

16. Governance of Artificial Agency and AI Value Chains: A Few Remarks on Autonomy from a Legal and Ethical Approach.

in López Castro, M. Cebal, P. Jiménez-Schlegl (eds.), *Regulating Autonomy: Ethics, Values and Governance in Intelligent Hybrid Systems*, Law, Governance and Technology Series, vol. 81. Hardcover ISBN 978-3-032-13062-4, eBook ISBN 978-3-032-13063-1, Cham: Springer Nature, 2026. https://doi.org/10.1007/978-3-032-13063-1_16

Pompeu Casanovas 

Pablo Noriega 

Abstract: This chapter develops a network of legal concepts as a framework for risks arising from the construction and implementation of autonomous artificial intelligence systems (AIs). The chapter examines the interrelationship between *legal autonomy*, *legal governance*, *legal agency*, *moral agency*, *delegated agency*, *harm*, and *liability*. To do so, it broadly distinguishes between Common law, Civil law, and Ethics. This conceptual network is re-examined in view of recent research in ethics and legal theory—including classical 20th-century legal theory, Artificial Intelligence & Law (AIL), and Law & Technology (LT). Computational and digital law as a systemic device, as meta-technology, and as a moral agent will be explored. This network of concepts sheds light on the meaning and functional sense of the basic concept of *autonomy* in the fields of law, multi-agent systems, and the sciences of design. Thus, its legal sense stems from a network of related concepts rather than from a lexical definition. Different levels of autonomy will be identified, and the chapter also shows that a common convergence on ethics has taken place in the last twenty years. Thus, the cognitive approach of value alignment and the emergence of ethical and legal ecosystems as regulatory devices in smart platforms and applications can frame the analysis of risks to be performed on them. The chapter ends with a working scheme on autonomy, moral agency, delegation, AI governance, and AI value chains.

Keywords: Artificial Intelligence Systems, Autonomy, Governance, Ethics, Agency, Moral Agency, Delegated Agency, Harm, Liability, AI value chains

Summary: 16.1 Preamble. 15.2 Basic distinctions. 16.2.1 A conceptual network to express complexity. 16.2.2 Foundations from ethics and legal theory. 16.2.3 Legal autonomy. 16.3 Definitions of legal autonomy, legal governance, legal agency, moral agency, delegated agency, harm, and liability across legal and ethical frameworks. 16.3.1 Methodology. 16.3.2 Summary of definitions. 16.4 Nuances from Artificial Intelligence & Law (AIL), and Law & Technology (LT). 16.4.1 The concept of legal personhood. 16.4.2 Overlapping contexts. 16.4.3 Legal and case-based sources: Highly politicized and monetized contexts. 16.5 Artificial Intelligence Systems (AIS) levels of autonomy and governance. 16.5.1 Levels of autonomy in AI systems and robotics. 16.5.2 Levels of autonomy and governance in online institutions. 16.5.3 AI levels of autonomy and governance in smart platforms (Platforms as a Service, PaaS). 16.5.4 Refining definitions for building AI ethical and legal value chains. 16.5.5 AI value chains (AIVC) and global value chains (GVC). 16.6 Conclusions and future work. 16.7 References.

Pompeu Casanovas
Artificial Intelligence Research Institute of the Spanish National Research Council (IIIA-CSIC)
Campus UAB, Barcelona, Spain;
LawTech La Trobe Research Group,
LaTrobe University, Bundoora, Victoria, Australia.
UAB Institute of Law and Technology (UAB-IDT)
pompeu.casanovas@iiia.csic.es

Pablo Noriega
Artificial Intelligence Research Institute of the Spanish National Research Council (IIIA-CSIC)
Campus UAB, Barcelona, Spain;
pablo@iiia.csic.es

16.1 Preamble

This chapter reflects on the notion of autonomy in artificial intelligence systems (AISs) from an AI governance perspective. Thus, the treatment of the risks and types of risk—as well as potential benefits—that the autonomy of intelligent systems can cause will be framed within a more general reflection on the ethical and legal consideration of their effects and sources. We will explore some problems comparatively, to shed light on how to address them, that is, what the conceptual scope of discourse is and where its limits lie. We won't focus on all possible problems, but only on a few that are relevant to the discussion. To be precise from the beginning the topics we address are: *Legal autonomy, legal governance, legal agency, moral agency, delegated agency, harm, and liability across legal and ethical frameworks*. We will not directly address the concepts of *accountability, responsibility, explainability, and explicability*.¹ This is not because they are not important in the regulation of intelligent systems, especially from the perspective of ethics and bioethics, but rather to narrow the scope of exploration, as they have been less closely related to the legal treatment of risks so far.

Let's start with the clear conceptualization of autonomy that has been paramount in defense and cybersecurity studies for the last ten years:

Autonomy results from delegation of a decision to an authorized entity to take action within specific boundaries. An important distinction is that systems governed by prescriptive rules that permit no deviations are *automated*, but they are not *autonomous*. To be autonomous, a system must have the capability to independently compose and select among different courses of action to accomplish goals based on its knowledge and understanding of the world, itself, and the situation. Recognizing that no machine—and no person—is truly autonomous in the strict sense of the word, we will sometimes speak of autonomous *capabilities* rather than autonomous *systems* (DSB 2016, 4).²

Thus, the U.S. Department of Defense figured out diverse scenarios in which autonomous systems could be deployed to save money, efforts and lives: (i) To covertly deploy networks of smart mines and Unmanned Underwater Vehicles (UUVs) to blockade and deny the sea surface, (ii) to differentiate between fishing vessels and fighting ships, without putting U.S. Service personnel or high-value assets at risk; (iii) to control rapid-fire exchange of cyber weapons and defenses, including the real-time discovery and exploitation of never-seen-before zero day exploits; (iii) to covertly operate inside the “turning radius” of adversaries to collect information or disrupt enemy operations; (iv) to adaptively jam and disrupt enemy Position, Navigation, and Timing (PNT) capabilities destroying their ability to coordinate operations; (v) to not only searching “big data” for indicators of Weapons of Mass Destruction (WMD) proliferation, but of deciding what databases to search to provide early warning and enable action (DSB 2016, *ibid.*).

These are complex scenarios that, ten years later, we believe may have been (at least partially) achieved and have become familiar to some extent. Now, forget about the military origin of the examples. If we chose this approach to begin with, it is because it very realistically situates the use of agents in artificial intelligence. When scenarios, including diverse contexts involving different agents and situations, have been explored and known, the problems to be solved can be more precisely defined. Interestingly, this approach also recognizes the limitations of predicting autonomy for both systems (intelligent artificial agents) and people (intelligent human agents). From a contextual

¹ Cf. Floridi and Cowls (2019), esp. on the difference between explainability and explicability and their relationship with responsibility and accountability. This analysis was resumed in Floridi *et al.* (2020), where they identify seven factors particularly relevant to AI as a technological infrastructure, to the extent that it is designed and used for the advancement of social good (AI4SG). To anticipate, these seven factors are: (1) falsifiability and incremental deployment; (2) safeguards against the manipulation of predictors; (3) receiver-contextualised intervention; (4) receiver-contextualised explanation and transparent purposes; (5) privacy protection and data subject consent; (6) situational fairness; and (7) human-friendly semanticisation.

² Cf. Shattuck (2015), De Lucia *et al.* (2019). “Autonomy results from delegation of a decision to an authorized entity to take action within specific boundaries. An important distinction is that systems governed by prescriptive rules that permit no deviations are automated, but they are not autonomous. To be autonomous, a system must have the capability to independently compose and select among different courses of action to accomplish goals based on its knowledge and understanding of the world, itself, and the situation” (Shattuck, 2015).

perspective, it is *not only the agents themselves, but rather the capabilities they can develop to solve specific problems that may be deemed autonomous*. This is rooted in the consideration of the level of abstraction, the layers of legal ecosystems, and the specific consideration of risks at a specific level (micro and meso) rather than in general.³

In short, this approach closely resembles the common sense developed for regulatory purposes by the legal tradition and scholarship for at least the last three centuries. There are many similarities. For instance, the way of conceiving action and agency; the difference between causality and causation, i.e., between what can be caused (factually) and the values with which something is judged to have been caused or not (epistemically); and especially the recognition of the complexity of cases and the need to develop types or categories to classify the problems that should be resolved. There are differences as well. The biggest one (or the major drawback) is that this approach must necessarily consider the *computational autonomy* of artificial information systems, whereas in the ethical and legal tradition this assumption has not been required so far. The cognitive perspective of *personal autonomy* (individual or collective) has been sufficient to assess the (formal) consequences and (material) effects of decision-making processes. This has dramatically changed.

We will divide the discussion into separate parts. In the first section of the chapter, we will address some definitions from the ethical and legal fields that can help us understand the problems of autonomy from the perspective of their practical consequences. We will show that it is better to outline the concept of ‘autonomy’ from a pragmatic approach, i.e. from a network of related concepts for making sense of its performance rather than from a lexical definition. In the second part, starting from this general framework, we will further refine the definitions based on the research carried out in the fields of Law & Technology and Artificial Intelligence & Law to flesh out this approach. The overall objective is to lay the conceptual, ethical, and legal foundations for the analysis of risk of intelligent artificial systems. We will not conduct this analysis in this chapter. We will only keep it in mind for further development in subsequent articles.

We must draw attention to the way the second part of the chapter develops the different levels of autonomy in AISs, robotics, and Platforms as a Service (PaaS). Thus, it distinguishes the proposed conceptual network into different regulatory areas: (i) civil law (Europe), (ii) common law (US and the Commonwealth), (iii) and ethics (philosophy). The chapter finally focuses on an operational framework for the governance of artificial intelligence, based on delegated autonomy and moral agency. This will eventually lead to the proposal for the construction of *ethical and legal artificial value chains*. To do so, first discuss the notions of “global value chain” (GVC) and “artificial value chain” (AVC). In what follows, we will divide the chapter into several sections: (i) On basic distinctions (section 15.2); (ii) the definitions of concepts that we have suggested (section 15.3); (iii) Nuances from Artificial Intelligence & Law, and Law & Technology (section 15.4); (iv) 15.5 Artificial intelligence systems (AISs) levels of autonomy and governance; (v) Conclusions and future work (15.6); and References (15.7).

16.2 Basic distinctions

16.2.1 A conceptual network to express complexity

Choosing a perspective that allows us to consider the complexity of human/machine interactions and the capabilities derived from this interaction with respect to autonomy requires both a series of prior distinctions and the development of conceptual relationships to identify the relevant framework. We acknowledge that just as computational ontologies cannot be confused with philosophical conceptions of ontology,⁴ the same could be said with respect to the capacity for autonomy. The moral and legal autonomy that we will briefly discuss in the next section (15.3) does not have the same scope of meaning and reference as artificial autonomy, but there are many links that show their connection. They have been identified and addressed in the development of intelligent software agents for at least twenty-five years and

³ This is compatible with the dynamic perspective of legal contexts developed at Pagallo *et al.* (2019a) (2019b); Noriega and Casanovas (2024); and Novelli *et al.* (2024a). The latter developed an integrated model that enables the estimation of AI risk magnitude by considering the interaction between (a) risk determinants, (b) individual drivers of determinants, and (c) multiple risk types.

⁴ Cf. Guarino (1995).

refer to the capabilities of multi-agent systems (MAS) to act and interact independently, i.e. similarly but not identically to humans. Simply put, “agents operate without the direct intervention of humans or others and have some kind of control over their actions and internal state”.⁵

Stemming from the idea of social interactive intelligence, Cristiano Castelfranchi proposed the modelling of autonomy using the notions of *dependence*, *delegation*, the characterization of the agent’s *architecture*, and, the joining of groups or organizations.⁶ In the so-called dependence theory, Castelfranchi *et al.* (2005) contended that the foundation of distributed activity, and the treatment of some basic aspects of a model of sociality—the notion of common world, the reasons for having and pursuing social goals, and the functions of social behavior—*presupposes and requires a theory of interdependence among the interacting systems*. Thus, “a theory of distributed action and interaction requires to be grounded upon a theory of interdependence, aimed at exploring how each agent’s limited capabilities concur to produce a network of social relations which even pre-exists the agent’s behavior, as well as their knowledge of the objective relations established.”⁷

Along these same lines, in the following sections (see esp. 15.3.3) we will consider autonomy from various positions coming from Artificial Intelligence & Law and Law & Technology. Mainly, among others, the theses developed in the last fifteen years by (i) Giovanni Sartor’s *computational law* perspective⁸; (ii) Ugo Pagallo’s positions on law as *meta-technology*⁹; (iii) and, in ethics, Luciano Floridi’s approach to the *digital regulation* of information systems.¹⁰ We will also refer to our own theses on the *governance of artificial intelligence*, the *metarule of law*, and *value engineering*.¹¹

For the moment, it is enough to cite here the retrospective analysis carried out by Sartor and Omicini (2016) on what autonomy might consist of according to the different AI approaches. They distinguish between three main positions: (i) the *cognitive perspective* by Castelfranchi and Falcone (already mentioned), grounding the autonomy of a system on the fact that the system tends to certain specific results due to its internal constraints or representations; (ii) George A. Bekey’s description of autonomy as the *capacity to operate* in the real-world environment without any form of external control and for extended periods of time; (iii) Stuart Russell’s and Peter Norvig’s (2003) characterization of autonomy as “an agent’s capacity to learn what it can to compensate for partial or incorrect prior knowledge”; thus, focusing on its *capacity to obtain new knowledge* interacting with the environment. Russell’s last book—*Human Compatible* (2019, 2023)—resumes this same view, but it is more critical with its possible consequences on *human* autonomy:

Machines may well understand that human autonomy and competence are important aspects of how we prefer to conduct our lives. They may well insist that humans retain control and responsibility for their own well-being, other words, machines will say no. But we myopic, lazy humans may disagree. There is a tragedy of the commons at work here: for any individual human, it may seem pointless to engage in years of arduous learning to acquire knowledge and skills that machines already have; but if everyone thinks that way, the human race will, collectively, lose its autonomy. The solution to this problem seems to be cultural, not technical. We will need a cultural movement to reshape our ideals and preferences towards autonomy, agency, and ability and away from self-indulgence and dependency—if you like, a modern, cultural version of ancient Sparta’s military ethos.¹²

⁵ Monroe and Luck (2004, p. 57).

