented in a way that contributes to an inclusive economy.

**Case**: The challenge of making sure that AI systems contribute to an inclusive economy can be illustrated by the complexity of evaluating their distributive impact. To begin, AI systems will push further the process of work automation, which is not new and could have positive distributive effects thanks to the overall productivity gains of Schumpeterian creative destruction (Schumpeter [1950] 2008; Stiglitz 2018). Indeed, just like any technological innovation, AI systems will lead to the destruction of previous business models and their replacement with more productive ones. If productivity gains are redistributed, everyone can benefit from this process. However, this will also restructure entire labor markets and

types of work, as Amazon did with retail. One worry is that AI systems are now able to replace not only less qualified workers like truck drivers but also highly qualified workers such as lawyers and radiologists. (Indeed, AI systems are now able to find loopholes in contracts and read magnetic resonance images.) While such technology could augment the productivity of workers or change their task composition, they could also de-professionalize jobs because poorly-paid, less qualified workers with AI systems could outperform current professionals. In the long run, this process could still improve the productivity of the economy, but how labor market restructuration will impact wages and unemployment depends on how productivity gains are distributed, which depends on government policy (Moradi and Levy 2020; Stiglitz 2018).

One problem is that workers losing their jobs or relocating because of degrading working conditions have to bear important transition costs over the short term. A mitigating strategy involves Keynesian economic policies to absorb these costs collectively, like unemployment benefits, retraining programs to improve labor skills, and public investments to create new jobs (Keynes [1936] 2001; Moradi and Levy 2020; Stiglitz 2018). Yet, the expected pace of automation could make this strategy less effective, especially for poorer countries that cannot afford social benefits and retraining programs, and could be confronted with premature deindustrialization, in which countries start to lose their manufacturing jobs before they had the chance to properly develop or transition towards a service economy (Rodrik 2015).

Besides automation, the development of AI systems can increase distributive inequalities in various other ways. One problem is access: While AI systems are expected to help individuals in various ways (e.g. improve risk mitigation in financial decisions and personalize education to help kids learn in schools), unequal access to these technologies could increase existing socio-economic inequalities. Another problem is risk-shifting: Businesses can use AI to predict more accurately fluctuating consumer demand and then shift this risk onto workers by scheduling on-call or split shifts, and generally more precarious and unstable schedules. An important problem is the monopolistic tendencies in the ownership of the technology itself and the infrastructure necessary to run AI systems: This can lead to market power inequality in the tech industry and increase inequalities in society. Therefore, how we choose to organize and regulate the use of these technologies and how we design taxes and benefits will make a difference in the distributive impact of AI (Moradi and Levy 2020; Stiglitz 2018).

**Institutionalist argument**: The important point from an institutionalist perspective is that realizing a just distribution of resources and, in this case, mitigating the distributive impact of AI should be done as efficiently as possible. Governments are, in many cases, the best agents to realize this end. The argument

is that they are often better than private agents (even when well-intentioned) at centralizing information to coordinate the kind of collective action required to realize a just distribution of resources. Private agents with the best intentions are still more likely to face coordination failure. To understand why, we must first discuss the importance of centralized information and coordination to realize distributive justice and, second, why governments are often better at this than private agents.

First, centralizing information is crucial to realizing a just distribution of resources. As Joseph Heath (2006, p. 315–316) notes, "much of contemporary social contract theory has been marked by… a tacit conceptual privileging of gains from trade as the primary mechanism of cooperative benefit." Instead, he underlines that social cooperation can produce efficiency gains in various other ways, such as economies of scale, risk pooling, and information transmission (Heath 2006, p. 319–322, p. 327). In particular, when aiming at realizing a just distribution of resources, information transmission is crucial. This is because we must make sure that everyone gets their due, no more, no less. This raises at least two challenges. To begin, if we take the example of a conception of fairness giving some priority to the worst-off (Rawls 2001), such a principle cannot be realized without centralized information and coordination because allocating by mistake some resources to better-off people instead of the worst-off constitutes an injustice. Moreover, institutional roles may require private agents to act in ways that do not directly aim at solving injustices – such as competing in the market – but end up contributing to a just society. This means that it is hardly possible to evaluate how just or unjust a particular action is unless we have a global view of the whole system (Rawls 1993, p. 267-69; Rawls 2001, p. 54; Scheffler and Munoz-Dardé 2005, p. 249–50). Thus, information transmission matters both to identifying everyone's fair share and to coordinating everyone's effort at distributing resources fairly.

