Regulating Autonomy: a Conceptual Introduction

Daniel López-Castro (corresponding), Pablo Jiménez-Schlegl (corresponding), Manuel Cebral-Loureda (corresponding)

These authors contributed equally to this work.

Daniel López Castro [1], [2]
[1] Instituto de Filosofía (IFS), CSIC, Madrid, España.
daniel.lopez@csic.es
[2] Universidad Nacional de Educación a Distancia (UNED), Madrid, Spain.
dlopez948@alumno.uned.es

Pablo Jiménez-Schlegl
Institut de Robòtica i Informàtica Industrial (CSIC-UPC), Barcelona, Spain
ORCID: 0000-0003-3627-4938
pjimenez@iri.upc.edu,

Manuel Cebral-Loureda
School of Humanities and Education. Tecnológico de Monterrey, Mexico.
ORCID: 0000-0001-6359-2427
manuel.cebral@tec.mx

Keywords: AI Ethics, Artificial Intelligence, Values, Agency, Autonomy, Governance, Regulation

1. Introduction

The coming years will see the deployment of increasingly sophisticated and capable artificial intelligence (AI). This may imply an increase in the autonomy of artificial systems that raises questions about how to integrate them into our societies in an appropriate manner. The premise of this book is analogous to Gartner's Hype Cycle. We are prone to overrate the immediate advancements of AI, yet conversely, we often undervalue the long-term impact of its incremental implementation. The deep and transformative digitalization of our societies brings with it a multitude of challenges that extend beyond the technical realm to encompass key ethical, political, and social dimensions. The delegation of increasingly important tasks to autonomous systems presents a major challenge for humanity, one that directly threatens the autonomy of its citizens. For this reason, the development and regulation of AI technologies must involve all relevant stakeholders, not only large corporations and governments, but also the scientific community and the public. Only in this way can we mitigate the risks posed by these increasingly powerful AI systems while avoiding the opportunity cost of not using them.

The term "regulation" does not necessarily imply the creation of a law to be imposed. Regulation is a wider concept. It entails the establishment of a more expansive intermediate domain for reflection, experimentation, and implementation. This space of regulation is where ethical perspectives and reflection on values converge, together with the analysis of how autonomous AI systems affect and intervene in governance systems. We cannot forget that it is not only the law that regulates technology. On the contrary, technology itself—through the architectures designed by its promoters—plays a significant role in regulating our societal structures and the manner in which we relate to others and to ourselves.[1]

The protagonists of this book are AI systems. The chapters will explore various facets of these systems, reflecting the fact that there is no single class of AI, but rather constellations of them. The expression *artificial intelligence* functions as an umbrella concept that encompasses a plurality of techniques and aspirations. Throughout its history, different approaches to AI—such as the symbolic and sub-symbolic (or connectionist) paradigms—have each predominated for a time, producing promising results. In the last few years, machine learning in general and LLMs in particular have once again placed AI at the center of public debate. These approaches coexist and will continue to do so, often leading to the development of hybrid or multi-agent systems. The contributors to this volume will, in their analyses, inevitably draw upon their experiences with particular types of AI systems. But all have in common, according to OECD,[2] that:

[1] See, e.g., Lessig (1999) on the regulatory power of code, and Floridi (2014) on how the infosphere shapes our reality.
[2] Grobelnik et al. (2024).

> [...] they are machine-based systems that, for explicit or implicit objectives, infer, from the input they receive, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy [...].

Autonomous robots deserve special mention, as the physical embodiment of AI, capable of perceiving and acting on the real world. In their case, decisions taken by the AI translate directly into actions.

2. Autonomy in focus: machine, human and hybrid

The core concept in this volume is autonomy. According to the OECD's definition:

> AI system autonomy means the degree to which a system can learn or act without human involvement following the delegation of autonomy and process automation by humans. Human supervision can occur at any stage of an AI system's life cycle, such as during AI system design, data collection and processing, development, verification, validation, deployment, or operation and monitoring. Some AI systems can generate outputs without specific instructions from a human.

Regulation must address the critical questions of what skills can be delegated to AI systems, under what conditions, and to what extent.