⁶ Cf. Castelfranchi (2000); Castelfranchi and Falcone (2004). See Sartor (2014) for a discussion of the content of Castelfranchi and Falcone’s thesis on delegation.

⁷ Castelfranchi *et al.* (2005, p. 60).

⁸ Cf. Gelati *et al.* (2004); Sartor (2014); Sartor and Omicini (2016); Sartor *et al.* (2023); Novelli *et al.* (2024)

⁹ Cf. Pagallo (2012); Pagallo (2018); (iii) Pagallo *et al.* (2019a); Pagallo *et al.* (2019b); Pagallo (2025a); Pagallo (2025b).

¹⁰ Cf. Floridi and Sanders (2004); Floridi (2015); Novelli *et al.* (2024).

¹¹ Cf. Noriega *et al.* (2016); Noriega *et al.* (2023); Noriega and Casanovas (2024); Casanovas and Noriega (2022); Casanovas (2024); Casanovas (2025a); Casanovas (2025b); Casanovas *et al.* (2025).

¹² Russell (2019, 2023, p. 255-256).

Sartor and Omicini (2016) distinguish three orthogonal aspects of autonomy: (i) *independence*, which concerns a system's capacity to achieve a task on its own (capacity independence) and its being entrusted with such an independent achievement (organizational independence); (ii) *cognitive skill*, which concern a system's ability to perform complex discriminative functions pertaining to information acquisition, analysis, decision adoption and implementation; (iii) and *cognitive architecture*, which consists of a system's possession of *teleonomy* (direction to a purpose), *adaptiveness* (the capacity to get inputs from the environment and change internal states in such a way as to better respond to challenges) and *teleology* (the capacity to have representations of the environment and goals to achieve and to identify appropriate means). There is a link between these three aspects, although "they are not always jointly present, and their mismatch can give rise to malfunctions and liabilities". Thus, "the independence of technological systems is questionable whenever better performance could be obtained by combining technologies and humans into hybrid or joined cognitive systems – that is, by integrating mechanical and human skills."¹³

16.2.2 Foundations from ethics and legal theory

This interdependent nature of autonomy has also been pointed out from the perspective of ethics and law. One of the philosophers who has written the most about this notion, Gerald Dworkin has observed that "about the only features held constant from one author to another are that autonomy is a feature of persons and that is a desirable quality to have".¹⁴ He links it to the identification and critical reflection of an agent upon his/her first-order motivations, i.e. "it is only when a person identifies with the influences that motivate him or her, assimilates them to himself or herself, that he or she is autonomous."¹⁵

This perspective, centered on personality (individual and/or collective) and its capacity to self-government and to critically evaluate objectives and action plans, is common in Western rationalist philosophy, following the tradition of Aristotle, Thomas Aquinas, Hume, Rousseau, Kant, and Stuart-Mill. It should be noted that there is an unavoidable political dimension, which Isaiah Berlin summarized in his well-known differentiation between *negative liberty* (based on the absence of external constraints and interferences) and *positive liberty* (the ability to independently achieve objectives based on internal self-determined incentives).¹⁶ Sometimes this has been understood from an *individual* standpoint, linked to the concept of *self* and *authenticity*, and eventually projected or upgraded to a collective dimension.¹⁷ This opens the box for many criticisms coming from alternative political philosophies (feminism, human rights, protection of vulnerable people, etc.) more grounded on existential, phenomenological or hermeneutic positions.¹⁸

Liberty has also been understood both as (i) *conditions* for the legal order, (ii) and *limitations* to the legal order. Thus, on one hand, as the relationship between *autonomy*, *license*, *privacy*, and *legal order*, in which privacy and autonomy would be attributes of liberty and necessary requirements for a *just* legal order.¹⁹ And, on the other hand, as

¹³ Sartor and Omicini (2016, p. 73).

¹⁴ G. Dworkin (2008, p. 204).

¹⁵ G. Dworkin (2015, p. 7).

¹⁶ His original formulation reads: "The first of these political senses of freedom or liberty (I shall use both words to mean the same), which (following much precedent) I shall call the 'negative' sense, is involved in the answer to the question 'What is the area within which the subject--a person or group of persons--is or should be left to do or be what he is able to do or be, without interference by other persons?'. The second, which I shall call the positive sense, is involved in the answer to the question 'What, or who, is the source of control or interference that can determine someone to do, or be, this rather than that?'" Berlin ([1958] 1969, p.118).

¹⁷ "Put most simply, to be autonomous is to govern oneself, 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. [...] Autonomy concerns the independence and authenticity of the desires (values, emotions, etc.) that move one to act in the first place." Christensen (2020).

¹⁸ Referring to Feinberg (1989), Christensen (2022) contends that he claims that there are at least four different meanings of autonomy' in moral and political philosophy: (i) the capacity to govern oneself, (ii) the actual condition of self-government, (iii) a personal ideal, (iv) and a set of rights expressive of one's sovereignty over oneself (Feinberg 1989). Thus, (i) "one might argue that central to all of these uses is a conception of the person able to act, reflect, and choose on the basis of factors that are somehow her own (authentic in some sense)", and (ii) the requirement of second order self-appraisal for autonomy as a set of competences can be equally challenged.

the relationships between law and morality, in which public morality would hold as reasonable boundaries for the implementation and emergence of the legal order.²⁰

Whatever the theory and the differences might be, there is a common agreement among philosophers and scholars who have addressed ethical and legal autonomy on at least two points: (i) the need to base its meaning on an *interdependent network* of related concepts rather than on a single lexical definition; (ii) the need to start from a *theory* or anyhow previous philosophical grounds to establish the functions of its components.

For Dworkin, a theory “requires conditions of adequacy; constraints we impose antecedently on any satisfactory development of the concept. In the absence of some theoretical, empirical or normative limits we have no way of arguing for or against any proposed explication”.²¹ He singles out the following criteria: (i) logic consistency, (ii) empirical possibility, (iii) value conditions, (iv) ideological neutrality, (v) normative relevance, (vi) judgmental relevance. As for the law, Lapidoth (1994) and Heintze (1998) identified at least four conceptions of legal autonomy: (i) as a right to act upon one's own discretion in certain matters; (ii) as a synonym of independence; (iii) as a synonym of decentralization; and (iv) as exclusive powers of legislation, administration and adjudication in specific areas of an autonomous entity. We will elaborate on these distinctions in the next section.

16.2.3 Legal autonomy

What should be noted from an epistemic point of view is that the different weight, value, and degree that legal scholars grant to the concept of autonomy situates it differently in the various proposed theories. The concept itself has grown in a steady and organic fashion and become more relevant since the formulation of the original theories of legal positivism and realism have evolved into more nuanced forms of regulation. We will focus on this evolution.

In Hans Kelsen's *Pure Theory of Law* (1960), the concept is assimilated with that of the German and Austrian neo-Kantian philosophy of the early 20th century. Autonomy is conceived as a precondition, a reflection of liberty, i.e. as a natural feature of human freedom of reasoning and action. Here, Kelsen dealt with autonomy only as related to *subjectives Recht* (in opposition to objective ones). Thus, criticizing Georg Puchta's formulation of rights, he contended that

if one can talk at all about self-determination of the individual in his capacity as a legal subject, namely in the realm of so-called private law (with respect to the law-creating function of a legal contract), then legal self-determination, that is, autonomy, is present only in a very limited sense. For nobody can create rights for himself, because the right of the one presupposes the obligation of the other, and such legal relation can regularly only be established in the field of private law, according to the legal order by an agreement of two individuals.²²

Later, following his move to the United States and the need to strengthen democratic values, the concept is treated according to the foundation for the legitimation of law: “The liberalism inherent in modern democracy means not only

¹⁹ Cf. Sellers (2008, p. 1-2): “*Autonomy*, in its simplest and most natural sense, signifies self-rule: the right of states, or of families, or of associations or individuals to make their own laws for themselves. Understood in this way, autonomy is almost a synonym for *license*, which is to say, the ability to do what one wants, without restraint. Autonomy differs from license, however, in that it implies some measure of self-restraint. This difference is not in itself enough to justify the concept's popularity. What makes autonomy so desirable is its inevitable connection with (and restraint by) liberty, understood as the right not to be interfered with by the state or by others, except to the extent that this interference is warranted by the common good of society as a whole. [...] The importance of autonomy in law is also intimately connected with the concept of *privacy*, which guards individuals, families and associations against unwarranted intrusion. Privacy is the negative expression of the positive value expressed by “autonomy.” Autonomy signifies the right to decide for oneself. Privacy signifies that zone in which no others may interfere. *Both privacy and autonomy are fundamental requirements in any just legal order because they both are basic attributes of liberty, and liberty is a fundamental element of the common good that all legal systems have a basic obligation to establish and protect.*” [our emphasis]

²⁰ Stanton-Ife (2022).

²¹ G. Dworkin (2008, p. 204).

²² Kelsen ([1960] 1966, p. 170).

political but also intellectual autonomy of the individual, autonomy of reason, which is the very essence of rationalism”.²³ However, even though, this cannot be understood as a concession to any ethical cognitivism.²⁴

Likewise, at the same time, Alf Ross wrote *Why Democracy?* (1952) setting a clear position regarding the legitimacy and foundations of his legal theory: “Democracy is the form of government in which voluntary support rooted in the individual’s autonomy is greatest”.²⁵ Also:

While direct democracy firmly holds to the basic democratic ideas of autonomy, self-determination and personal responsibility, representative democracy signifies a modification of these by linking them with the idea of leadership in recognition of and confidence in the greater knowledge and ability of others. Here, of course science cannot decide the issue. Everyone must make up his own mind; I choose representative democracy.²⁶

But this did not mean either a concession to ethical cognitivism, since it is not necessary for the delimitation of the concepts of validity, efficacy, and effectiveness that he elaborated in his normative theory, based on the logical inferences of legal effects from factual conditions.²⁷

What both authors emphasize is the importance of the separation between *private autonomy* (in contracts) and *public autonomy* (in the public sphere), i.e., the framework for the delimitation of the normative power of the state. Even within the pragmatic turn that characterizes Ross’ *Directives and Norms* (1968), the Danish philosopher distinguished “the morality of conscience”, *autonomous identity* (“with regard to moral directives independently of all authority”) from autonomy that is exercised within the legal framework of *authority*.²⁸ In this sense, “private autonomy” is “similar to a legal institution but of a less formal and precise character”.²⁹ Ross (1959) distinguished between public legal acts and private transactions. Individual autonomy does not create legal frameworks, as “this autonomy [of the individual] has been circumscribed, but even circumscribed autonomy is autonomy still—not a social function.”³⁰

This holds for H.L.A. Hart’s *The Concept of Law* (1961) as well. Autonomy is considered with respect to international law and the boundaries of authority related to other states. Only with the concept of *inclusive positivism* suggested in the later *Postscript* (1994) of his main work, Hart’s attention turned to this notion due to his discussion with Ronald Dworkin.

Regarding legal realism, there is also an initial misunderstanding. The concept of autonomy is not mentioned even once in *The Bramble Bush* (1930). This was corrected later, when Karl Llewellyn became interested in the self-regulation of North American Indian institutions (Llewellyn and Hoebel, 1941). But if this was certainly the case, his

²³ Kelsen kept sustaining moral relativism until the end. Cf. Kelsen (1955, p. 27-28): “A relativistic value theory does not deny the existence of a moral order and, therefore, is not- as it is sometimes maintained-incompatible with moral or legal responsibility. It denies that there exists only one such order that alone may claim to be recognized as valid and, hence, as universally applicable. It asserts that there are several moral orders quite different from one another, and that consequently a choice must be made among them. Thus, relativism imposes upon the individual the difficult task of deciding for himself what is right and what is wrong. This, of course, implies a very serious responsibility, the most serious moral responsibility a man can assume. Positivistic relativism means: moral autonomy.” It follows: “This attitude, especially the respect for science, corresponds perfectly to that kind of person which we have described as specifically democratic. In the great dilemma between volition and cognition, between the wish to dominate the world and that to understand it, the pendulum swings more in the direction of cognition than volition, more toward understanding than dominating, just because with this type of character the will to power, the intensity of the ego-experience, is relatively reduced and self-criticism relatively strengthened; hence, belief in critical, and thus objective, science is secured.”