Second, governments are often better than private agents (even when well-intentioned) at coordinating the fair distribution of resources. First, they can centralize information about the actions of all citizens more efficiently (e.g. thanks to national statistics and tax agencies). This allows them to make accurate judgments about who the worst-off are and how to maximize their situation. Second, they can design coherent and efficient policies securing everyone's simultaneous access to their fair share. Regarding the distributive impact of AI, governments can regulate tech companies, force them to pay their fair share of taxes, secure everyone's fair access to the benefits of AI, and provide workers with the training necessary to take advantage of the opportunities of the digital economy. In fact, the superior capacity of governments in coordinating the distribution of resources is a general problem in business ethics that extends beyond the distributive impact of

AI. As Vogel (2005, p. x) notes, "while there is a rich tradition of public policy analysis that evaluates the costs, benefits, and impact of government decisions, nothing comparable exists for the realm of 'private policy' in which CSR is located." The difficulty to identify how the AI industry can contribute to a fair and inclusive economy comes from the need to have a global picture of the overall distributive effect of AI. Therefore, even if we assume that private actors have the best intentions, realizing an inclusive economy cannot be left to uncoordinated private initiatives but requires adequate public policies to avoid coordination failure.

Finally, there is another sense in which governments can realize justice "more efficiently": They allow for a lower "mental load" for each of us. Indeed, they allow us "to abstract from the enormous complexities of the innumerable transactions of daily life and frees us from having to keep track of the changing relative positions of particular individuals" (Rawls 2001, p. 54). Given the scale of modern economies and the complex ways in which new technologies such as AI impact the distribution of resources, gathering the information necessary to know how best to contribute to distributive justice is too demanding for any single individual or business and, as Scheffler argues, may simply be beyond their capacity (Scheffler and Munoz-Dardé 2005, p. 244). Instead, the institutionalist approach and its division of moral labor allow us to realize the same goal – a fair society – in a less demanding way. Just institutions relieve us of the mental load of constantly wondering whether our daily choices contribute or not to social justice (Murphy 1999). For Rawls (1993, p. 269) "if this division of labor can be established, individuals and associations are then left free to advance their ends more effectively within the framework of the basic structure, secure in the knowledge that elsewhere in the social system the necessary corrections to preserve background justice are being made."

**Conclusion**: Using the institutionalist argument, I assumed premise 1, according to which realizing a just society as quickly as possible should be everyone's moral priority. The arguments above support premise 2, namely, that governments, in many cases, are in principle more legitimate, stable, and efficient than private agents at realizing justice. I noted that institutionalists only need to demonstrate that public institutions are often better than private agents. Indeed, if this is true, then premise 3 holds and there is an opportunity cost of relying on suboptimal methods: Precious time and resources would be wasted on illegitimate, unstable, and inefficient means, which can at best delay justice and at worse prevent it. Therefore, premise 4, in many cases, helping governments build adequate regulation should be the ethical priority of all private agents, including the businesses in the AI industry.

# 4 Justice Failure and Businesses' Obligation to Help Build AI Regulation

In this section, I respond to an important objection against the institutionalist approach to AI ethics. Authors in the institutionalist tradition argue that when just institutions are already in place individuals merely have a duty to maintain them (Rawls 1971, p. 115; Rawls 2001, p. 201). For example, this means maintaining the government's capacity to enforce adequate AI regulation. The claim that this approach relieves private actors from any responsibilities and is not demanding enough is misleading because, contra O'Neill (2001, p. 181), institutionalist approaches ask private agents to actively maintain just institutions, not merely conform to their demands. Even skeptics of institutionalist approaches, like Porter (2009), acknowledge that such a duty is very demanding and requires individuals and businesses to play an active part in public debates and deliberative processes and to exercise scrutiny over public decisions and regulatory compliance (p. 177). The real issue arises when institutions are not already just.