2.1 Machine Autonomy

To further delimit the meaning of autonomy when speaking about machines, it is often useful to contrast this term with the notion of automation. Both terms refer to machines that are able to move without constant and direct involvement of a human operator. Taking the usual meaning of automatic, "capable of operating without external control or intervention,"[3] all the autonomous machines are clearly automatic, whereas the inverse is not necessarily true. Etymology may be quite insightful in this case: automatic derives from the Greek word *αὐτόματος* (autómatos, "self-moving, moving of oneself, self-acting, spontaneous"), which is composed by *αὐτός* (autós, "self, myself") and *μέμαα* (mémaa, "to wish eagerly, strive, yearn, desire")[4]. On the other hand, autonomous comes from *αὐτόνομος* (autónomos, "living under one's own laws, independent"), where the second part of the word refers to "a system of rules or laws about a particular field".[5] This is implicitly stating that automatic machines move following their very nature (their "desires"), they cannot do differently, whereas autonomous machines follow "their own rules or laws" which both points at independence and at some kind of higher cognitive process. This contrast is even more stressed by the second meaning of automatic, namely "done out of habit or without conscious thought"[6].

[3] Wiktionary, "Automatic" (2025).
[4] Ibid.
[5] Wiktionary, "Autonomous" (2025). As for the etymology, the term is redirected to "Autonomy".
[6] Wiktionary, "Automatic" (2025).

Consider a conventional industrial robot, working in a protected cell, in a controlled environment. This robot follows the routine specified in its program, including motions along given trajectories and actions such as opening and closing the gripper in predefined positions. The robot may even be able to react to environmental circumstances, such as the presence or absence of a workpiece, by closing the gripper around it in the first case or waiting in the second. It certainly follows a rule, but this rule has been provided by a programmer, who has foreseen the possible occurrence of these events. This robot is not able to modify this rule by itself, or to add others, by learning from failures due to modified circumstances such as slight misplacements of the workpiece or the presence of two of them instead of one. It is possible to design the robot cell in a way that such events may simply not happen. The robot behaves in this case just as an automatic machine.

But the picture is entirely different when the robot has to perform its activity in a natural or a human-populated environment (be it domestic, urban, a hospital, a museum…). The robot will have to be mobile, in the first place, instead of fixed as the industrial robot. Uncertainty is consubstantial to such environments, as they contain too many possible events for a programmer to foresee them all. Purely reactive behaviors may still be successful for simple tasks such as avoiding an obstacle along the trajectory, but the robot will have to be equipped with enhanced sensory devices. This includes interoceptive ones, such as a battery state measurement unit, that tells the robot when to turn back to the recharging station. This means that the robot has to be able to localize itself and to find its way to its destination. We are describing some of the traits of the well-known Roomba® vacuum cleaning robot, and we can consider it an autonomous machine. This illustrates two important aspects: 1) automatic vs. autonomous is not just reactive vs. deliberative, and 2) autonomy occurs in different levels or degrees, as explained next.

A collection of cleverly designed reactive behaviors based on perception, embedded within an overall strategy (which in the case of the first vacuum cleaning robots was to randomly move around) and self-sustainment (going to the docking station by itself), are enough to define an autonomous machine in its most basic expression. More recent robots are able to build up a model of their environment and to travel across more efficiently while cleaning. Learning is a skill that enables higher degrees of autonomy, as it allows the machine to derive new rules governing its behavior. But autonomy is not fully exercised unless the machine is also able to make its own decisions, on a deliberative basis, which implies some kind of reasoning. This is even more so if the machine is capable of anticipating a whole course of action via task planning. In sum, higher cognitive functions are *conditio sine qua non* for higher degrees of autonomy. This is a necessary but not sufficient condition: in the case of a physical machine such as a robot, it has to be constructively enabled to act in the world (e.g., have arms and grippers for manipulation, means of displacement, etc.), and also to have the freedom to act.