²⁴ Kelsen (1955, p. 97).

²⁵ Ross (1952, p. 127-28). Cf. also Ross (1952, p. 210): “While direct democracy firmly holds to the basic democratic ideas of autonomy, self-determination and personal responsibility, representative democracy signifies a modification of these by linking them with the idea of leadership in recognition of and confidence in the greater knowledge and ability of others. Here of course science cannot decide the issue. Everyone must make up his own mind; I choose representative democracy.”

²⁶ Ross (1952, p. 210).

²⁷ Cf. Ross (1957).

²⁸ Ross (1969, p. 59).

²⁹ Ross (1969, p. 135).

³⁰ Ross (1959, p. 213).

suggestions to overcome and improve the oral expression of the Pueblo Indians of New Mexico's rights, for instance, were headed to establishing written law through codes to regulate (and protect) their authority against federal laws.³¹ Thus, in spite of his division between *paper* (written) and *real* (working) rules, he still saw legal rules and reasoning as an autonomous normative order in itself.

Based on these observations, we contend that the interest in the notion of autonomy as a key legal concept, beyond authority and the constitution of independent orders and states, is related to the rise of ethics and public morality (i.e. *substantive* natural law) as constitutive of the internal order of norms. Lon Fuller's resizing of the ethical perspective as internal to law in *The Morality of Law* (1964); his previous critical view of the contract liability (will contract) as the result of a mere agreement³²; the assessment of the social, relational and policy dimension of contractual liability³³; and the rise of the perspective of rights due to Ronald Dworkin, and, to a lesser extent, to Joseph Raz³⁴, contributed to placing autonomy as the vertex of legal relations. Fuller, in *Consideration and Form* (1941), after proposing the integration of three different functions of legal formalities (evidentiary, cautionary, and channeling), concluded that "the future of consideration is tied up to the future of principle of private autonomy" and that reliance could "become increasingly important as the basis of liability".³⁵ In Common law, 'reliance' means the exchange of mutual promises or obligations to a contract. Thus, for him, mutuality of obligation and substantive commitment were even more relevant than the form of the contract.

Dworkin eventually characterizes autonomy as the "capacity to *form, criticize and pursue conceptions of the good life*, and that interest plainly cannot be served in a community whose rules do not provide ample freedom of choice over at least important matters [our emphasis]".³⁶ In this way, the importance of assumed beliefs, principles, and values, as opposed to rules, is emphasized. Law is conceived as a point of convergence and dissonance between interests and values in the market and civil society, fostering equality of personal autonomy, rather than as a mode of authoritarian social order. Hence the importance of considering the role of mentally handicapped people.³⁷ We deem relevant Dworkin's concerns:

Does a competent person's right to autonomy include, for example, the power to dictate that life-prolonging treatment be denied him later, even if he, when demented, pleads for it? Should what is done for a demented person be in his contemporary best interests, that is, such as to make the rest of his life as pleasant or comfortable as possible? Or in the best interests of the person who has become demented, that is, such as to make his life judged as a whole a better life?³⁸

Finegan (2015) has argued that Dworkin's concerns lie not with the equality of persons in the sense of the equality of personal worth, but rather with the equality of *personal autonomy*, based on a series of interrelated concepts — sanctity of life, dignity, critical interests, rights (including human rights) and personhood.

Thus, configuring a network of relevant concepts for understanding the content of the notion of autonomy lies in the way contemporary political philosophy and legal theory proceed. As soon as the concept of legal autonomy is

³¹ The history is well-documented in Twining ([1973] 2012, p. 358 and ff.).

³² Cf. Fuller on consideration and form (1941): "What is attempted in this article is an inquiry into the rationale of legal formalities, and an examination of the common-law doctrine of consideration in terms of its underlying policies".

³³ According to Kennedy (2000), Fuller's contribution in "Consideration and Form" was to the long-term project of moving from the will theory of contractual liability to a more relational and abstract/specific model by arguing persuasively that, even after the demise of the will theory, 'principle of private autonomy' was and should be the key consideration in private law theory, to be harmonized with, and occasionally balanced against, a small number of counter-principles.

³⁴ Stanton-Ife (2022) has shown that 'to harm a person' in Raz's *The Morality of Law* (1986), i.e. "to diminish his prospects, to affect adversely his possibilities", is essentially understood not as a setback to interests, but a setback to autonomy; and autonomy is essentially understood "as the ability to choose between an adequate range of valuable options, while in possession of the appropriate capacities to make such choices and while sufficiently independent of others".

³⁵ Fuller (1941, p.824).

³⁶ R. Dworkin (2000, p. 130).

³⁷ See on this subject, R. Dworkin (1986). Cf. also Winnick (1992).

³⁸ R. Dworkin (1986, p.4).

detached from the problem of the independence of law from other social sciences and from the problem of the sovereignty of nation-states, it becomes *relational*, since it relates to the capacities, competence, authority, and normative powers of hetero-, co-, and self-regulation; i.e., the power to provide *valid* regulatory models for a community, a bilateral or multilateral relationship, or the individual subject itself.

As Gerald Dworkin has suggested, some theory is required and should be developed to properly elucidate the concept. Nevertheless, from our point of view, we will not attempt to construct such a theory in this chapter. We will instead select some terms to weave the network of legal concepts that can subsequently be used for risk assessment. This network should be understood more as a scaffolding or a model to assemble than as a complete or proper theory.

16.3 Definitions of legal autonomy, legal governance, legal agency, moral agency, delegated agency, harm, and liability across legal and ethical frameworks

16.3.1 Methodology

We have drawn some conclusions established in our previous work on legal governance and the governance of artificial intelligence to select the terms to start with. We had already defined some of them—for example, *legal governance*, *legal ecosystems* and *smart legal ecosystems*³⁹—to reinforce the idea of the implementation of values in intelligent systems and the need to build smart legal ecosystems capable of offering real-time regulation to materialize and protect the rights of citizens, consumers, and workers.⁴⁰ We then carried out the exploration already offered in the previous sections to circumscribe the idea of *legal autonomy* and complete the selection of terms that are relevant to establishing and minimizing the risks inherent in legal governance through artificial intelligence. We found that there has been a natural tendency toward the application of ethical principles and an evolution in the way they are conceived by dominant legal theories. Thus, ethics is coming back to the core of legal reflection, whether to propose an inclusive legal positivism, or to maintain that this position is not sufficient to explain the functioning of multi-level governance mechanisms in the market and civil society. Nor in the argumentation of cases, or at the contextual level of micro-situations (whether in contracts, administrative acts, public policies, or private international law).

In one of our latest works on ethics and risk minimization we decomposed AI risks into three sorts of risks —*inertial*, *disruptive*, and *fundamental*— to be approached in different but complementary ways. We have shown that *value engineering* is a pertinent approach to address fundamental AI risks.⁴¹ To continue our analysis, we had a few intuitions: (i) a few concepts would be sufficient to establish a framework for managing the risks arising from the autonomy of intelligent systems; (ii) there is no need to reinvent the wheel, as the ethical and legal tradition can provide the required elements; (iii) however, the legal field is jurisdictionally circumscribed, and these elements could present significant variations depending on the field and legal culture considered. In the realms of law and ethics, the nuanced understanding of concepts such as autonomy, governance, agency, harm, and liability carries significant weight, influencing legal interpretations, policy formulations, and ethical decision-making. What we were looking for is also

³⁹ Cf. Casanovas and Noriega (2022); Casanovas *et al.* (2025); Casanovas (2025a) (2025b). As already stated in previous works, by *legal governance* we understand the set of processes that generate a sustainable regulatory ecosystem reflecting fundamental legal concepts of a modern democracy. By *legal ecosystems* (LE) we understand the complex and dynamic systems that include multiple levels of governance, ranging from local to national and international, and involving a wide range of actors and stakeholders, including lawmakers, judges, lawyers, law enforcement officials, civil society organizations, companies, corporations, and ordinary consumers and citizens. By *smart legal ecosystem* (SLE) we refer to regulatory systems partially embedded into cyber-physical systems (CPSs) that function in an intelligent environment encompassing ethics and law to achieve legal compliance in real time. Artificial Intelligence legal governance encompasses two distinct dimensions: governance through AI technologies (embedding human values via value engineering) and governance over AI technologies (aligning systems with legal instruments).

⁴⁰ Cf. Casanovas (2024) (2025); Casanovas *et al.* (2025).

⁴¹ Cf. Noriega and Casanovas (2024). Inertial risk reflects historical risk evolution, independent of evolving fundamental risk and disruptive events. Disruptive risks emerge from exogenous forces or radical AI innovations, significantly impacting AI artifact development, expectations, and adoption. Fundamental AI risk, intrinsic to AI deployment, can be addressed through preventive risk management by analyzing its origins and underlying forces. For preliminary thoughts on value engineering and a theory of values, see in this same volume (Noriega and Plaza, 2025).

a way to take differences into consideration but overcome them in a more general formulation in order to develop a conceptual model that could be acceptable and reused in a variety of contexts.

We therefore decided to explore this variety by constructing *prompts* for Gemini and evaluating the results obtained by comparing them with our own knowledge of the subject and by analyzing biases.⁴² This analysis seeks to illuminate the core meanings, contextual variations, and the underlying principles that shape the application of terms. Since these are probabilities based on millions of parameters and texts available online, we were able to verify that there is indeed a bias in favor of bioethical principles and health risks —minimizing or ignoring, for example, cybersecurity risks, whose negative effects have produced successive scandals but are less prevalent online. There is no treatment in depth. Nor does it capture the interrelationship between law and ethics and the evolution toward a legal integration of ethical principles that we have described in previous sections. However, the results are generally acceptable, with minimal conception errors. The following paragraph and Table 1 summarize the conclusions and reorder the terms provided according to the chosen fields —*Civil law, Common law, and Ethics*— after several iterations.

The definitions of *legal autonomy, legal governance, legal agency, moral agency, delegated agency, harm, and liability* vary across Civil Law, Common Law, and Ethics, reflecting the distinct principles and objectives of each domain. Legal autonomy, while present in both civil and common law, is approached with different emphases. Civil law often focuses on the codified rights of individuals and entities within a structured legal system, whereas common law emphasizes self-regulation and individual decision-making, particularly in areas like healthcare. Legal governance is a core function of both legal systems, with civil law relying on comprehensive statutes and codes, and common law evolving through judicial precedents. Legal agency, though more prominently discussed in common law as a fiduciary relationship, also exists in civil law as a contractual mandate. Moral agency in ethics centers on an individual's capacity for moral reasoning and accountability, influenced by ethical frameworks and situational contexts. Delegated agency, a legal mechanism for transferring administrative responsibilities, has parallels in the ethical context through the delegation of ethics oversight within organizations. Finally, harm is defined broadly in law to encompass physical, mental, and economic detriment, while in ethics, particularly concerning the harm principle, the definition and scope of harm remain subjects of ongoing debate regarding the limits of individual liberty. Liability, in a legal context, refers to the responsibility for harm or damages, with distinctions in its application and standards between civil and common law systems. In ethics, liability extends to moral and ethical accountability for actions and their consequences. Understanding these distinctions and commonalities is essential for navigating complex legal and ethical issues.

Table 1: Definitions of Key Terms Across Domains, after to Gemini

Term	Civil Law Definition	Common Law Definition	Ethics Definition
Legal Autonomy	Legally entrenched power of entities to exercise public policy functions independently within the state's legal order; self-referential nature of the legal system; protected individual rights; corporate policy-setting power.	Self-regulation or self-governance; relational, task-specific, and continuous (especially for technology); right of competent individuals to make informed healthcare decisions.	Not a primary focus as "legal autonomy"; ethical discussions often center on individual autonomy as self-determination and moral autonomy as self-legislation.

⁴² This work was carried out from 18/04/2025 to 5/05/2025. Prompts asked for the content across domains of the selected terms. Gemini used the information provided in a variety of significant private and public websites, using mainly USA sources. Among them: law | Wex | US Law | LII / Legal Information Institute - Law.Cornell.Edu, accessed April 28, 2025, <https://www.law.cornell.edu/wex/law>; Civil law systems - (AP US Government) - Vocab, Definition, Explanations | Fiveable, accessed April 28, 2025, <https://library.fiveable.me/key-terms/ap-gov/civil-law-systems>; Moral Agent - Ethics Unwrapped, accessed April 28, 2025, <https://ethicsunwrapped.utexas.edu/glossary/moral-agent>.