While governments may be, in principle, more legitimate, stable, and efficient than private agents at realizing a just society, they often fail at their task in non-ideal, real-world circumstances. Abraham Singer (2018, 2019) famously coined the concept of "justice failure" to refer to these non-ideal circumstances. His argument is a counterpoint to the idea of the division of moral labor according to which, within a larger scheme of social cooperation, markets ought to pursue efficiency and leave the pursuit of justice to governments and the welfare state. He argues that just as market failures lead to suboptimal efficiency in actual markets, justice failure leads to suboptimal fairness in actual welfare states (Singer 2018, p. 97). For example, authoritarian states lack democratic legitimacy and weak states lack the independence or the effective enforcement capacity necessary to secure public trust and accountability. Even strong democracies are sometimes ill-equipped to monitor the industry and secure stable compliance to public regulation, especially at the global level (O'Neill 2001, p. 182). Finally, governments often lack information about the fast-paced evolution of business and technology making regulations ineffective (Néron 2010, 2016, p. 716).

This is a particular iteration of a long-standing debate in political philosophy regarding "non-ideal theory" (Valentini 2012). There are various ways to understand "ideal" and "non-ideal" circumstances that I cannot fully summarize here. For the purposes of this article, I want to clarify that I do not call circumstances "ideal" where all relevant agents comply with the demands of justice applying to them (Rawls 2001), nor do I call a utopian society "ideal justice" where it is defined independently of feasibility constraints (Cohen 2003; see also Valentini 2012, p. 654).

In my view, we must always accept individuals as they are and, accordingly, think of institutions as they should be. Even in ideal circumstances, the burdens of judgment will likely lead to reasonable disagreements about what is right and wrong (Rawls 2001, p. 35), and conflicts of values and interests, weakness of the will, and dynamic preference inconsistency will remain (Heath 2006, p. 324). Therefore, we must define just institutions while keeping in mind the possibility of non-compliance as well as realistic feasibility constraints. This is why, even in a just society, we still need a constitution and checks and balances to prevent abuses of power by government officials, a judicial system to adjudicate conflicts, and protections of the rule of law. This is also why we must accept that corporations face competitive constraints that limit their capacity to engage in corporate social responsibility. This would be unnecessary if we assumed a society of angels. Instead, I define circumstances as "ideal" where institutions are designed adequately to realize justice and circumstances as "non-ideal" where there is "justice failure", that is, governments are failing at realizing justice (Singer 2018, p. 97).

Under this definition, ideally, when public institutions are just, most of us can focus on doing business while respecting the law. But in non-ideal, real-world circumstances, when institutions are unfair, absent, or ineffective, the nature of our ethical obligations change. We can no longer be content with following the law but we should take on more responsibility in realizing a just society. On this ground, some object that the institutionalist approach cannot guide individuals in non-ideal circumstances (O'Neill 2001, p. 182). This is particularly relevant to discussions on AI and new technologies because, until regulations are adequately updated, businesses seem to have no choice but to self-regulate.

An institutionalist response is that we also have a duty to help build just institutions when they do not already exist (Hsieh 2009; Rawls 1971, p. 115; Rawls 2001, p. 201). But, contra Berkey (2016, p. 732), the fact that governments sometimes fall short is not a green light to the individualist approach. Instead, the same reasons justifying why governments are, in principle, the best instruments to realize justice should continue to guide agents in non-ideal circumstances and justify that the best strategy and top priority should be to build just institutions as quickly as possible. In the business sector, this means that firms should get involved in the political process and engage in political CSR (Baumol 1974; Hsieh 2009; Néron 2010, 2016; Scherer and Palazzo 2011; Singer 2019). As Vogel (2005) points out, "corporate responsibility should be about more than going 'beyond compliance'; it must also include efforts to raise compliance standards. In fact, the most critical dimension of corporate responsibility may well be a company's impact on public policy. A company's political activities typically have far broader social consequences than its own practices" (p. 171). Therefore, the institutionalist

approach can continue to provide ethical guidance to private agents in the AI industry even in case of justice failure: Helping governments build adequate AI regulation should often be their ethical priority in non-ideal circumstances.