What we have illustrated via the robot example applies for non-embodied AI as well. An accounting program will behave automatically with the fed data, whereas an AI-based stock exchange recommender system will be able to learn from previous experiences, and modify

its behavior accordingly. If allowed to operate in a stock market, it becomes truly autonomous. This leads towards the next question: to which extent will we allow AI systems to be autonomous, in exercising their autonomy in the real (physical or symbolic) world? In the next section, we will present different ways humans may supervise and control (allow or inhibit) the autonomous actions of an AI system. To this end, we have to be well aware that AI systems autonomy will possibly conflict with human autonomy. What has been conceived as human augmentation may become, via delegation of skills and/or decision making, human limitation.

2.2 Human Autonomy

In recent years, numerous organizations, including business, governments, and civil society, have convincingly advocated for the importance of placing the human being at the center of AI systems design (see, for example, human-centered AI). In this same vein, we assert that any framework for regulating the autonomy of AI and robotics must be grounded in a robust understanding of human autonomy.

The concept of autonomy plays a fundamental role in various fields of social and human sciences. Directly or indirectly, it is a central theme in discussions regarding freedom, responsibility, sovereignty, human rights, and the nature of human agency. All of these issues are crucial when considering how we relate to artificial systems, both at the individual and collective levels. While it has evolved significantly over time, it is rooted in the foundations of Western ethical and political thought.

Initially, the term “autonomy” was used in ancient Greece to refer to the self-governance of city-states. While Greek philosophers explored related topics (e.g., Plato or Aristotle), the term autonomy as such was not explicitly attributed to individuals, as long as they were firstly understood as citizens, so as part of a bigger entity. Autonomy, understood as a virtue related with the individual capacity to govern oneself, will not be widely regarded as a fundamental moral and political value until modernity.

In this context, one of the most important theoretical developments of the concept of autonomy can be found in the work of Kant. For the German philosopher, this concept serves as the fundamental principle that shapes all morality. Autonomy does not imply following external rules, or doing so to seek rewards, fulfill desires, or avoid punishments (heteronomy); rather, it involves acting based on principles that one has rationally determined for oneself. This self-imposition of a universal moral law would be the foundation of all moral obligation. This conception of autonomy (which we can understand as *moral autonomy*) implies recognizing each individual’s capacity to govern themselves. Therefore, the concept of autonomy in Kantian ethics not only serves as a principle for determining our own moral obligations but also reinforces the idea that all individuals, by virtue of their autonomy, deserve moral consideration.

Most contemporary theories of autonomy are somewhat neutral with respect to this universal moral law. These conceptions, which we can frame within the label *personal or individual*

autonomy, could be summarized as saying that to be autonomous is "to be directed by considerations, desires, conditions, and characteristics that are not simply imposed externally upon one, but are part of what can somehow be considered one's authentic self".[7] To illustrate this conception, consider the influential work of Gerald Dworkin.[8] The American philosopher seeks to describe how individuals exercise their autonomy. According to his hierarchical model, people have a series of commitments (interests, values, beliefs, desires, etc.) that they are capable of evaluating through reflexive processes. On the one hand, there would be first-order desires (which have the agent's actions as their object) and, on the other hand, second-order desires (which have first-order desires as their object). A person acts autonomously when, after critical reflection, she identifies with a first-order desire and acts upon that endorsement.

The concept of autonomy has been challenged, particularly from a feminist perspective, for centering or even promoting a harmful individualistic ideal that disregards the importance of social relationships and cultural patterns in the constitution of the self. These critiques have given rise to a fruitful conception of autonomy that can be framed under the label of *relational autonomy*. According to this perspective, the liberal ideal of the self-made man is incomplete. On the contrary, human beings develop their most essential characteristics through interactions with others and with their environment. Thus, a relational conception of autonomy can direct our attention to those relationships that enhance or diminish our agency.

Despite being a complex concept, replete with nuances and often misunderstood, autonomy, in its various forms, plays a fundamental role in our lives and institutions. Particularly in its normative aspect, autonomy and other closely linked concepts (freedom, dignity, responsibility) act as firewalls against interventions generally deemed undesirable, such as oppression or paternalism. Moreover, even considering its critiques, this rich body of reflection provides the essential starting point for regulating intelligent systems whose artificial autonomy may conflict with our own. The challenge, therefore, is to ensure these systems are designed not merely to be functional, but to genuinely enhance, rather than diminish, human autonomy.