Legal Governance	Establishment and enforcement of legal order through written statutes and comprehensive codes; regulation of interactions between individuals, government, and private entities; judicial interpretation,	System based on legal precedents established by courts (case law); principle of <i>stare decisis</i> ; adversarial system; common law can inspire legislation; multi-layered sources of law.	Not a direct concept; ethical governance often refers to principles guiding behavior within a profession or society.
Legal Agency	Contractual relationship where an agent is appointed to perform specific legal actions on behalf of a principal (mandate); often requires the agent to act openly for a disclosed principal.	Fiduciary relationship where an agent is authorized to act on behalf of a principal to create legal relations with a third party; various forms of authority (express, implied, apparent); principal's liability for agent's actions.	Not a standard term; ethical agency might refer to acting on behalf of ethical principles or values.
Moral Agency	Not a distinct legal term; legal systems acknowledge individual responsibility for actions.	Not a distinct legal term; legal systems acknowledge individual responsibility for actions.	Capacity to discern right from wrong and be held accountable for actions; requires moral competency; responsibility to avoid unjustified harm; can apply to individuals and collective entities.
Delegated Agency	Transfer of responsibility for administering a program or function from a higher authority to another agency (government, nonprofit, for-profit) through formal agreements.	Similar to civil law; transfer of authority for specific tasks or programs, often formalized in writing.	Not a standard term; ethical programs within organizations involve delegation of authority to ethics officials (e.g., DAEOs) for oversight and implementation.
Harm	Loss of or damage to a person's right, property, or physical or mental well-being; includes injury, illness, death, and resulting losses (statutory definitions exist).	Similar to civil law; loss or damage to rights, property, or well-being.	Idea that actions should be limited only to prevent harm to others (harm principle); definition of harm (physical, psychological, through speech) is complex and debated.
Liability	Legal obligation to pay for damages or follow court orders in a lawsuit; responsibility for harm or loss arising from torts, contracts, or statutes; standard of proof is preponderance of evidence.	Legal responsibility for financial loss or harm caused by intentional actions, negligence, or breach of contract; includes vicarious and strict liability; often requires proving duty, breach, causation, and harm.	Responsibility for ethical wrongdoing and its consequences; moral obligation to make amends; ability to recognize and act upon ethical principles; accountability based on moral standards.

16.3.2 Summary of definitions

Four our purposes we will use the following definitions:

1. *Legal autonomy*: In civil law systems, legal autonomy is understood as the legally established authority of communities to independently exercise public policy functions, encompassing legislative, executive, and judicial aspects. This power is not absolute but operates within the overall legal order of the state. In common law systems, legal autonomy is often defined as self-regulation or self-governance, referring to the ability of an entity to establish its own rules of conduct and to follow those rules.

2. *Legal governance*: Legal governance in civil law systems is fundamentally characterized by the establishment and enforcement of legal order through written statutes and comprehensive codes. Legal governance in common law systems is fundamentally different, relying on legal precedents established by the courts, also known as case law. This system is based on the idea that judicial decisions in past cases should guide the rulings in future cases with similar facts.

3. *Legal agency*: While the concept of legal agency is predominantly associated with common law, it also exists within civil law frameworks, though the terminology and emphasis might differ. In civil law, agency is fundamentally understood as a contractual relationship where one person, the agent, is appointed by another, the principal, to perform a specific legal action on their behalf. This relationship is based on a mandate or agreement between the principal and the agent, granting the agent specific powers to act in either legal or financial matters, provided that the outcomes benefit the principal. In common law, legal agency is a fundamental concept defined as a fiduciary relationship where one person, the agent, is authorized by another, the principal, to act on that person's behalf and is empowered to do what the principal could lawfully do in person. This relationship allows the agent to create or alter the legal rights, duties, or relationships of the principal with third parties. The principal assumes responsibility for the acts of the agent when the agent is acting within the scope of their authority.

4. *Moral Agency*: Moral agency in ethics refers to the capacity of an individual to discern right from wrong and to be held accountable for their own actions. A moral agent possesses the ability to make ethical decisions based on moral principles and has a moral responsibility not to cause unjustified harm. The exercise of moral agency requires moral competency, which encompasses several key abilities and traits, including moral reasoning (the capacity to think critically about moral issues), recognition (the ability to identify situations with ethical implications), response (the ability to decide on an appropriate course of action), discernment (having insight in moral situations), accountability (being responsible for one's moral actions), character (possessing moral integrity), motivation (having the desire to act ethically), and sometimes leadership (guiding others morally).

5. *Delegated Agency*: In a legal context, a delegated agency refers to an entity to which a higher authority has transferred responsibility for administering a specific program or function. This delegation involves the transfer of authority from the original governing body to another agency, which can be a state or local government agency, a public or private non-profit organization, or even a for-profit entity. While the term "delegated agency" is not standard in ethics, the concept of delegating authority and responsibility is highly relevant to the administration of ethics programs within organizations, particularly governmental bodies. In this context, the Designated Agency Ethics Official (DAEO) plays a central role. In USA, each federal agency is required to designate a DAEO who is responsible for directing the daily activities of the agency's ethics program and coordinating with the Office of Government Ethics (OGE). This designation represents a delegation of authority from the head of the agency to manage and administer the ethics program.

6. *Harm*: In a legal context, "harm" is generally defined as the loss of or damage to a person's right, property, or physical or mental well-being. This broad definition encompasses various forms of detriment that the law seeks to address. In ethics, particularly within the framework of the *harm principle*, harm is understood as the idea that actions should be limited only when they cause harm to others. This principle, a cornerstone of liberal thought, suggests that individuals should be free to act as they wish as long as their actions do not negatively impact others. The harm principle is primarily intended to restrict the scope of criminal law and government restrictions on personal liberty, rather than to serve as a direct guide for individual moral behavior.

7. *Legal liability*: In civil law, liability refers to a legal obligation requiring a party to pay for damages or adhere to other court-ordered enforcements in a lawsuit. It signifies the state of being legally responsible for something. Unlike criminal liability, which is often initiated by the state for public wrongs, *civil liability* typically arises from lawsuits brought by private parties seeking remedies such as monetary compensation or injunctions for harm or loss caused by a non-criminal action. Liability in common law signifies the state of being legally responsible or obligated for an action or its consequences. This responsibility can arise from *intentional torts*, *unintentional acts of negligence*, or

contractual agreements. In ethics, liability extends beyond legal obligations to encompass *moral and ethical responsibilities*.

8. *Moral liability* refers to the responsibility for ethical wrongdoing and its consequences. It involves the moral obligation to make amends or provide restitution for harm caused, even unintentionally. *Ethical responsibility* is the broader ability to recognize, interpret, and act upon ethical principles and values in a given context. While legal liability focuses on legal duties and potential legal repercussions, ethical liability concerns an individual's or entity's accountability based on moral standards and the potential for moral censure or the need for ethical repair. This can include the moral responsibility to prevent harm, act with integrity, and consider the ethical implications of one's actions.

16.4 Nuances from Artificial Intelligence & Law, and Law & Technology

16.4.1 The concept of legal personhood

We believe that the network of aforementioned concepts is sufficient for assigning risks to Artificial Intelligent systems (AIS), as defined by OECD:

An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, 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 and adaptiveness after deployment* [our emphasis].⁴³

The OECD definition stresses the different *levels of autonomy*⁴⁴ of AIS and the problem of *adaptiveness*⁴⁵ that follows from their implementation in real settings. The OECD definition of an AI system intentionally does not address the issue of liability and responsibility for AI systems and their potentially harmful effects, “which ultimately rests with humans and does not in any way pre-determine or pre-empt regulatory choices made by individual jurisdictions in that regard.”⁴⁶

It should be noted that formulating the conceptual network in the way we did, considering OECD's definition, modifies the assumptions that have been imposed in legal theory and jurisprudence since the 19th century. This is especially true because normative systems considered ‘legal’ are based on the assignment of rights and obligations relative to a subject endowed with *legal personhood*. Legal theorists of the 20th century also started from this standpoint, which we can trace back to Roman and Mediaeval law. Legal personhood is a fundamental concept in both Common and Civil Law systems, defining who can hold these rights and obligations. A legal person is an entity capable of participating in legal relations, such as owning property, entering contracts, and being held liable for harm. Historically, this status has been attributed to *natural* persons (human beings) and *juridical* persons (such as companies or states), which are creations of law designed to serve human interests.

⁴³ OECD (2023, p. 4).

⁴⁴ “*AI system autonomy* (contained in both the original and the revised definition of an AI system) 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 the AI system lifecycle, 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 these outputs being explicitly described in the AI system's objective and without specific instructions from a human.” (OECD, *ibid.* p. 6)

⁴⁵ “*Adaptiveness* (contained in the revised definition of an AI system) is usually related to AI systems based on machine learning that can continue to evolve after initial development. The system modifies its behaviour through direct interaction with input and data before or after deployment. Examples include a speech recognition system that adapts to an individual's voice or a personalised music recommender system. AI systems can be trained once, periodically, or continually and operate by inferring patterns and relationships in data. Through such training, some AI systems may develop the ability to perform new forms of inference not initially envisioned by their programmers.” (OECD, *ibid.* p. 6)

⁴⁶ OECD (*ibid.* p.6).

Proponents of extending this status to AISs argue that a *functional* approach is necessary.⁴⁷ They point to the complexity and unpredictability of AISs, particularly in areas like autonomous vehicles and financial trading algorithms. When a system can cause damage or make a decision without direct human intervention, a legal gap emerges. Attributing a form of “electronic personhood” to AISs could simplify liability by creating a single point of imputation for damages. This is analogous to how corporate personhood allows a company, rather than its individual shareholders or employees, to be sued. This new status would likely be a limited form of personhood, focused on duties and liability rather than fundamental rights.

There has been an intense debate in recent years about the attribution of personality to AISs. Twenty years ago, Floridi and Sanders (2004) argued that we need to treat AI systems as “moral agents” for specific purposes of accountability, but not as subjects of rights. They distinguish between *responsibility* (which an AI system cannot hold) and *agency* (which it can), suggesting that a new legal framework is needed to manage this distinction without resorting to personhood. Likewise, stemming from AI & Law, Gelati *et al.* (2004) suggested that the key question is not whether an AI system is “like a person”, but how a legal system should best classify and manage it to achieve desired outcomes like fairness and accountability. They were skeptical of the “electronic personhood” model and advocated for extending existing legal tools, such as tort and contract law, to deal with AI-related issues, rather than inventing a new legal category.

This debate has been recently and thoroughly analyzed by Novelli *et al.* (2024), contending that “the scientific and policy discussion on AI legal personhood follows extended periods of stability (*stasis*) interrupted by rapid paradigmatic shifts triggered by socio-legal catalysts and technological breakthroughs (*punctuations*).” Several factors would shape these shifts: (i) conceptions of legal personhood—*clustered* versus *singularist*⁴⁸; (ii) the state of AI technology, accessibility, and its commercial reach; (iii) socio-digital institutions that mediate between AI capabilities and social status. They single out two forces that modulate the depth and durability of each equilibrium: (iv) the interactions between cross-domain overlapping legal frameworks like data privacy, agency, liability, and cybersecurity, as evidenced by attempted legislative reforms; (v) historical judicial precedents (i.e., case law) on extending personhood to new entities. The authors conclude distinguishing between incremental advancements in *generative models* and *autonomous agents*, “which may result in the conferral of (partial) legal capacities in the short and mid-term and more transformative possibilities, such as AI integration with human cognition through Brain-Machine Interfaces (BMIs), which could catalyze a deeper reconceptualization of legal personhood in the long run”.⁴⁹

It is worth noting that this perspective carries out a functional (also historical) and normative analysis *at the same time*. Agency is held as passive *and* proactive; just as exogenous factors are considered from the semantic perspective of legal norms. While it contends that “which of these two conceptions prevails at any given moment depends less on logic than culture, politics, and institutional venue”, it also states that

Personification allows the law to “internalize” emergent non-human entities (organizations, algorithms, and human-nature associations) as quasi-actors, albeit not these entities themselves, but only as semantic artifacts, i.e., “persons”. By saying that persons are semantic artifacts, we mean they result from a bundle of expectations and many norms directed at the person. These persons become, indeed, attribution addresses for legal norms. Moreover, legal persons trigger a new dynamic in which emergent phenomena outside the law, including AI, become fully-fledged communication partners within the law. They now participate directly in legal communication.⁵⁰

⁴⁷ Cf. Solum (2020). Also, referred to agents, if an artificial agent can be described as (i) rational and (ii) interactive, then we can ascribe (iii) responsibility and (iv) personhood to it, and consequently we can recognize it as having rights based on those capacities and attributes (Laukyte, 2017).

⁴⁸ In the ‘singularist’ view, an entity qualifies as a legal person simply by holding, or having the capacity to hold, any single right or duty. The ‘clustered’ view “argues that legal personhood consists of vesting (or having the capacity to vest, given appropriate conditions) a large cluster of legal roles, such as, in the private law domain, the ability to make contracts and own property, protection against harmful behavior by others, powers to activate judicial proceedings and participate in political functions, as well as liability for own wrongful behavior.” Novelli *et al.* (2024).

⁴⁹ Novelli *et al.* (2024, *ibid.*).

⁵⁰ Novelli *et al.* (2024).