**Legitimacy**: The AI industry must first help improve government's legitimacy by contributing to improving the transparency, oversight, and accountability of the regulatory process. When the countries in which they operate lack legitimate administrations, they should get involved in the political process, helping citizens in establishing just democratic procedures, supporting the development of a thriving civil society, and fighting corruption (Hsieh 2009, p. 260–64; Singer 2019, p. 244; Vogel 2005, p. 171–173). Even in well-functioning democracies, the AI industry can support public debates by encouraging stakeholder consultation in the regulatory process. They can do so by providing public officials, academics, and civil society with the opportunities, resources, and information necessary to engage in effective discussions about AI regulation. Good examples are the global Partnership on AI, announced in 2016 and now gathering all big tech companies as well as members of academia and civil society, or the Royal Society's initiative in the UK, You and AI, supported by DeepMind in 2018, which provided an opportunity for public debate on AI (Hassabis 2018). But improving the legitimacy of public regulation also requires powerful actors such as Google, Amazon, Facebook, Apple, and Microsoft to refrain from activities that make governments less legitimate, such as financing political campaigns, using economic threats to promote self-serving regulation, or engaging in corruption (Baumol 1974; Singer 2019, p. 245).

**Stability**: The AI industry must help governments secure stable compliance. One way consists in supporting binding "hard" legislation, instead of self-regulation, which has a greater impact by guaranteeing the compliance of entire industries rather than only self-regulating firms (Baumol 1974; Hsieh 2009, p. 269; Vogel 2005). Businesses should also gather data about the performance of their AI systems before and after deployment to help the government monitor compliance with public standards (Yeung, Howes, and Pogrebna 2020). And while businesses should refrain from pushing self-serving agendas, when democracy itself is at risk, businesses could condition further investment on the establishment of democratic safeguards. Indeed, one way in which businesses can exercise pressure on governments is through the exercise of their property rights: "A firm might move out of a state in response to the passage of a law it does not favor, or it may threaten to move out of a state if such a law is passed. This may cause the state's citizens to revise or edit their political decisions" (Moriarty 2016; see also Hsieh 2009).

**Efficiency**: The AI industry must finally help governments implement more efficient policies. To begin, they should simply pay their taxes to help fund effective retraining programs and generous enough social benefits to collectively

absorb the transition costs generated by disruptive AI technologies (Singer 2019; Stiglitz 2018). But businesses should also participate positively in lawmaking by offering valuable information about the AI systems they are developing to help anticipate potential ethical issues and find technical and regulatory solutions. Indeed, a problem facing governments is that regulation is slow-moving and can struggle to keep up with the fast-paced development of AI and new technology in general. This is already true in ideal circumstances and the problem worsens in non-ideal circumstances. Therefore, the AI industry may have an obligation to cooperate with governments to anticipate potential problems raised by new technologies in order to update legislation in time. AI firms such as DeepMind already have an Ethics & Society unit running initiatives of this sort (Hassabis 2018). The AI industry can also help governments find loopholes or update obsolete legislation (Dignum 2020; Yeung, Howes, and Pogrebna 2020). As Néron (2016) notes, "regulatory systems are sometimes slow to evolve and adapt to complex and rapid changes, especially in some industries producing innovative technologies. Facing change and complexity, governments and regulatory bodies need to work in partnership with industry's key actors, and sometimes rely heavily on industry research and information" (p. 716).

Therefore, in non-ideal circumstances, the first-best strategy and top priority for businesses and other private agents should often consist in supporting legitimate government intervention and improving regulation as quickly as possible. Only when the first-best strategy is no longer possible can the second-best strategy become permissible as a last resort: using suboptimal means such as self-regulation. Even then, when collaboration with governments is not possible and only self-regulation remains, private agents should aim at improving legitimacy, stability, and efficiency.