2.3 Human-Machine Autonomy

Just as the concept of relational autonomy invites us to cease thinking of individuals as isolated from their context, in this book we do not consider AI systems in isolation, but forming part of a *hybrid* system, composed of artificial agents and human beings. In other words, AI systems are neither developed nor deployed in a vacuum; they operate in conjunction with human individuals or collectives. Therefore, in this hybrid context, autonomy is not merely an attribute of the AI system. It also refers to the human users' freedom and their capacity to engage in the coordination of hybrid systems (with human and non-human agents), as may appear in shared decision-making, assistive use of technology, or technology understood as a way of augmenting human capabilities.

[7] See Christman (2020).
[8] See Dworkin (1988).

A behavior of an agent can be conceived as a continuous loop or cycle involving three activities: perception, cognition, action. The world[9] state, which includes both the environment and the agent's own state, is perceived and interpreted. This perception enters a cognitive process where, according to a certain purpose or goal, a suitable response is elaborated, which translates into an action. This action modifies the world in order to bring its state closer to the goal. This is checked again via perception and the loop goes on. If we are speaking of reactive behaviors, the cognitive part is skipped (or, if you prefer, hardwired), perceptions are directly linked to actions, whereas in deliberative behaviors, cognition involves reasoning and planning (processing in the case of an AI system, and thinking in the case of a human). In hybrid systems where the AI part takes over the execution of actions, the involvement of the human in the AI system's loop belongs to one of the following three categories, which we illustrate with the example of an autonomous robotic weapon, an Unmanned Aerial Vehicle (UAV) equipped with a camera and missiles:

- Human-in-the-loop: The human is constantly monitoring and even operating some functions of the AI system. Except for routine tasks, the AI system is not autonomous. In the case of the UAV, the human teleoperates the machine in take-off and landing, fixes the objective and provides the firing command, the UAV just flights in autopilot mode during the approach to the objective's position and return to the base.
- Human-on-the-loop: The human is in constant surveillance of the AI system, which works autonomously, and has to validate crucial decisions. In the UAV example, the human operator validates or not the identified objective and /or provides the firing command.
- Human-out-of-the-loop: The AI system operates fully autonomously, the human is a mere spectator in the best case. In our example, this means that the weaponized UAV, once deployed, does everything on its own: flying, selecting the objective, firing.

These levels of increasing autonomy of the AI system come together with an increasing delegation of decision-making to the machine. However, this should not mean a delegation of responsibility as well. While there is a growing consensus that artificial systems can possess agency—the capacity to act causally in the world—attributing autonomy to them in the same rich, normative sense we attribute it to human beings is a far more controversial, and for many, dangerous assertion. This is not a trivial matter, as in an interdisciplinary environment, it is very likely that a lawyer, an engineer, and a philosopher are talking about different things when they discuss autonomy.

In this book, we shall explore this expanded understanding of autonomy and regulation through the following sections.

[9] This term is used in a conventional fashion and refers just to what can be perceived by the agent and on which its actions have an immediate effect, it does not clearly refer to the whole planet.

3. Structure of the volume

3.1 The ethical dimension: foundations, education and social implications

Ethics has emerged as a prominent topic within the study of AI. This trend has been accompanied by the publication of numerous corporate and institutional documents outlining principles and theoretical frameworks, sometimes contradictory, which have sown confusion among stakeholders. The concept of autonomy, in its various forms, is essential for a nuanced understanding of the implications of the gradual implementation of artificial autonomous systems in our societies. Moreover, ethics examines our actions, which are influenced by the use of technologies, so the discipline itself is forced to evolve.

In this first section devoted to ethics and autonomy, fundamental questions on this topic are revised in the first two chapters. Manuel Cebral-Loureda, in **Chapter 2, "The Ontological Freedom of AI (as a Limit to its Regulation),"** explores the consequences of a system being intelligent and how its very conceptualization forces us to recognize its ontological freedom. Based on this idea, a system cannot be regulated purely externally, at least not in a deterministic way that reduces its ability to think autonomously. All of this redefines the position of human beings in relation to technology, forcing us to set aside anthropocentrism, which understands machines as being absolutely subordinate to human beings. The recognition of AI ontological freedom is against.

