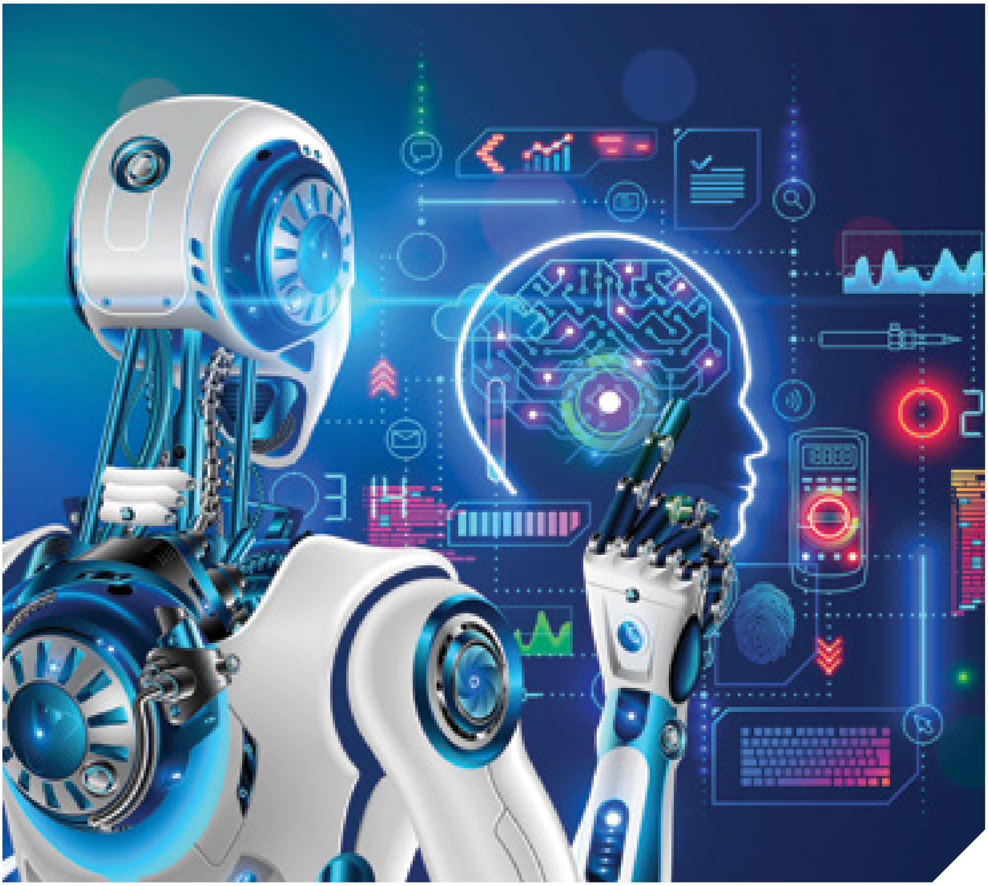

Human, Machine, State

Toward the Regulation of Artificial Intelligence

THE ISRAEL
DEMOCRACY
INSTITUTE

Amir Cahane | Tehilla Shwartz Altshuler

THE ISRAEL DEMOCRACY INSTITUTE

Human, Machine, State

Toward the Regulation of Artificial Intelligence

Amir Cahane | Tehilla Shwartz Altshuler

Series & Cover Design: Studio Alfabees
Typesetting: Nadav Shtechman Polischuk
Chart Graphics: Dana Berger
Printed by Maor Wallach Print

ISBN: 978-965-519-433-3

The Israel Democracy Institute
4 Pinsker St., P.O.B. 4702, Jerusalem 9104602
Tel: (972)-2-5300-800
Website: http://en.idi.org.il
Online Book Store: en.idi.org.il/publications
E-mail: orders@idi.org.il

All IDI publications may be downloaded for free, in full or in part, from our website.

The views expressed in this book do not necessarily reflect those of the Israel Democracy Institute.

Abstract

In recent years, artificial intelligence (AI) systems have increasingly become part of the fabric of daily life. They recommend travel routes and the next song to be played, support medical diagnoses, and, lately, even take an active part in doing homework. Public entities around the world are assimilating algorithmic systems that make, or support, administrative decisions on resource allocation, planning, crime prediction, and protection of the public space—from personal digital assistants to autonomous cars, from robots that carry out simple tasks to monitoring, detection, and forecasting systems.