**Legitimacy**: AI industry leaders should set aside their own moral beliefs and follow legitimate public guidelines. Examples include the Universal Declaration of Human Rights or the United Nations' Global Compact, which provide guidelines for corporations regarding human rights, working conditions, corruption, and the environment. The Global Compact Board has some legitimacy by being constituted of people representing the United Nations members, businesses, civil society, and unions. In the case of AI ethics, in particular, similar international deliberative processes have produced best practices standards for the responsible development of artificial intelligence (on AI governance by human rights-centered design, see Yeung, Howes, and Pogrebna 2020). Examples include UNESCO's global consultation on AI and the Montreal Declaration for the Responsible Development of Artificial Intelligence. Involving stakeholders affected by AI systems in the deliberative process is one way to improve firms' legitimacy in private governance and self-regulation (Dignum 2020; Yeung, Howes, and Pogrebna 2020).

**Stability**: AI businesses should create self-binding mechanisms within their own organization, including impact assessment, independent oversight, verification and accountability mechanisms, and internal sanctions (Kroll 2020, p. 181–196; Vogel 2005, p. 164). Yet, voluntary self-regulation remains unstable since the operation of market forces disincentivizes business organizations to comply with their own ethical standards, unless proper oversight mechanisms hold decision-makers accountable and impose consequences for non-compliance (Slee 2020; Yeung, Howes, and Pogrebna 2020).

**Efficiency**: Finally, while AI businesses should directly support new regulations, in the meantime, they can also pave the way through self-regulation by following the rules expected in upcoming legislation. This can facilitate the swift adoption and implementation of future regulations (Néron 2016, p. 716; Vogel 2005, p. 168–173). Moreover, when they engage in self-regulation or corporate social responsibility in particularly burdened societies, they should focus on initiatives on which they have the most information and coordination capacity and take seriously the responsibility to act in coordination with other businesses and civil society in order to put people at the centre of development efforts, build local capacity (Hsieh 2009, p. 263–264) and avoid pulling in different directions or stopping collective efforts that can delay the realization of justice.

**Conclusion**: While governments often remain the first-best option to implement ethical standards in the AI industry, in non-ideal circumstances well-intentioned private actors can offer a valuable contribution, either by helping build just public institutions or, when this is not possible, by abiding by voluntary ethical standards. That many private actors are well-intentioned does not mean that they can easily realize justice on their own, because they are not in the best position to secure legitimate standards, stable compliance, and efficient collective coordination. This is why they should channel their efforts into political action and collaborate with governments, because social justice is often best served when voluntary standards become legally binding (Vogel 2005, 163).

# 5 Conclusion

I argued that the institutionalist approach in philosophy and business ethics can provide useful guidance to business actors and practitioners in the AI industry. I concluded that, in many important cases, helping governments build adequate regulation should be the ethical priority of private agents, including businesses in the AI industry. The ethical framework presented here differs from other authors in

two important ways. On the one hand, I propose that business actors like the AI industry should help build just institutions when they do not exist. Therefore, I agree to some extent with Abraham Singer's "justice failure" approach, which claims that businesses have a larger set of political obligations (Singer 2018) compared to authors proposing to restrict firms' political obligations to tackling market failures (Baumol 1974; Heath 2014). Focusing on a market failures approach allegedly identifies clearer goals for firms' political behavior (Néron 2016, p. 725), but in a pluralistic society people have different values and interests, including about the role that markets should play in society, so promoting perfectly competitive markets is not necessarily clearer or less controversial. On the other hand, Singer and others argue that firms have a large variety of ethical obligations and special duties towards stakeholders and believe that political duties are only one obligation among others social and distributive duties (Moriarty 2016; Néron 2010; Scherer and Palazzo 2011; Singer 2019). I argued instead that, in many cases, the priority of AI businesses leading the fourth industrial revolution should be to cooperate with governments and international bodies to improve public regulation, except in exceptional circumstances when this first-best strategy is impossible and self-regulation is the only option.