On the other hand, Florian Richter honors the title of his **Chapter 3 "The Concept of Co-Action: Its Roots and Its Place in the Philosophy of Technology"** by exploring the foundations of this notion and exploiting its significance as for responsibility attribution in hybrid human-AI systems. The role of autonomous AI in education and its ethical implications is analyzed in the next three chapters. It should be stressed that the educational sphere is far from being just one more application domain of AI, but a crucial one, as it shapes the stance and attitudes of the human actors of tomorrow on AI. In particular, as for shared (or not) autonomy.

Gerardo Castañeda points at the potential dangers in child development, from a developmental psychology point of view, in **Chapter 4 "'Am I or AI, that is the question': the psychological development experiment of 21st century"**. From prenatal and infancy until adolescence periods of the "Generation AI" are scrutinized as for the risks posed by AI, bearing recognized children's rights in mind.

Chapter 5 “Ethical Dilemmas in Educational Interventions” by Hector G. Ceballos and Juan Andrés Talamás, advocates for maintaining the “human-in-the-loop” of education, the involvement of human experts in educational settings that use AI tools, to avoid possible biases while at the same time respecting students' rights, freedom, and dignity. A critical issue is that the educational and other personnel involved in the use of this technology are sufficiently knowledgeable about their functioning.

A broader view is taken in the last chapter of this section, **Chapter 6 “On the Social Consequences of Generative Artificial Intelligence”**, where Sabur Butt and Hector G. Ceballos review the consequences, both positive but also negative and/or dangerous, of the adoption of Generative AI tools in the domains of trust, economic structures, education, creativity, parenting, mental health, social behavior, climate change, the global economy, and national security.

3.2 The agential dimension: embodied AI and assistive settings

Robots are programmable machines, and as such, able to incorporate the cognitive skills AI has to offer. The other way round, robots can be considered embodied AI, that is, they provide the possibility to AI to have a physical presence in the world: to gather information from the surroundings directly through their sensors (including computer vision) and to act on the environment. Robots will obviously participate from the growing presence of AI in human society. They are the bodies that will bring human interaction with AI to a new level. In this context, their appearance, the external aspect of these bodies, is anything but irrelevant, as it may condition human expectations beyond the actual behavior of the robot.

This is the topic addressed by Pablo Jiménez-Schlegl in **Chapter 7, “Embodiment and Appearance: Robot Bodies and Human Responses”**, by reviewing the different shapes taken by fictional and real robots in recent history, systematically dissecting and analyzing the constituent features of these embodiments, and finally discussing their effects on human acceptance, trustworthiness, engagement, as well as the ethical risks linked to robot appearance. This is especially sensitive in the assistive robotics domain, because of the vulnerability of the end-users. However, in these settings, actual robot autonomy is a far more relevant factor than the perceived autonomy.

Belén Liedo delves deeply into this issue in **Chapter 8 “Ethics of Care and Assistive Robotics: How Should Robotic Care Look Like?”** and presents a relational autonomy framework for the ethics of care that builds on coordination of the assistive robot with other (human) caregivers. Relational autonomy is a fundamental concept in the assistive robotics scenario, which can be framed within a wider discussion on robot autonomy.

This broader picture is tackled in Pablo Jiménez-Schlegl, Júlia Pareto, and Carme Torras' **Chapter 9 "Robot ethics: Assessing risk from the perspective of autonomy"**. Robots are characterized by the type of control that governs their performance, which is the premise to the kind and degree of autonomy they can achieve. Human perception of this quantification of robot autonomy and its relation to ethical risks levels are other topics addressed in this chapter, with examples in social assistive robotics.

The healthcare scenario is approached in **Chapter 10**, Aníbal M. Astobiza, Ramón Ortega Lozano, Marcos Alonso and Jorge Linares' **"AI agents for health. Agency and autonomy in AI-driven healthcare"**. The chapter focuses on clinical decision-making by autonomous AI-agents in human-AI collaborative settings, with special regards on accountability and patient autonomy.