However, despite their inherent advantages, algorithmic systems may menace human rights and basic freedoms unless there is oversight of their use, development, and deployment. These hazards may emerge at various stages of their development and use—from defining the purpose of the system, via reliance on incomplete, erroneous, corrupted, or biased data, to non-application of post-deployment oversight of the systems' outputs. Furthermore, the more AI develops, the more its systems tend to present new capabilities that were neither intended nor predicted by their developers. Some of these capabilities are inspiring, and therefore called "sparks," but others have the potential to cause harm, such as carrying out offensive cyber actions, manipulating people by discursive means, and disseminating erroneous and misleading artificial

information. Therefore, the ability to identify these capabilities and limit their associated hazards has become a supremely important challenge.

What is artificial intelligence? What are its advantages? And what fears does it evoke, particularly when used by the authorities? These questions are the focus of this book.

Decision-makers, industry, academia, and civil-society organizations in Israel and around the world have all identified AI as a disruptive technology for which national strategy and regulatory policy must be prepared in advance. The end of the previous decade saw the publication of dozens of ethics documents regarding AI which sought to lay down principles for the development, use, and application of algorithmic systems. The ethical values proposed in these documents are founded on several principles: transparency; fairness; damage prevention and safety; responsibility and accountability; privacy; promoting the common good and prioritizing people; and freedom and autonomy.

Ethical values, however, are not enough. To assure the upholding of human rights and basic freedoms, these principles must be moored in legislation. Indeed, we are now seeing around the world initial signs of legislation that seeks to regulate the use and development of AI technologies. Some follow an across-the-board legislative model applied to AI systems at large, as in the European AI Act; others are regulatory patchworks that target specific sectors and uses of AI, such as laws that aim specifically to cope with algorithmic discrimination in systems that hire and promote workers.

This book offers guiding principles for the creation of rights-oriented AI policy and of a toolbox for AI regulation in Israel.

Main Recommendations

Guidelines for the creation of rights-oriented AI policy

Prioritizing people. The main purpose of developing artificial intelligence and learning systems should be serving the human race—individually and collectively—in a way that enhances its welfare. This principle of human freedom and autonomy connects with the principles of promoting the common good and preventing harm, and underscores all the more the importance of the principle of human-centering, freedom, and autonomy. What this anthropocentric outlook means in practice is that the development, deployment, and use of intelligent systems should be based on an approach that considers the defense of basic rights and civil liberties a paramount principle, rather than simply paying lip service to them while actually giving preference to principles such as "promoting innovation," pursuing economic interests such as the advancement of high-tech industry, "making Israel a global technological leader," or even making public-sector processes more efficient.

Prioritizing democracy. AI-based systems have a great deal of potential to infringe on democracy in its broad sense by affecting public discourse and circulating ideas; serving as an instrument of control, surveillance, and policing; and sowing doubt and subverting the very ability to determine reality and distinguish between original and counterfeit and between truth and falsehood. Therefore, the principle of "democracy at the center" should be given much weight, even if this sometimes comes at the expense of technological progress and innovation.

Digital literacy among decision-makers. Digital literacy means the ability to analyze the market and understand the direction in which technology is developing, at least in the near term. For example: Where do the tech giants keep their research and development funds? What patents have they registered in order to secure new technological developments? It also includes understanding commercial and regulatory possibilities for guiding technological development and, by extension, understanding policymakers' responsibility to influence technological development and not merely observe it from the sidelines. A liminal stratum is needed between understanding technology and making technology policy—a framework for understanding the implications of technological systems, being able to imagine the new possibilities that they offer, and gauging their implications for social ethics and the contours of the judicial method. The frequent lack of framework often results in lacunae in understanding, particularly in matters that have broad implications such as AI.

Reconceptualizing systems and capabilities in the AI field. The concept of artificial intelligence is a highly powerful politico-technological metaphor. Comparing AI systems to the human brain creates proximity and similarity, leading to the social assimilation of the idea that machines operate like the human brain, perform human actions the same way people perform them, and, in fact, compete with people. We recommend that machines' actions and the traits attributed to them be conceptualized in a manner that is not contingent on this comparison.

Developing rights for the objects of AI decisions. Basic rights have to be rethought in two senses. First, the constitutional theory of existing human rights, such as freedom of speech and the right to privacy, needs an injection of new meanings. And second, new digital rights, unneeded in the past, should be created: foremost the rights of individuals who come into contact with algorithm-based machine systems.

Israel must not become a "digital backyard." In recent years, proposals for AI regulation have been advanced in leading countries and the European Union, in what is presumably the onset of a global trend. Even if there are values-based differences among regulatory mechanisms in different places—whether in the choice of systems defined as dangerous or in the

declared goals of the legislation itself—they also have much in common. Therefore, a situation should not be allowed to develop in which Israel lacks legislation that is well aligned with accepted legislation abroad. Admittedly, a lower regulatory standard than the convention abroad may stimulate innovation, but not necessarily of the desired kind; it may turn Israel into a technological backyard, a place where systems are created whose development, dissemination, or use are banned in other western countries.