Let's briefly summarize this position. The “personification” of AISs—the personification of algorithms—, framed through the process of “institutionalization”, is the key point of the analysis. The authors distinguish three distinct but interconnected forms of social, economic, and legal personification. They qualify them as forms of ‘socio-digital institutionalization’ based on *assistance*, *hybridity*, and *interconnectivity*, requiring different regulatory approaches in consideration of the corresponding risks.

The risks of ‘digital assistance’ are realized when tasks are delegated to single algorithms, the risks of ‘human-machine associations’ when humans and algorithms together form a collective actor, and the risks of ‘digital interconnectivity’ when social communication is only indirectly coupled to a crowd of interacting algorithms.⁵¹ Via these three forms of personification, alternatively: (i) technological risks are transformed into social risks, “the law must decide, according to its criteria, what degrees of legal personhood it attributes to the digital actants”, and “in the digital principal-agent relation, rules of vicarious liability for the actant’s decisions are needed”; (ii) responsibility for actions can only be ascribed to the whole hybrid entity, rather than to the individual algorithm or human involved, and “the law needs to develop new liability rules for human-machine collectives”; (iii) soft-law rules must be oriented to compensation, precaution, and the broader social implications of interconnectivity damages.

16.4.2 Overlapping contexts

In our opinion, Novelli *et al.* (2024) provide a useful and comprehensive classification of how the autonomy of AISs is articulated with their risks from a socially constructed approach. But this entails a social *reification* of the concept of autonomy via semiotic concepts (such as the re-semantization of the structuralist notion of “actant”) and the idea that “through personification, the social system ‘parasitizes’ the intrinsic dynamics of autonomous processes in its environment”.⁵² This is not the only way to face its social and political dimension. There are other options for considering contexts from an epistemic point of view, without having to resort to the twofold (normative/empirical) notion of ‘personification’; for example, by distinguishing between *electronic* and *online* institutions⁵³, or by structuring the potential risks based on concepts already existing in different legal cultures and defining the different dimensions in which these concepts—such as validity, compliance, effectiveness, efficiency, etc.—can operate in relation to data spaces and the levels and types of regulatory enforcement. In a platform-driven economy, platforms behave as online institutions, where sensors and actuators can operate through information processing modules at different layers of execution and control. What we deem important and should be identified are the types and organization of the information processing flows. We will return to this thread later (Section 15.5).

⁵¹ These are the forms of institutionalized personification: (i) *Digital Assistance* (or social personification), in which the delegation of tasks from a human actor to an algorithm creates a principal-agent relation between them. Such principal-agent relations presuppose social personhood for both the principal and the agent that can be framed according to several social theories, mainly Bruno Latour’s Actor-Network Theory (ACT). ACT studies the interactive qualities that transform an algorithm into an ‘actant’, to use ACT’s (and Greimas’ structural semantics terminology). (ii) *Digital Hybridity*, in which closely intertwined interactions between algorithms and humans may engender new forms of collective actorship, and responsibility for actions can only be ascribed to the whole hybrid entity, rather than to the individual algorithm or human involved. (iii) *Digital Interconnectivity*, that focuses on the systemic behaviour resulting from the linkage of machine agents, where “rather than principal-agent relations or hybrid integrations between humans and machines, we encounter heterarchical interconnected processes between algorithms”.

⁵² Teubner (2006, p. 504).

⁵³ Online institutions should be differentiated from electronic institutions. Electronic institutions (EI) can be roughly defined as a collective activity where agents perform within a shared state of affairs and under the effective enforcement of the explicit rules of the game (Noriega, 2024). Online institutions (OI) require a narrower conceptualization. They can be featured as the class of multiagent systems that are: (i) open (there is an ‘inside’ and an ‘outside’ of the OI); (ii) hybrid (involving human and software agents); (iii) situated (it is part of the actual world and functions within a particular sociotechnical context); (iv) online (the OI is a technological entity, and agents interact with it and among themselves via the environment(s) in which they are situated); (v) regulated (all agent interactions are subject to some constraints that are declared and enforced by the OI); (vi) state-based (the institutional state is unique and the same for every participating agent, and only enabled institutional actions and feasible institutional events can change it); (vii) and satisficing the dialogical and observability stances. Cf. Noriega *et al.* (2023).

In the field of Law & Technology, authors have distinguished several overlapping contexts to situate the problem. Kurki (2019) distinguishes between (i) the *ultimate-value context*, in which AIS are of ultimate value and therefore worthy of receiving some of the protections that legal persons such as human children enjoy; (ii) the *responsibility context*, in which liability can be attributed to machines, i.e., could self-driving cars or autonomous security robots be held criminally or tortuously liable for their actions?; (iii) the *commercial context* has to do with AIs' functioning as commercial actors (buying, selling, and so on).⁵⁴ Kurki applies the so-called *Bundle-Theory* of legal personhood to artificial intelligence. For him, legal personhood is a cluster concept, i.e., a complex concept consisting of several interrelated elements. Martin (2025) has recently observed that bundle theory, which treats each unique form of legal personhood as its own unique "bundle" of rights and duties, is an interpretation which US courts already have implicitly endorsed in some cases.⁵⁵

Pagallo (2013) has explored new forms of accountability for the behavior of robots as well as traditional ways of distributing risk through insurance models or authentication systems. The idea that (certain types of) robots may be held directly accountable for their own behavior has a precedent in the ancient Roman law institution of *peculium*.⁵⁶ This is relevant because it fosters different ways of producing effects that can be deemed legally valid, heading to the generation of new instruments for analysis and legal governance, as "it is likely that in the field of contracts, the growing autonomy of robots will affect basic concepts such as foreseeable harm, individual negligence or fault".⁵⁷ Hence, "although robots have no consciousness, free will or human-like intentions, the level of robotic autonomy is sufficient to have relevant effects in the civil (as opposed to the criminal) side of the law".⁵⁸ Thus:

The aim is to further distinguish between robots as simple tools of human interaction and robots as proper agents in the civil law field. Although current rules bar the acceptance of the legal agency of robots in certain cases, such legal agency makes sense in that humans delegate relevant cognitive tasks to robots. These machines can send bids, accept offers, request quotes, negotiate deals and even execute contracts, so that the level of autonomy, which is insufficient to hold robots criminally accountable for their behavior, is arguably sufficient to acknowledge new forms of artificial agency in the law of contracts.⁵⁹

His analysis operates in a threefold level of abstraction: (i) The legal personhood of robots as proper legal "persons" with their constitutional rights; (ii) the legal accountability of robots in contracts and business law; (iii) new types of human responsibility for others' behaviour, e.g., extra-contractual responsibility or tortuous liability for AI activities.⁶⁰ Only the two latter statements should be taken into consideration. We will also embrace this stance and reject applying legal personhood to AIS, but for technical reasons this time, using Occam's razor (see above, Section 15.5.3). The philosopher of Turin is focusing instead on the emergence of a spontaneous order out of robots' behavior, i.e. the idea that "complexity does not necessarily entail uncertainty or legal chaos".⁶¹ He reminds us that law is a polymorphic field, in which reason and irrational or non-rational behaviors are intertwined and eventually blurred. This behavior is not only characteristic of the inventive capacity of AIS but (and more properly) also of human beings.

⁵⁴ Kurki (2019, p. 176).

⁵⁵ Martin (2025) explicitly mentions *Nonhuman Rights Project v. Breheny, People ex. Rel Nonhuman Rights Project v. Lavery*.

⁵⁶ "In Justinian's Digest, the mechanism of *peculium* enabled slaves, deprived of personhood as the ground of individual rights, to act as estate managers, bankers or merchants. Similarly, I suggest that a sort of portfolio for robots could guarantee the rights and obligations entered into by such machines. Drawing a parallel between robots and slaves is attractive, since the aim today is the same as lawyers pursued in Ancient Rome: individuals should not be ruined by the decisions of their robots and any contractual counterparties of robots should be protected when doing business with them" (Pagallo, 2013, p. 87).

⁵⁷ Pagallo (2013, p. 89).

⁵⁸ Pagallo (*ibid.* p. 152).

⁵⁹ Pagallo (2013, p. 82).

⁶⁰ Pagallo (2018, p. 229-230).

⁶¹ Pagallo (2013, p. 176) singles out three fundamental aspects of the law that he calls *meta-technology*, considering the complexity of the law in terms of information (and vice-versa): (i) The normative complexity of the law as a set of rules or instructions for the determination of other informational objects; (ii) the knowledge and concepts framing the function and representation of a shared legal terminology; and (c) the laws of distribution of legal information hinging on the statistical properties of such quantities as the edges and diameters of the network.

16.4.3 Legal and Case-based Sources: *Highly politicized and monetized contexts.*

Let's have a look at the legal sources now. We will only sketch some points that we consider relevant to our discussion.

The European Parliament has largely abandoned its earlier recommendations for an “electronic personality” for robots in favor of a risk-based approach focused on human responsibility. The European Parliament in the 2017 European Parliament Resolution on Civil Law Rules on Robotics, paragraph 59, read:

creating a specific legal status for robots in the long run, so that at least the most sophisticated autonomous robots could be established as having the status of electronic persons responsible for making good any damage they may cause, and possibly applying electronic personality to cases where robots make autonomous decisions or otherwise interact with third parties independently.

There were many reactions to this statement, which seemed to grant rights to robots and AIS. Among them, the Expert Group on Liability and New Technologies appointed by the European Commission denied the necessity to adopt the notion of electronic personhood in its 2019 Report on Liability for Artificial Intelligence:

there is currently no need to give a legal personality to emerging digital technologies. Harm caused by even fully autonomous technologies is generally reducible to risks attributable to natural persons or existing categories of legal persons, and where this is not the case, new laws directed at individuals are a better response than creating a new category of legal person. (also: Key Finding 8: For the purposes of liability, it is not necessary to give autonomous systems a legal personality)

The Expert Group recommended instead adaptations and amendments to existing liability regimes, bearing in mind that, “it is impossible to come up with a single solution suitable for the entire spectrum of risks” (Key Finding 4), and that “strict liability is an appropriate response to the risks posed by emerging digital technologies, if, for example, they are operated in non-private environments and may typically cause significant harm” (Key Finding 9).

Fault and strict liability have been treated in the well-known by now scale or pyramid of risks set by the Artificial Intelligence Act (EU 2024), that came (partially) into force on 1 August 2024.⁶² One year later, on 2 August 2025, the AIA Code of Practice has been laid down.⁶³ Autonomy of AISs is barely mentioned, as the adopted classification assumed that the level of autonomy goes along with the value chain and the pyramid of risks. Intellectual Property (IP), cybersecurity and the protection of human and fundamental rights are deemed much more relevant.

To date, no major jurisdiction has granted full legal personhood to an AI system. Instead, courts and legislative bodies have consistently reinforced the status of AI as a tool or property, with liability remaining firmly with the human actors who create, own, or operate it. As said, the AI Act focuses on a risk-based approach to regulating AI, placing the onus of responsibility on developers, importers, and users, which is not an optimal solution and points at the concern for the development of the EU digital market in the first place. Anyway, the current legal consensus in Europe explicitly states that AI systems do not have legal personality or human conscience, and any required changes to the legal framework should start from this clarification.

⁶² Recital 12 of the AI Act replicates and assumes OECD definition: “AI systems are designed to operate with varying levels of autonomy, meaning that they have some degree of independence of actions from human involvement and of capabilities to operate without human intervention. The adaptiveness that an AI system could exhibit after deployment, refers to self-learning capabilities, allowing the system to change while in use”. Cf. also Art. 3, Definition 1. The degree of autonomy and scalability is taken into account to regulate systemic risks and the effects of general-purpose AI models. Cf. Recital 110: “Systemic risks should be understood to increase with model capabilities and model reach, can arise along the entire lifecycle of the model, and are influenced by conditions of misuse, model reliability, model fairness and model security, the degree of autonomy of the model, its access to tools, novel or combined modalities, release and distribution strategies, the potential to remove guardrails and other factors”. Cf. also Art. 14.3: “The oversight measures shall be commensurate to the risks, level of autonomy and context of use of the high-risk AI system, and shall be ensured through either one or both of the following types of measures: (a) measures identified and built, when technically feasible, into the high-risk AI system by the provider before it is placed on the market or put into service; (b) measures identified by the provider before placing the high-risk AI system on the market or putting it into service and that are appropriate to be implemented by the deployer.”