## References

Ajunwa, I., and R. Schlund. 2020. "Algorithms and the Social Organization of Work." In *The Oxford Handbook of Ethics of AI*, edited by M. D. Dubber, F. Pasquale, and S. Das, 805–22. Oxford: Oxford University Press.

Anderson, E. 2017. *Private Government: How Employers Rule Our Lives (and Why We Don't Talk About It)*. Princeton: Princeton University Press.

Andrus, M., U. Bhatt, and A. Xiang. 2020. "Multistakeholder Approaches to Explainable Machine Learning." Partnership on AI. Retrieved from https://www.partnershiponai.org/multistakeholder-explainableml/.

Andrus, M., E. Spitzer, and A. Xiang. 2020. "Working to Address Algorithmic Bias? Don't Overlook the Role of Demographic Data." Partnership on AI. Retrieved from https://www.partnershiponai.org/demographic-data/.

Baumol, W. J. 1974. "Business Responsibility and Economic Behavior." In *Managing the Socially Responsible Corporation*, edited by M. Anshen, 59–71. New York: MacMillan.

Berkey, B. 2016. "Against Rawlsian Institutionalism about Justice." *Social Theory and Practice* 42 (4): 706–32.

Cambridge Consultants. 2019. "Use of AI in Online Content Moderation." *Ofcom*. Retrieved from https://www.ofcom.org.uk/.

Cashore, B. 2002. "Legitimacy and the Privatization of Environmental Governance: How Non–state Market–Driven (NSMD) Governance Systems Gain Rule–Making Authority." *Governance* 15 (4): 503–29.

Cohen, G. A. 2003. "Facts and Principles." *Philosophy & Public Affairs* 31 (3): 211–45.

Diakopoulos, N. 2020. "Transparency." In *The Oxford Handbook of Ethics of AI*, edited by M. D. Dubber, F. Pasquale, and S. Das, 197–213. Oxford: Oxford University Press.
Dietsch, P. 2011. "Asking the Fox to Guard the Henhouse: The Tax Planning Industry and Corporate Social Responsibility." *Ethical Perspectives* 18 (3): 341–54.
Dietsch, P. 2015. *Catching Capital: The Ethics of Tax Competition*. Oxford: Oxford University Press.
Dignum, V. 2020. "Responsibility and Artificial Intelligence." In *The Oxford Handbook of Ethics of AI*, edited by M. D. Dubber, F. Pasquale, and S. Das, 215–31. Oxford: Oxford University Press.
Dubber, M. D., F. Pasquale, and S. Das, eds. 2020. *The Oxford Handbook of Ethics of AI*. Oxford: Oxford University Press.
Duhigg, C. 2012. "How Companies Learn Your Secrets." The New York Times Magazine. Retrieved from https://www.nytimes.com/.
European Union. 2020. "What is GDPR, the EU's New Data Protection Law?" gdpr.eu. Retrieved from https://gdpr.eu/what-is-gdpr/.
Fairfield, J. A. T., and C. Engel. 2015. "Privacy as a Public Good." *Duke Law Journal* 65 (3): 385–457.
Franceschi-Bicchierai, L. 2020. "Amazon Is Hiring an Intelligence Analyst to Track 'Labor Organizing Threats'." Vice. Retrieved from https://www.vice.com.
Freeman, R. E. 1984. *Strategic Management: A Stakeholder Approach*. Boston: Pitman.
Friedman, M. 1970. "The Social Responsibility of Business is to Increase its Profits." The New York Times Magazine. Retrieved from https://www.nytimes.com/.
Gabriel, I. 2020. "Artificial Intelligence, Values, and Alignment." *Minds and Machines* 30 (3): 411–37.
Gebru, T. 2020. "Race and Gender." In *The Oxford Handbook of Ethics of AI*, edited by M. D. Dubber, F. Pasquale, and S. Das, 253–69. Oxford: Oxford University Press.
Gorwa, R., R. Binns, and C. Katzenbach. 2020. "Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance." *Big Data & Society* 7 (1): 1–15.
Gosseries, A. 2017. "Publicity." In *Stanford Encyclopedia of Philosophy*. Stanford: Stanford University. https://plato.stanford.edu/entries/publicity/.
Hassabis, D. 2018. "The History, Capabilities and Frontiers of AI." In *You and AI, April 30, 2018, London. Retrieved from*, https://royalsociety.org/science-events-and-lectures/2018/04/you-and-ai-history/.
Heath, J. 2006. "The Benefits of Cooperation." *Philosophy & Public Affairs* 34 (9): 313–51.
Heath, J. 2014. *Morality, Competition and the Firm: The Market Failures Approach to Business Ethics*. Oxford: Oxford University Press.
Hern, A. 2017. "Royal Free Breached UK Data Law in 1.6m Patient Deal with Google's DeepMind." The Guardian. Retrieved from www.theguardian.com.
Hsieh, N. 2009. "Does Global Business Have a Responsibility to Promote Just Institutions?" *Business Ethics Quarterly* 19 (2): 251–73.
Keynes, J. M. (1936) 2001. *The General Theory of Employment, Interest, and Money*. Boston: Houghton Mifflin Harcourt.
Klinova, K. 2020. "What's the Responsibility of the AI Industry in Ensuring that AI Serves to Create an Inclusive Global Economy?" Partnership on AI. Retrieved from https://www.partnershiponai.org/whats-the-responsibility-of-the-ai-industry-in-ensuring-that-ai-serves-to-create-an-inclusive-global-economy/.
Kroll, J. A. 2020. "Accountability in Computer Systems." In *The Oxford Handbook of Ethics of AI*, edited by M. D. Dubber, F. Pasquale, and S. Das, 181–96. Oxford: Oxford University Press.