Future-proof regulation. Technology is advancing rapidly; in countries such as Israel, where legislative processes are very slow, an especially wide gap is being created. Therefore, regulation should not target any specific technology, as that is a sure prescription for regulatory obsolescence. Instead, an effort should be made to establish guiding principles and general definitions that would invest future enforcement with flexibility.

A hybrid regulatory framework: principles, rights, and legislation. Principle-based frameworks present a matrix of ethical core principles; rights-based frameworks focus on protecting the human rights and liberties of those affected by AI-based technological applications; and legislation-based frameworks make it possible not to rely solely on voluntary regulation predicated on the goodwill of economic actors. The three frameworks do not clash with each other; they should be integrated into a hybrid structure that combines soft regulation (ethical principles) with a risk-management approach manifested in rigid legislative provisions and regulatory rules.

Flexibility in the timing of regulatory intervention. In certain cases, there are clear advantages to early regulation, introduced before products based on a certain technology enter the market. If a technology is perceived as especially dangerous—physically (as in autonomous cars) or ethically (such as artificial creation of content that incites to terrorism)—

then advance regulation makes sense. Even in less extreme cases, prior intervention may be useful in shaping the directions of research and in planning resource investment in development. Furthermore, since investment is relatively small and sunk costs are smaller at this stage, regulatory intervention may meet more limited resistance from interested parties. Contrastingly, in certain cases it may be better to wait and cope with problems when they arise instead of trying to anticipate them.

Sectorial vs. across-the-board regulation. Across-the-board regulation attains goals of governance, creates regulatory harmony, and, accordingly, may enhance public trust and increase regulatory certainty for industry. Sectorial regulation, conversely, allows the use of existing regulators and their enforcement powers, does not require the establishment of new institutional frameworks, facilitates more accurate tailoring of enforcement arrangements and methods to a given industry, enhances regulatory clarity and certainty, and also allows stakeholders in each sector to participate in formulating these arrangements. The problem with sectorial regulation, however, is that it may result in discrepancies among industries, create inconsistent standards, exacerbate gaps among regulators, and leave behind unregulated spheres that fall into the cracks. Therefore, we recommend a combination of across-the-board and sectorial regulation, such as having a general regulator with instructional and advisory powers vis-à-vis sectorial regulators.

The AI regulation toolbox

Regulation of learning systems should be based on an understanding of their "lifecycle." To create effective regulation of learning systems, all components of their lifecycle should be taken into account. Given that the principles of regulation—such as fairness, privacy, transparency, accountability, and risk management—are manifested in different

contexts in each component of the lifecycle, inattention to these components may result in excessive regulation of certain elements and disregard of others, compromising regulatory effectiveness.

An integrated approach, by contrast, gives consideration to the full set of lifecycle components and their interrelations. The purpose of a learning system and the framing of the problem it addresses, for example, should influence the choice of model used (whether or not to allow the choice of a more opaque model in terms of the ways it makes decisions). Evaluation outcomes at the stage at which the model is built energize risk-assessment processes—which, in turn, require decisions on how the model should be trained, deployed, and protected. The purpose of the model affects the choice of user interfaces: Should a model that dispenses medical advice notify users that it may be wrong? Should users be told that they are connecting with an artificial system and not with a human one?

An important part of understanding the lifecycle of learning systems relates to the need to monitor them after they are deployed in the real world (e.g., when systems are integrated into a product or an interface). This is because a learning system—unlike other products, such as pharmaceuticals—can, by its very nature, change even after it is applied due to the feedback-loop that it receives from its users.

Development of risk-management methodologies. The application of risk-management methodologies to algorithmic systems, despite being vitally needed, is still in its infancy. We propose a risk-management model that requires a double observation in order to assess the dangerousness of a system, first to assess the potential dangerousness of the system as designed and then to assess the strength of the alignment of its task with its outcome, namely, the potential of a system to manifest its dangerousness outside the role its designers intended for it.

This double layer assessment should be the basis for decision-making. To make it practicable, rules of governance and safety need to be formulated, relating *inter alia* to responsible training (whether to train a new model that exhibits early indications of danger—and in what way) and responsible application (whether, when, and how to implement models that may be dangerous); requisite levels of transparency and documentation in cases of models that may pose extreme danger; and the auditing and cyber-security systems that should be applied to them.