3.3 The axiological dimension: embedding values into AI

In recent years, the notion of "embedding" human values into machines has been gaining traction. Whether through social science methodologies like Value Sensitive Design or more technical approaches such as AI alignment, the concept of value seems poised to play a pivotal role in the design and regulation of emerging technologies. The idea of value-based design or alignment can easily be interpreted in terms of control and is thus closely tied to the tension between human and artificial autonomy. However, it remains unclear what we mean when we speak of human values, particularly in the context of technology. This section addresses these issues theoretically, through the examination of methodological and conceptual frameworks, and practically, through case studies.

The section opens by interrogating the very term at the heart of the debate. In **Chapter 11, "Exploring 'Value Alignment',"** Daniel López-Castro argues that the concept, often presented as a technical problem, is also and fundamentally a site of conceptual struggle. Through a genealogical analysis that traces the term from transhumanist forums to corporate safety teams, the chapter reveals three distinct and competing conceptions of value alignment, each serving a different function: narrative, instrumental, or normative.

Chapter 12, "Towards an Axiology of Intelligent Hybrid Systems," then moves from this critical mapping to a constructive theoretical proposal. Mario Toboso-Martín offers an innovative approach to the axiology of technology, arguing that values are not pre-existing entities to be "emdedded," but are instead dynamic "mediations" that emerge from the "imbrication" between the system and its circumstance. The chapter presents a formal model to operationalize this insight, distinguishing between the "estimative" importance of contextual requirements and the "evaluative" realization of a system's qualities.

To connect such foundational theories with the complexities of implementation, **Chapter 13, "Four Settings and a Proposal (For the Exploration of Value Alignment in AI),"** by

Pablo Noriega and Enric Plaza, approaches the value alignment problem from a complementary perspective. The authors deconstruct this challenge into four complementary problem settings: the design of collective coordination systems, the modeling of value-driven agents, the experimental use of policy sandboxes, and the governance of AI "in the wild." These settings form the basis of their proposal: to launch a research program where the systematic study of these paradigmatic problems can ground a new theory of values inspired by and for the design of AI.

Connecting with the relevance of context expressed in Chapters 11 and 12, Nardine Osman et al. show in **Chapter 14**, **"Instilling Organisational Values in Firefighters through Simulation-Based Training"**, how to integrate value-alignment in the evaluation of the trainees' decision-making in an evolving firefighting scenario, via a firefighter simulation-based training framework developed by the authors. AI is used to generate an optimal strategy for each situation, to which the trainees decision-making is compared. As in other educational/formative scenarios, AI contributes in this case to the human trainee's gain of autonomy: it is not just about knowledge acquisition, but about understanding and internalising value-aligned decision-making, which works towards ethical grounded autonomy.

The inquiry into value is then extended beyond a purely anthropocentric frame in **Chapter 15, "What is the Ecological Value of AI Used for Environmental Purposes and to What Extent is It Ethical?"**. Cristian Moyano-Fernández challenges the straightforward narrative of "AI for sustainability," arguing that its promises often obscure a contested understanding of "ecological value". By shifting the analysis to include the autonomy of non-human nature, the chapter reveals the trade-offs of environmental AI, highlighting how these interventions can conflict with core values like ecological integrity and resilience.

3.4 The regulatory dimension: frameworks for governance and law

Although the most widespread theoretical conceptualizations primarily concern autonomy in a more personal or individual sense, there is no doubt that this ideal serves a fundamental function on a social scale. On one hand, autonomy may refer to the ability of groups or communities to maintain their practices, languages, and social structures without external interference. On the other hand, in an increasingly complex global context, national and supranational institutions can play an important role in safeguarding their own autonomy and that of their citizens. In this section, institutional and legislative initiatives aimed at regulating the capabilities of AI systems, as well as the attribution of responsibility will be addressed.

In **Chapter 16**, **"Governance of Artificial Agency and AI Value Chains: A Few Remarks on Autonomy from a Legal and Ethical Approach"**, Pompeu Casanovas and Pablo Noriega examine the concepts of legal autonomy, legal governance, legal agency, moral agency, delegated agency, harm, and liability with a focus on AI agents, under the lens of

Civil Law, Common Law and Ethics, not in isolation but as an interconnected network. This framework allows to elucidate the functional meaning of the basic concept of autonomy from an ethical and legal standpoint, and, complementarily to Chapter 8, to identify different autonomy levels for AI systems to frame the associated risks and the ethical and legal tools that can be used to confront them on a regulatory basis.