Lambrecht, M. 2020. ""Free Speech by Design: Algorithmic Protection of Exceptions and Limitations", in the Copyright DSM Directive." *JIPITEC – Journal of Intellectual Property, Information Technology and E-Commerce Law* 11 (1): 68–94.

LeCun, Y. 2020. "Deep Leaning, Neural Networks, and the Future of AI." TED2020. Retrieved from www.ted.com/.

Llansó, E., J. van Hoboken, P. Leerssen, and J. Harambam. 2020. *Artificial Intelligence, Content Moderation, and Freedom of Expression*, 1–30. A working paper of the Transatlantic Working Group on Content Moderation Online and Freedom of Expression.

Manheim, K., and L. Kaplan. 2019. "Artificial Intelligence: Risks to Privacy and Democracy." *Yale Journal of Law and Technology* 21: 106–88.

McCarthy, T. 2020. "Zuckerberg Says Facebook Won't Be 'Arbiters of Truth' after Trump Threat." The Guardian. Retrieved from www.theguardian.com.

Moradi, P., and K. Levy. 2020. "The Future of Work in the Age of AI." In *The Oxford Handbook of Ethics of AI*, edited by M. D. Dubber, F. Pasquale, and S. Das, 271–88. Oxford: Oxford University Press.

Moriarty, J. 2016. "Business Ethics." In *Stanford Encyclopedia of Philosophy*. Stanford: Stanford University. https://plato.stanford.edu/entries/ethics-business/.

Murphy, L. B. 1999. "Institutions and the Demands of Justice." *Philosophy & Public Affairs* 27 (4): 251–91.

Néron, P. 2010. "Business and the Polis: What Does it Mean to See Corporations as Political Actors?" *Journal of Business Ethics* 94: 333–52.

Néron, P. 2016. "Rethinking the Ethics of Corporate Political Activities in a Post-Citizens United Era: Political Equality, Corporate Citizenship, and Market Failures." *Journal of Business Ethics* 136: 715–28.

Nissenbaum, H. 1998. "Protecting Privacy in an Information Age: The Problem of Privacy in Public." *Law and Philosophy* 17 (5/6): 559–06.

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

O'Neill, O. 2001. "Agents of Justice." *Metaphilosophy* 32 (1/2): 180–95.