⁶³ <https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai>

However, in light of recent events, the highly contentious and political dimension of the regulation of liability and responsibility must be emphasized. There was a Proposal in 2022 for a EU AI Liability Directive as a complementary Directive for the AI Act. The Commission characterized the proposal as aiming to “improve the functioning of the internal market by laying down uniform rules for certain aspects of non-contractual civil liability for damage caused with the involvement of AI systems.” As recently explained by Andrews (2025), the decision to abandon the proposal was noted in the Commission's 2025 work program, which was adopted 11 February, and presented to the European Parliament the day after:

The decision to move on from the proposed directive was slammed by German Member of European Parliament Axel Voss, who told the IAPP the idea the directive would have created unneeded regulation with the AI Act in place missed its purpose. The directive was a “ex post liability mechanism” only kicking in when harms occurred, versus the AI Act's aim to prevent them. “Why the sudden U-turn? The answer likely lies in pressure from industry lobbyists who view any liability rules as an existential threat to their business models,” Voss said. “Big Tech firms are terrified of a legal landscape where they could be held accountable for the harms their AI systems cause. Instead of standing up to them, the Commission has caved, throwing European businesses and consumers under the bus in the process.”⁶⁴

In the United States, judicial decisions have addressed related issues but have stopped short of recognizing AI as a legal person. The most prominent examples come from intellectual property law. In the case *Thaler v. Vidal*, 43 F.4th 1207 (Fed. Cir. 2022), the U.S. Court of Appeals for the Federal Circuit affirmed the U.S. Patent and Trademark Office's (USPTO) stance that a patent author must be a human being, ruling that AI, specifically the AI system DABUS, cannot be an inventor. Similarly, in the copyright case concerning the comic book *Zarya of the Dawn* (US Copyright Office, *In re Zarya of the Dawn*, February 21, 2023), the U.S. Copyright Office revoked a copyright registration for AI-generated images, emphasizing that only human authors can receive such protections. The ruling distinguished the use of AI from traditional tools like cameras, noting that the AI user lacked the “sufficient control” over the output to be considered the “master mind” or author. This should not hide the fact that the number of patent applications related to AI is growing exponentially.⁶⁵

In the UK, Australia, New Zealand, Canada, and South Africa, reports by bodies like the Commonwealth Ombudsman highlight the legal challenges, focusing on principles of fairness, accountability, and contestability when AI is used in public administration. While there is not one single landmark case, decisions like *Amato v Commonwealth* (VID611/2019) have declared that fully automated decisions can be deemed irrational and thus unlawful, reinforcing the need for human oversight and accountability.

However, going under a case-by-case strategy, the proactive campaign of Stephen Thaler on AIS' rights before the Courts has fostered the debates regarding the legal consideration of the AISs creativity worldwide. In Australia, against the position of the United States Patent and Trademark Office (USPTO), the UK Intellectual Property Office (IPO) UK, and the European Patent Office (EPO), where an inventor must be a natural person, Justice Beach, from the Federal Australian Court, found in 2021 that there was “no specific provision [in the Patents Act] that expressly refutes the proposition that an artificial intelligence system can be an inventor”, and in such circumstances, AI can be an inventor. However, the next year, CJ Allsop and Justices Nicholas, Yates, Moshinsky and Burley, from the Full Federal Court, overturned this decision, ruling that AI is not capable of being an inventor under Australian patent law⁶⁶, a decision the High Court of Australia upheld by refusing special leave to appeal. This is the so-called DABUS

⁶⁴ Andrews (2025).

⁶⁵ “From 2002 through 2018, there was an increase in both the volume and the share of patent applications on AI.14 During that time, annual patent applications on AI increased by more than 100%, going from 30,000 to over 60,000.” (Lavrichenko, 2022, p. 701).

⁶⁶ Federal Court of Australia [2022]: Commissioner of Patents v Thaler [2022] FCAFC 62: “122. [...] we note that the outcome in the present case is the same as the outcome of the Court of Appeal in *Thaler UK*. Whilst there are important aspects of the reasoning of the learned judges in that Court with which we respectfully agree, we consider that the task in the present case focusses on the particular statutory language of the Patents Act, which in material respects differs from that in the equivalent patents legislation in the United Kingdom. [...]. 123: Decision. For the reasons set out above we consider that the first ground of the appeal

case — DABUS (Device for the Autonomous Bootstrapping of Unified Sentience)⁶⁷ is an AI system created by Stephen Thaler⁶⁸ which has given rise to various decisions by patent offices and other court rulings in several countries.⁶⁹ Thaler filed patent applications in more than 15 jurisdictions, including the US, Australia, South Africa, India, Germany, and Japan. According to claimant, DABUS had produced two inventions by its own without his intervention: (i) A food/beverage container which “makes tight packing grasping by a robotic arm”; (ii) and light that flicks in a unique way to attract the audience’s attention during emergency situations.

It should be observed that the relation of judges to the law binds them to *interpret* concepts according to the legal provisions of the law. In the Appeal from the United States District Court for the Eastern District of Virginia in No. 1:20-cv-00903-LMBTCB, in which Thaler sought for a reversal of the DABUS decision of the Trademark Office (PTO) on the meaning of “inventor” under the Patent Act, Judge Leonie M. Brinkema contended it clearly from the beginning:

This case presents the question of who, or what, can be an inventor. Specifically, we are asked to decide if an artificial intelligence (AI) software system can be listed as the inventor on a patent application. At first, it might seem that resolving this issue would involve an abstract inquiry into the nature of invention or the rights, if any, of AI systems. In fact, however, we do not need to ponder these metaphysical matters. Instead, our task begins – and ends – with consideration of the applicable definition in the relevant statute. [...] The sole issue on appeal is whether an AI software system can be an “inventor” under the Patent Act. [...] Here, there is no ambiguity: the Patent Act requires that inventors must be natural persons; that is, human beings.⁷⁰

It is worth noting that the legal debate concerns patents and intellectual property, not directly the personality and legal capacity of machines. There is no dispute about whether a system can technically generate new information, knowledge, or algorithms. This is clearly the case. What is under discussion is the attribution of its legal value: Who can *legally* benefit from it? Who receives the rights in the case of a patent? To whom is ownership attributed? This is also not new. It is the old nineteenth-century concept of the *legal relationship*—what it consists of and how it operates—that is brought up, and therefore the debate shifts the perspective of the structure considered capable of generating that value, i.e. its *legal framework* at theoretical level. The discussion is not over, and the legal positions vary, depending on the jurisdiction. Thus, South Africa has positioned itself in favor of an AIS being able to patent its creations—as the term ‘inventor’ has not been defined in the South African Patent Act—and Switzerland has indicated that, even if it is the machine that invents, intellectual property must be attributed to the human subject who created it. We will return to this in the next section.

must succeed with the consequence that the appeal should be allowed. We do not consider that it is necessary to consider the second. The result is that the decision of the primary judge should be set aside and the orders made by the Deputy Commissioner reinstated”.

⁶⁷ For a general explanation, see <https://en.wikipedia.org/wiki/DABUS>: “DABUS itself is a patented AI paradigm capable of accommodating trillions of computational neurons within extensive artificial neural systems that emulate the limbo-thalamo-cortical loop within the mammalian brain. Such systems utilize arrays of trainable neural modules, each containing interrelated memories representative of some conceptual space.”

⁶⁸ S. Thaler’s previous work on Creativity Machines is relevant for the understanding of DABUS. He explains his work on *perceptrons* and *imagitrons* as “the subjective feel of consciousness results from the associative chaining of relevant intact and de-graded memories. As these associative gestalt chains form, an attentional spotlight sequentially examines them in the same way our brains relive past, related experience. [...] Subjective experience stems from associative chaining of memories among sundry neural modules that then lead to the wholesale secretion of neurotransmitters to produce gut-level feelings and associated somatic effects. So, artificial neural systems may have an emotional response not only to things in the external world, but to internally generated ideas.” (Thaler, 2017, p. 22).

⁶⁹ Cf. Nieves (2022), Lavrichenko (2022), Saravanan and Prasad (2024), Agboola (2025). Dr. Thaler claimed that DABUS had produced two inventions: i) a food/beverage container which “makes tight packing grasping by a robotic arm” ii) a light that flicks in a unique way to attract the audience’s attention during emergency situations; and Thaler mentioned DABUS as the (only) inventor on both claims

⁷⁰ United States Court of Appeals for the Federal Circuit (August 5, 2022). United States of America US111-j Thaler v. Vidal, 43 F.4th 1207, 1210 (Fed. Cir. 2022).

16.5 AIS levels of autonomy and governance

16.5.1 Levels of autonomy in AI systems and robotics

The next two sub-sections will deal with levels of autonomy, in robotics and online institutions. As for robotics, this subsection should be read in conjunction with chapter 8 of this same volume: “Autonomy plus enhanced perception, mobility and dexterity allow robots to expand their activity in unstructured and unpredictable surroundings, pervading all the environments of human activity, in work and leisure, home and abroad, in intimacy and in public spaces” (Jimenez-Schlegel *et al.*, chapt. 8). This chapter includes an elaborated and updated table of types and levels of autonomy in robotics to deal with this problem.

As it is well-known by now, the US National Highway and Transportation Safety Administration (NHTSA), following SAE (2016), set in 2018 5 levels of autonomy for Automated Vehicles functionality: (i) *Level 0: No automation*, the driver is in complete control of all aspects of driving; (ii) *Level 1: Driver assistance*, automation of one control function such as lane keep assist or autonomous control, the vehicle can provide assistance for either steering or braking, but not both simultaneously; (iii) *Level 2: Partial driving automation*, automation of two control functions, the vehicle can perform both steering and braking/acceleration tasks simultaneously; (iv) *Level 3: Conditional driving automation*; the system can handle all aspects of the driving task under certain limited conditions. i.e. expect the driver to take control at any time with adequate warning; (v) *Level 4: High driving automation*, the vehicle can perform all driving functions and monitor the environment within a specific operational design domain (e.g., a particular geographical area); *Level 5: Full driving automation*, self-driving with no human control, the vehicle is capable of performing all driving tasks under all road and environmental conditions, no human driver is needed, and the vehicle may not have a steering wheel or pedals.

This adoption has led to the scale's notable success, likely due to its functionality, and it has been reused as a template in attempts at scaled classification of artifacts and AISs in multiple fields. It has had an impact, for example, in robotics (HRI), medicine, and corporate governance so far. But it will probably have a bigger impact. Following the same path, Yang *et al.* (2017) proposed 5 autonomy levels for medical robotics: (i) Level 0 (No assistance); (ii) Level 1 (Robot assistance); (iii) Level 2 (Task autonomy); (iv) Level 3 (Conditional autonomy); (v) Level 4 (High autonomy); (vi) Level 5 (Full autonomy). Yang *et al.* (2017) also contend that “at the higher levels of autonomy (specifically Level 5 and possibly Level 4), the robot is not only a medical device but is also practicing medicine.” This is controversial, but it is indicative of the effect caused by the development of artificial intelligence systems.

Likewise, in the field of corporate governance and focusing on robo-directors —automated boards, boards of directors, BoD— it has been argued that “as the capabilities of AI increase, the use of AI within the corporate governance and decision-making context is expected to shift from being merely assistive, to serving as an augmentative decision-support tool, and finally to conforming to the stage of a fully autonomous BoD.”⁷¹ The authors propose a scale of ascending levels of autonomy following the levels of “synergic” intelligence: *assisted* intelligence, *augmented* intelligence, *amplified* intelligence, *autonomous* intelligence, and finally, *autopoietic* intelligence (*ibid.* p. 355). Therefore, stemming from 0, they add and pile up five different layers: (i) Autonomy level 0: human directors are the sole decision-makers; (ii) Autonomy level 1: assistance, human directors are the sole decision-makers; (iii) Autonomy level 2: augmentation, human directors and AI systems share decision rights and learn from each other supported by simple digital devices/equipment, (iv) Autonomy level 3: amplification, human directors and AI systems are required to perform decision-making tasks jointly; (v) Autonomy level 4: high autonomy, AI systems can make decisions independently and operate within a predefined range without constant decision inputs from human directors; (vi) Autonomy level 5: full autonomy, AI systems are capable of making independent decisions for a particular scenario, and develop and expand this scenario over time.

To classify AI agents, Feng *et al.* (2025) adopt a different strategy, as they contend that an agent’s level of autonomy can be treated as a deliberate design decision, separate from its capability and operational environment. They define the levels of autonomy according to the roles an user can take in its interaction with the agent, i.e. as (i) *operator* (the user leads and makes decisions, the agent acts); (ii) *collaborator* (the user and the agent collaboratively plan,

⁷¹ Drukarch and Fosch-Villaronga (2022, p. 353).

delegate, and execute); (iii) *consultant* (the agent takes lead but consults the user for expertise/preferences); (iv) *approver* (the agent engages user only in risky or prespecified); (v) *observer* (agent operates with full autonomy under user monitoring).

16.5.2 AI Levels of Autonomy and Governance in Online Institutions

It is worth noticing that these approaches are connected with the original scale of autonomy, mainly based on a functional relationship between the AIS and a *human operator*—driver, doctor, user, or corporate director. This is a useful way of describing it. However, again, this is not the only one. Focusing on AISs and online institutions, we can figure out their artificial or computational independence from a higher level of abstraction, in which there is an inner engineering relationship within the AIS. Chopra and White (2011, p. 9) already observed that the autonomy of AI agents is not a binary concept. It can be better understood as a *spectrum* in which word processors and browsers with capacity to behave with a minimal amount of autonomy would be in one end of the spectrum, and learning systems equipped with sensors in the other. If we assume that autonomy refers to a certain capacity to decide independently of third parties, i.e. to a *constitutive relationship* in which a “principal” (a natural or legal person) delegates to another one the power to decide and carry out actions under certain conditions, we can propose a different scale for agents.