In **Chapter 17, "The European Union Artificial Intelligence Act: Ethical Principles and the Regulation of AI for Social Welfare and Development,"** Barcenas and Sabag examine the evolution of the European Union's regulatory approach to AI and highlights the social and ethical challenges of rapidly advancing AI technologies. It explores key ethical principles such as justice, transparency, autonomy, and responsibility, and discusses how these principles can be incorporated into binding legal frameworks. The authors assess the technical, legal, and geopolitical challenges of AI governance and analyze how to reconcile innovation with fundamental rights through constructive dialogue in a globalized context.

Putting these broad regulatory principles to their ultimate test in a domain where consensus remains elusive, **Chapter 18, "Repurposing Lethal Autonomous Weapons Systems for Disarmament,"** has Sara Degli-Esposti intervening in the deadlocked debate on LAWS by arguing that its narrow focus on kinetic lethality is disconnected from the reality of hybrid warfare. The chapter's central proposal is to reconceptualize these systems by integrating Cyber Weapons Systems (CWS), thereby acknowledging that political goals are also achieved through other means. This crucial shift in definition breaks the current impasse by opening a novel conceptual path: repurposing these dual-use technologies away from purely offensive capabilities and towards applications that reinforce international peace and security.

Morales-Tiburcio et al. propose in **Chapter 19, "Advancing Responsible AI: An Analysis of Guidelines, Regulations, and Standards for Trustworthy and Responsible AI,"** that Implementing Responsible Artificial Intelligence (RAI) is crucial for ensuring AI technologies serve the public good. The authors analyze the evolution and application of RAI principles among various stakeholders presenting a visual framework based on a comparative analysis of recent documents. It underscores the practical challenges of applying ethical principles and emphasizes the need for context, prioritization, and integration throughout the AI lifecycle, providing a foundational guide for developers, policymakers, and researchers.

In **Chapter 20, "AI and Financial Inclusion in Mexico: A Challenging Relation,"** Ibarra-López and Moreno-Alvarado examine how financial inclusion in Mexico has interacted with AI, discussing its contribution to the development of financial operations and the potential disadvantages for consumers and traditional providers. The study focuses on Fintech companies and the implications of AI in financial processes. It explores regulatory challenges, particularly the current state of the Fintech Law and ways to enhance it, with a focus on anti-discrimination efforts.

The book's four sections guide the reader through a complete analysis built upon interconnected dimensions. The ethical dimension uncovers the core problems and

implications; the agential dimension examines where these problems take physical and interactive form; the axiological dimension provides the tools for a value-based analysis; and the regulatory dimension seeks comprehensive societal solutions. The original approach of this book consists in this multi-disciplinary view on AI-Human autonomy, structured around them.

Acknowledgements

This work was partially funded by the project the bilateral action BILTC22005 between the Tec (Tecnológico de Monterrey) and the CSIC (Spanish Research Council) named “Desarrollo de una estrategia de formación y seguimiento de los desafíos éticos del uso de la inteligencia artificial en una perspectiva 2030” (DesafIA2030) and MCIN/AEI/10.13039/501100011033 and the European Union NextGenerationEU/PRTR under grant no. TED2021-131295B-C31-C31 (Value Awareness Engineering).

References

Christman J (2020) Autonomy in Moral and Political Philosophy. The Stanford Encyclopedia of Philosophy

Dworkin G (1988) The Theory and Practice of Autonomy. Cambridge University Press, Cambridge

Floridi L (2016) The 4th revolution: how the infosphere is reshaping human reality, First published in paperback. Oxford University Press, Oxford

Grobelnik M, Perset K, Russell S (2024) What is AI? Can you make a clear distinction between AI and non-AI systems? https://oecd.ai/en/wonk/definition. Accessed 1 Sep 2025

Lessig L (1999) Code: And Other Laws of Cyberspace. Basic Books

(2025a) automatic. Wiktionary, the free dictionary

(2025b) autonomous. Wiktionary, the free dictionary