Porter, T. 2009. "The Division of Moral Labour and the Basic Structure Restriction." *Politics, Philosophy & Economics* 8 (2): 173–99.

Powers, T. M., and J. Ganascia. 2020. "The Ethics of Ethics of AI." In *The Oxford Handbook of Ethics of AI*, edited by M. D. Dubber, F. Pasquale, and S. Das, 27–51. Oxford: Oxford University Press.

Rawls, J. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Rawls, J. 1993. *Political Liberalism*. New York, NY: Columbia University Press.

Rawls, J. 2001. *Justice as Fairness*. Cambridge, MA: Harvard University Press.

Reuters. 2019. "Google Shifted $23bn to Tax Haven Bermuda in 2017, Filing Shows." The Guardian. Retrieved from www.theguardian.com.

Rodrik, D. 2015. "Premature Deindustrialization." NBER Working Paper No. 20935.

Sander, B. 2020. "Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights-Based Approach to Content Moderation." *Fordham International Law Journal* 43 (4): 939–1006.

Scheffler, S., and V. Munoz-Dardé. 2005. "The Division of Moral Labour." In *Proceedings of the Aristotelian Society, Supplementary Volumes*, 79, 229–84. Oxford: Oxford University Press.

Scherer, A. G., and G. Palazzo. 2011. "The New Political Role of Business in a Globalized World: A Review of a New Perspective on CSR and its Implications for the Firm, Governance, and Democracy." *Journal of Management Studies* 48 (4): 899–931.

Simon, H. A. 2000. "Government in Today's World of Organizations and Markets." *PS: Political Science and Politics* 33 (4): 749–56.
Singer, A. 2018. "Justice Failure: Efficiency and Equality in Business Ethics." *Journal of Business Ethics* 149 (1): 97–115.
Singer, A. 2019. *The Forms of the Firm*. Oxford: Oxford University Press.
Schumpeter, J. A. (1950) 2008. *Capitalism, Socialism and Democracy*. New York: Harper Perennial Modern Classics.
Slee, T. 2020. "The Incompatible Incentives of Private-Sector AI." In *The Oxford Handbook of Ethics of AI*, edited by M. D. Dubber, F. Pasquale, and S. Das, 107–23. Oxford: Oxford University Press.
Solove, D. J. 2007. "I've Got Nothing to Hide and Other Misunderstandings of Privacy." *San Diego Law Review* 44: 745–72.
Stiglitz, J. 2018. "The Future of Work." In *You and AI*, September 11, 2018, London. Retrieved from https://royalsociety.org/science-events-and-lectures/2018/09/you-and-ai/.
Valentini, L. 2012. "Ideal vs. Non-ideal Theory: A Conceptual Map." *Philosophy Compass* 7 (9): 654–64.
Vogel, D. 2005. *The Market for Virtue*. Washington: Brookings Institutions.
Waldron, J. 2012. "Political Political Theory: An Oxford Inaugural Lecture." In Public Law & Legal Theory Research Paper Series, Working Paper no. 12–26, New York University.
Weinberg, J. 2009. "Norms and the Agency of Justice." *Analyse & Kritik* 31 (2): 319–38.
Wenar, L. 2017. "John Rawls." In *Stanford Encyclopedia of Philosophy*. Stanford: Stanford University. https://plato.stanford.edu/entries/rawls/.
Yeung, K., A. Howes, and G. Pogrebna. 2020. "AI Governance by Human Rights-Centered Design, Deliberation, and Oversight." In *The Oxford Handbook of Ethics of AI*, edited by M. D. Dubber, F. Pasquale, and S. Das, 77–106. Oxford: Oxford University Press.
Zimmermann, A., E. Di Rosa, and H. Kim. 2020. "Technology Can't Fix Algorithmic Injustice." Boston Review. Retrieved Sept. 5, 2020, from: http://bostonreview.net/science-nature-politics/annette-zimmermann-elena-di-rosa-hochan-kim-technology-cant-fix-algorithmic.