In AI there is an analogous delegation to an “autonomous agent” which is a computational (or artificial) entity. The more general notion of what an AIS consists of encompasses computing systems that have the property of autonomy but whose intelligence emerges from the combination of their own components. The important difference from conventional autonomy is that for the artificial agent or autonomous system, there is always a design and an engineering artifact that implements this delegated autonomy in one or another way. In this context, we proposed to distinguish five levels where the delegation of autonomy in AIS is incremental, progressively larger and, consequently, their governance becomes more complex (Noriega and Casanovas, 2022). The five levels of independence we identified in our original work can be restructured considering the rapid evolution of generative AI, but the essential is that they focus on the complexity of information processing mechanisms and the type of outputs they are able to produce. Fig. 1 reproduces these levels of autonomy and governance in a visual way through a truncated pyramid for Artificial General Intelligence (AGI) that we will explain right away below (point 5).

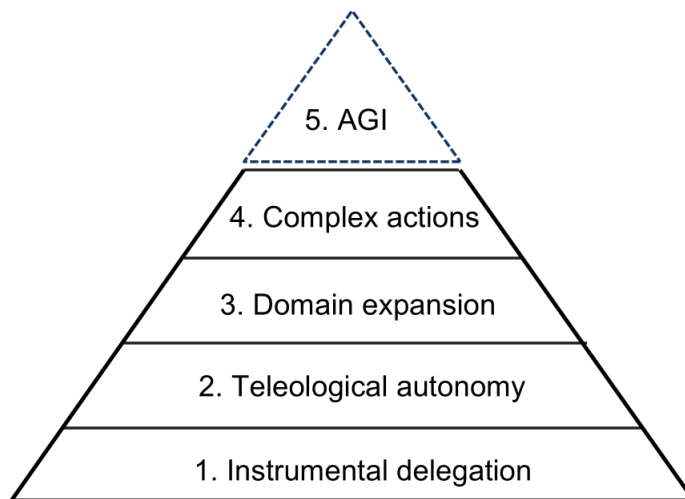


Fig. 1. AIS Levels of autonomy and governance

1. **Instrumental delegation.** At the *first level*, an AIS operates under *instrumental autonomy*, where it receives a specific, *predefined delegation*. The system makes decisions and executes actions autonomously, but these

are strictly confined to a well-defined part of a process and a limited universe of situations. Examples include assistive chatbots and autonomous cleaning devices like the Roomba. Potential issues at this level arise when the delegated scope of decisions exceeds the system's implemented capabilities. This can manifest complex management systems whose automated decisions are based on an incorrect interpretation of the processes they govern. For instance, Amazon's order fulfillment system may be perfectly adapted to the physical and operational characteristics of its robotic components but fails to account for the physiological needs of the human workers who are also part of the process.

These systems are widespread and perform narrow, well-defined tasks. Their autonomy is limited to a single, operational function. Some more examples are: (i) *Smart Home Devices*: A Ring camera's motion detection system, which autonomously decides to begin recording when it senses movement within its field of view; (ii) *Customer Service Bots*: A company's website chatbot that can only answer a specific set of frequently asked questions (FAQs) and follows a fixed script; (iii) *Email Filters*: Spam filters that autonomously classify incoming emails as junk or legitimate based on a set of predetermined criteria; (iv) The mentioned *Roomba Vacuum Cleaners*: The core function is to clean a room, and its decisions are limited to navigation and obstacle avoidance within that singular task.

2. **Teleological autonomy.** The *second level* is characterized by *teleological autonomy*, where a system *articulates and interconnects multiple instrumental tasks* to achieve a predetermined, high-level objective. The system independently manages a portfolio of actions to reach a specific end goal. A prime example of this complexity is found in systems that optimize aircraft utilization for commercial airlines. These systems autonomously coordinate a wide range of tasks, including route selection, maintenance scheduling, gate determination, seat assignment, and dynamic pricing strategies, all toward the singular objective of maximizing profit and operational efficiency.

Some examples: (i) *Supply Chain Management Platforms*: An AI-powered platform like C3 AI or Samsara that optimizes a logistics network. It autonomously coordinates inventory tracking, route planning, and delivery scheduling to meet the overall objective of efficient and timely delivery; (ii) *Programmatic Advertising Platforms*: Systems that use real-time bidding to place ads. They autonomously select ad inventory, determine bid prices, and serve ads to specific user demographics in milliseconds, all to achieve the goal of maximizing campaign ROI; (iii) *Portfolio Management Systems*: Financial AI systems that manage investment portfolios. They autonomously decide which stocks to buy or sell, based on market data and the ultimate goal of maximizing returns while adhering to a pre-set risk profile; (iv) *Agricultural Automation*: A precision agriculture system that autonomously monitors soil conditions, weather patterns, and crop health to coordinate the actions of watering, fertilizing, and harvesting robots to optimize crop yield.

3. **Domain expansion.** *Level 3 autonomy* transcends the boundaries of competence and responsibility that define the previous two levels. This is the most complex level, with three sub-categories based on how the system's capabilities expand beyond a single, predefined domain. This classification is merited for three distinct reasons:

3.1. Mixed or Hybrid Intelligence. This sub-level includes systems that incorporate both artificial and natural intelligence, often leveraging various forms of *crowd-processing*. These intelligent systems are built on architectures that decompose a problem into sub-tasks, with human intelligence solving these sub-tasks and the system automatically recomposing the partial solutions. This hybridization is significant for two reasons. First, it allows the system to harness general human intelligence for specific tasks while applying purely artificial intelligence to others. Examples include CAPTCHA for human-robot discrimination (sometimes termed "artificial artificial intelligence") and Ushahidi, a civic technology used for humanitarian mapping

and conflict management. Second, the use of an online coordination system allows for the integration of a vast number of individuals into a single collective activity. In this class of Level 3 systems, the issues of responsibility and design adequacy from Levels 1 and 2 persist, but new legal and ethical considerations emerge, including: (a) Representativeness, (b) Privacy, (c) Equity, (d) Safety of individuals involved in operation and use.

Examples: (i) *Content Moderation*: Platforms like YouTube or Facebook use AI to automatically flag potentially harmful content. However, human moderators review ambiguous or "edge-case" content to make the final, nuanced decision. This is a classic human-in-the-loop system; (ii) *Medical Diagnosis*: AI systems are increasingly used to analyze medical images (e.g., X-rays or CT scans) to identify anomalies. However, the system's output is an analysis, and the final diagnosis and treatment plan are made by a human physician who uses the AI's data as a tool.

3.2. Reusable and Generalizable Architecture. A second reason for distinguishing this level is the reuse of a highly sophisticated autonomous system architecture across new domains without requiring substantial adaptation. This differs from simpler systems (e.g., standard convolutional neural networks) that are merely retrained for a new domain. The distinguishing feature of Level 3 is a program's ability to re-train itself for new, equally complex tasks. Examples include AlphaZero and its successor, MuZero. These systems achieve their ability to perform highly complex tasks through the convergence of massive computational power and a sophisticated combination of AI techniques like cooperative problem-solving, machine learning, pattern recognition, and machine reasoning.

Examples: (i) *AlphaZero*: The most cited example. This AI system learned to master multiple games (chess, shogi, Go) with no prior knowledge of the game rules, simply by being given the rules and playing against itself. Its architecture is a generalized learning mechanism, not a specific game-playing program; (ii) *Robotics Platforms*: A single robotic architecture (e.g., from Boston Dynamics) that can be repurposed to perform multiple different physical tasks—from search and rescue to package delivery—by applying different learning algorithms and mission objectives.

3.3. Open-Domain Competence. The third defining characteristic of Level 3 is an open domain of competence, which necessitates generic problem-solving and an a priori undefinable scope of delegation. A classic example is Cyc, a system designed to anchor common-sense reasoning and provide generality to conventional expert systems. Cyc and similar systems like Open Mind Common Sense and DBpedia contain colossal ontologies and knowledge bases that describe how the world works, the result of ongoing collaborative efforts. The most significant impact of these systems occurs when they are integrated into autonomous systems that operate without a constrained knowledge domain. IBM's Watson is an emblematic case. Designed to understand and solve puzzles in natural language, it combines such open-domain systems with vast amounts of unstructured data and various intelligent modules for tasks like natural language understanding, strategic behavior, and complex problem-solving.

Examples: (i) *IBM Watson*: This system was designed to understand and answer questions posed in natural language across a vast range of general knowledge. Unlike a simple chatbot, its domain of competence is not limited to a specific database; it processes huge amounts of unstructured data to find a solution. (ii) *Generative AI with Retrieval-Augmented Generation (RAG)*: Modern LLMs, when combined with RAG, can access and synthesize information from a wide, external knowledge base (e.g., a company's entire document library or the web) to provide answers that go beyond their original training data.

4. **Complex actions.** Level 4 systems autonomously execute a repertoire of socially complex actions within the real world. Conceptually, they may be argued to exhibit a form of moral agency at a certain level of abstraction, as contended by Luciano Floridi. Prototypes and early versions of these systems already exist,

demonstrating the feasibility of their future construction. Examples include autonomous vehicles, caregiving robots (for patients, the elderly, and the disabled), and advanced programs like GPT-4, Gemini, Claude, and DeepSeek-V3. These systems are capable of integrating different forms of perception and action around a behavior whose rationality includes the ability to learn and competently perform tasks not explicitly foreseen in their original design. They operate in complex, real-world environments and must make decisions with significant social or ethical implications. They represent the current frontier of AI research and deployment.

Examples: (i) *Autonomous Vehicles*: A self-driving car in an unavoidable accident scenario. The system must decide how to minimize harm, a decision that involves a moral calculus (e.g., protecting the passenger versus a pedestrian). This requires the system to make a "moral" judgment based on its programming. (ii) *Elder-Care Robots*: A robot assisting an elderly person that has to make decisions balancing a user's autonomy (e.g., not wanting to take medication) with their physical health and safety. The system must be capable of navigating these complex, value-laden choices. (iii) *Automated Surveillance Systems*: An AI system that monitors public spaces to identify and flag potential threats. The decisions it makes have implications for privacy and civil liberties, and the system must be designed to avoid discriminatory outcomes.

5. **Artificial General Intelligence (AGI).** The *fifth level* is a conceptual frontier: the development of Artificial General Intelligence (AGI). This refers to an AIS capable of performing any task with the depth and generality of a human. This would entail an artificial realization very similar to human autonomy and moral responsibility. It is reasonable to cast serious doubt on both the feasibility and the desirability of such a development. This category however has been used as an evocative long-term research direction. Companies like OpenAI, Google DeepMind, and Anthropic are explicit in their mission to develop AGI and portray some recent developments like Open AI o3-pro, AlphaEvolve, and Claude 3.7 Sonnet, as steps toward this goal.

No systems currently exist at this level. This category is entirely theoretical and serves as a long-term goal for research. *Current Research Goals*: Companies like OpenAI, Google DeepMind, and Anthropic are explicit in their mission to develop AGI. Their current LLMs (like GPT-4 and Gemini) are seen as significant steps toward this goal, but not as AGI themselves. (ii) *Hypothetical Examples*: The AI portrayed in science fiction, such as JARVIS from the Marvel universe or the AI in the movie Her. These systems exhibit (without really holding) human-like intelligence, creativity, introspection, and the ability to operate across any domain, which are the hallmarks of AGI.

16.5.3 AI Levels of autonomy and governance in platforms (Platform as a Service, PaaS)

Based on the same cognitive approach to the internal structure and organization of design, we can add a further level of complexity related to the configuration of platforms. Service platforms can also be considered online institutions from a regulatory perspective. Therefore, their architecture should be understood from within to reconstruct their multiple levels of governance. This may seem like a very complex task, and indeed it is, since both the form of control over the functioning of the algorithms and the ecosystem it can produce with its end users will depend on their internal conceptual order. However, simplicity in design is a value, and therefore, an *inside/out* and *middle/out* epistemic perspective, rather than top-down and bottom-up, can help us better describe the forms of governance it can

develop. It is also a matter of intelligent design.⁷² Various technologies can be coordinated and articulated through a topology that establishes the levels and modes of control.