Data documentation, data governance, and post-deployment auditing procedures. The complexity of AI systems generally, and of learning systems particularly, imposes special difficulties on policymakers and regulators when it comes to formulating rules of liability and identifying the actual chain of causation that precipitates an infringement of rights, particularly in consideration of variances among sectors and among applications.

What all these have in common is that they are impossible without proper documentation. The basis for every factual examination of a concrete failure in AI systems is data governability and painstaking documentation of working procedures, information sources, labeling, models, coding processes, risk assessment and databases, and detection of discrepancies in each. Good documentation design also serves the interests of entrepreneurs and developers because it allows them to investigate failures and unexpected phenomena after their occurrence, and also to satisfy regulatory documentation obligations that originate elsewhere.

Development of tools to contend with biases and "fairness engineering." Although the algorithmic-bias problem defies prevention, particularly when the bias originates in reality itself as embodied in data, it may be identified and mitigated. Several strategies to mitigate algorithmic biases should be pursued, including having in place statistically fair procedures, diversifying human capital among developers of AI systems, and applying after-the-fact auditing procedures.

Coping with the challenges of algorithmic transparency: We suggest a model that is based on the classic conceptualization of transparency, but includes an alternative tailored to the technological limitations of algorithmic systems that cannot provide explanability for specific outputs. The model is meant for cases in which an output cannot be provided but society needs to see the obligation of transparency upheld so as not to thwart the development and use of certain technologies.

Recommendations for an AI regulatory institution in Israel

We recommend establishing an AI regulation authority in Israel for the purpose of advancing and systematizing the regulation of intelligent systems in the country, including the formulation of across-the-board legislation with an eye on corresponding developments abroad. This authority should make policy on the development, implementation, and use of AI-based products in Israel; provide sectorial regulators with professional guidance in order to assure the consistency of the rules that apply to the development, deployment, and use of these systems; and serve as a residual guiding entity, that is, be responsible for regulating those products in fields where there is no sectorial regulator.

The regulator should also provide state authorities with professional AI guidance. Among other things, it should express its professional opinion about government tender documents pertaining to artificial intelligence in order to make sure that the state procures AI systems that meet Israeli and foreign standards in the field, and should give the private market “soft” incentives to align its work with standards.

The regulator should also be tasked with consultation in matters of AI law and should be empowered to present the Knesset and the courts with its own position in these fields. In addition to the aforementioned roles, the proposed regulator should be a source of knowledge, instruction, and cooperation and should promote digital literacy among government players in these regards. It should also discuss the possible social effects of these technologies with local and international stakeholders in industry, government, and academia.

At the present time—at least for the next few years, until the field stabilizes—the AI regulator should be established in the form of a unit within the Regulatory Authority, because the latter's roles are highly suited to the evolving world of AI regulation. To cope with the challenges of flexibility, the burdened regulatory environment, and the technology, enough resources should be allocated to allow positions to be created for experts in technical fields and also in law and policy. Even in the absence of a comprehensive artificial intelligence law, the Authority should be budgeted by government resolution so that it can set up the AI regulator's unit.

Recommendations for the interim period until AI regulation is introduced

Supplementary legislation and legislatory amendments. Right now, even in the absence of a broad and dedicated artificial intelligence law, designated decision-makers and regulators should be obligated to update existing legislation and pass supplementary legislation—mainly to statutes such as the Competition and Consumer Protection Law, the Copyright Law, the Protection of Privacy Law, the Evidence Ordinance, and the Government Procurement Law.

Create a "pre-regulatory ecosystem." A "pre-regulatory ecosystem" should be created, in which the designated regulator, sectorial regulators, and supplemental regulators (such as the Protection of Privacy Authority and the Competition Authority) would issue guidelines and professional opinions while serving industry, its players, and the courts. These guideline documents would help industry set self-regulation standards on the assumption that future regulation would resemble the guidelines. They would also serve standards-setting companies in constructing their standards frameworks.

Under the influence of this pre-regulatory ecosystem, auditing patterns in various contexts would undergo fine-tuning, thinking about due caution standards would take place, an effort to encourage responsibility in developing and implementing learning systems would be made, sandboxes would be built, various regulatory trials would be undertaken, and a pool of experts who could serve industry, the regulators, and the courts in coping with new and complex issues would take shape. Concurrently, industry and its actors would be able to provide the various regulatory bodies with feedback, thus improving the guidelines. In addition, in the absence of regulation, the courts could use the guidelines for inspiration as they interpret conflicts presented to them. Their rulings would also enrich the body of knowledge and enable players and regulators alike to enhance and polish their guidelines until regulation can be formulated.

www.idi.org.il