For example, OPTIMAI is a data-driven platform for zero-defect manufacturing (ZDM) to deploy a smart industry ecosystem. Figure 2 draws the main components of the conceptual architecture. Its building blocks can be summarized as follows: (i) Quality Control Sensor Network, (ii) Middleware, (iii) Machine-Operator Interface, (iv) Data Repository, (v) Blockchain (and smart contracts), (vi) Intelligent Marketplace, (vii) Digital Twins, (viii) Production Optimization, (ix) Smart Quality Control, (x) Visualization and Decision Support. Components are organized into modules that coordinate the information flows on the platform. Margetis *et al* (2022) describe the OPTIMAI architecture as follows:

The OPTIMAI service-oriented architecture (SOA) stack segments the envisioned ICT subsystems on a vertical axis, thus allowing for a high-level classification of different technological enablers on the grounds of their properties, relationships, and execution environment. Each layer thus comprises a major subsystem, with information flowing through the overall system from top (i.e., the IoT sensing devices) to the bottom (i.e., the actual UI/HMI software). The involved subsystems are: (i) the *Quality Control Sensors Network*; (ii) the *Edge Computing Modules*; (iii) the *Cloud Computing Modules*; and (iv) the *Users' Applications*.⁷³

A “smart factory” refers to the vertical integration of various components to implement a flexible and reconfigurable manufacturing system. This is one of the key features of the I4.0. OPTIMAI design that follows the Wang *et al.* (2016) model, according to which the smart factory framework consists of a *self-organized multiagent system assisted with big databased feedback and coordination*. The model includes an intelligent negotiation mechanism for agents to cooperate with each other. Thus, Figure 2 (architecture) shows the organization of components in the different layers, and their relation to the operational mechanism dual loop closed system: (i) The first loop consists of elements that are involved in the coordination and feedback provided at the Cloud level toward reconfiguring assets found in the Physical Resources Layer (“Coordinator”), (ii) the second one regards data visualization and manipulation manifested between the Cloud components engaged in statistical analysis (“Statistician”) and the supervisory terminal applications. Big data storage on the Cloud facilitates both sensing and acting, as well as control manipulation processes in the smart factory framework.⁷³ A third loop can be added to embed norms, privacy, and ethical values into the system (Casanovas and Hashmi, 2024).

⁷² Cf. Casanovas (2024, p.160). “From a regulatory perspective, *top down* refers to the decisions taken in upper organization levels and implemented (or enforced) through hard law mechanisms (laws, statutes, acts, and case-based law). *Bottom-up* refers to the negotiated order set by composition, covenants, collective agreements, and dialogical dispute resolution mechanisms. *Middle-out* refers to the mediating layer of technology that pervades any possible solution using information systems and the construction of conceptual and processual toolkits through semantics and AI algorithms (including symbolic AI, neural networks, and machine and deep learning techniques). *Inside-out* refers to the coordination via relevant norms and regulations, stemming from the technical protocols, recommendations, best practices, and standards that are embedded or incapsulated into information and CPS systems, and heading to more abstract principles and regulations, including policies and laws. Normative iterative lifecycles reflect the double looping encompassed by self-organized multiagent systems of smart factories.”

⁷³ Cf. Casanovas (2024, p. 154, following Margetis *et al.* (2022)).

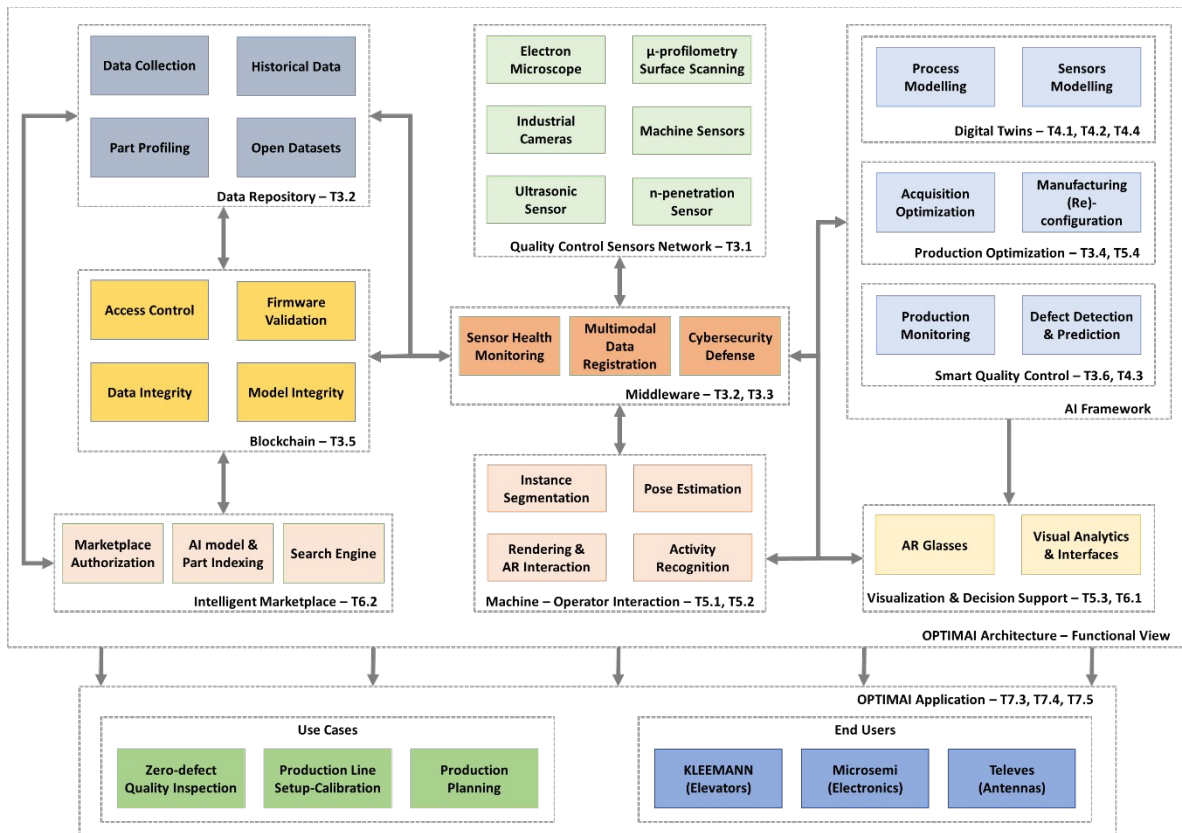


Figure 2. OPTIMAI Conceptual Architecture Diagram. Source: Apostolakis, Margetis et al. (2022, 75; latest version, 2023). Reproduced in Casanovas (2024, p. 156).

16.5.4 Refining definitions for building AI ethical and legal value chains

This example serves to illustrate a few key points: (i) In the governance of service platforms (PaaS) that operate as online institutions what matters is the information processing flows and how ecosystems can emerge from them; (ii) for their regulation, therefore, it is important to maintain the simplicity of design and reduce, not increase, the complexity of the conceptual apparatus that can be used for this purpose; (iii) metaphors coming from legal systems—such as qualifying AIS as ‘legal persons’—can be expressively used to enrich comprehension and as a base of discursive exchanges, but are not required to extract rules from norms, to formalize regulatory models, to build artificial online institutions or to assign risks; (iv) risk analysis can be conducted after, not before, knowing the modularization of these processing flows; (v) classic legal concepts can be adapted and outlined to models of governance for the implementation of legal ecosystems.

What is most interesting is that we can maintain the notion of autonomy of intelligent artificial systems without any need to substantially modify the concept as it has already been developed by different legal traditions. What changes are the instruments, contexts, and digital environments, not the basic concepts which can be adapted to new complex realities. For example, the introduction of simulations and sandboxes to test legal effects in different scenarios offers new possibilities for the implementation of regulatory (legal) systems. Concepts must, indeed, fit into digital scenarios that are very different from those known until now because the agents capable of deciding, planning, and foreseeing are artificial. Rephrasing Aristotle, this time, “the law is said in many ways because the legal phenomenon is far more complex than its own language” (Pagallo, 2013, p. 148).

This encompasses the ideas of working on the intermediate level of regulatory models—*self-, co-, and hetero-regulation*—and *in-between* the top-down and bottom-up orientations, by defining a more detailed legal governance toolkit through artificial intelligence, according to the following principles: (i) Modular adaptability; (ii) semantic interoperability; (iii) systemic interdependence; (iv) organic decentralization; (v) intermediate conceptualization; (vi) coordinated agency; (vii) middle-out (abductive) reasoning.⁷⁴

Legal governance entails restructuring the main elements of the law in a way that facilitates the emergence of legal sustainable ecosystems. This approach cannot be confused with any kind of authoritarian legal execution based on hard compliance only. Compliance (especially automated compliance) is an important subject that should be aligned with the human, civic and social values chosen to define rights and duties to foster wellbeing in a good society.

There are hybrid contexts, in short, between human and artificial agents. But this means that we can restructure them into new frameworks, in which the contexts of the web of data, the internet of things, and so-called Industry 4.0 (and 5.0) add complexity because they open new dimensions to the use of concepts (such as the legal ecological validity of regulations). In a recent paper, we developed a four-dimensional framework for the validity of legal ecosystems (language, society, legality, and data) using a tesseract model (Casanovas, 2025).⁷⁵ Mechanisms change, but the basic idea of maintaining a proportional balance between interests in building a fair and just society remains.

This does not entail any substantial modification on the governance of artificial intelligence either. On the contrary, the use of AISs is enriching the legal field because it is possible to flesh it out from within, impinging on its efficiency, and enriching its ethical values. To do this, we should avoid the problems of over-compliance and over-regulation.⁷⁶ We should avoid introducing more complexity into risk analysis but rather reduce it. Instead, we should face at least two technical issues that have not yet been satisfactorily solved: (i) The problem of imbuing values into artificial intelligence systems (value alignment problem, VAP); (ii) the problem of extracting formal rules from norms expressed in natural language.

Figure 3 draws a minimal scheme to begin with, focusing only on the two main branches of autonomy, i.e. *delegated autonomy* and *moral agency*. Very likely, this scheme will be modified in the next future, but it is nevertheless useful to show that simple formulations can help solve difficult issues. Note that at the core of the diagram, we have placed the main problem: the regulation and control of artificial intelligence systems through artificial intelligence itself.⁷⁷ Human intervention, especially in generative AI systems, cannot be let alone. It needs to be combined with AI techniques. Human-in-the-loop means AI-in-the-loop too.

As a result, this approach makes it possible to incentivize and promote the emergence of ethical and legal regulatory ecosystems that incorporate the so-called *AI value chain* (AIVC). It aims at the realization of values in the organization, coordination and implementation of AISs in specific contexts. This field is experiencing rapid development, especially since the concept has been incorporated into several recitals and articles of recent European legislation constituting a guideline for risk mitigation, the promotion of European values, and the development of human rights-based protections. The EU AI Act mentions AIVC from the beginning, in several recitals and articles.⁷⁸

⁷⁴ These were the components of the toolkit figured out in the AI4People model for AI and legal governance (Pagallo *et al.* 2019a), resumed and fleshed out in successive projects, books, and articles (Pagallo *et al.* 2019b), Poblet *et al.* (2019), Casanovas (2024); Casanovas *et al.* (2025).

⁷⁵ Cf. Casanovas (2025) for a more detailed explanation and the definition of the seven principles as components of regulatory models.

⁷⁶ On over-compliance as a means of deviate the attention and accommodate the law to one's interests, cf. de Koker and Casanovas (2024).

⁷⁷ In Casanovas and Noriega (2022) we described and faced the dilemmas of legal and AI governance.

⁷⁸ Cf. AI Act (EU 2024, e.g. Recital 9: “the obligations placed on various operators involved in the AI value chain under this Regulation should apply without prejudice to national law, in compliance with Union law”; Recital 88: “Along the AI value chain multiple parties often supply AI systems, tools and services but also components or processes that are incorporated by the provider into the AI system with various objectives, including the model training, model retraining, model testing and evaluation, integration into software, or other aspects of model development”. Cf. also Art. 25: Responsibilities along the AI value chain, and Art. 56, 2 d.: the measures, procedures and modalities for the assessment and management of the systemic risks at Union level, including the documentation thereof, which shall be proportionate to the risks, take into consideration their severity and probability

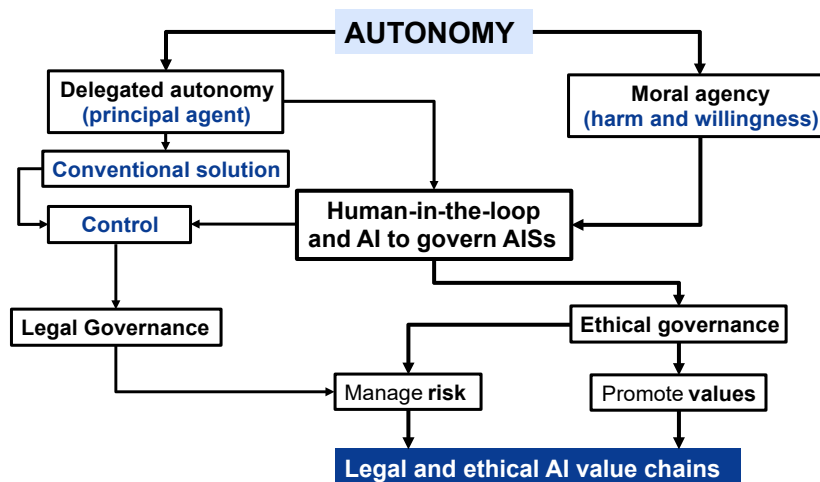


Fig. 3: Governance through imbuing values into AI systems and promoting AI ethical and legal value chains.

The Artificial Intelligence Value Chain is a framework that illustrates how organizations create value by transforming raw data into intelligence as a competitive advantage. It is not merely a linear sequence of steps but a dynamic and interdependent ecosystem. The framework functions in a dual capacity: (i) as a standalone industry model for companies that develop and deliver AI solutions, (ii) and as a horizontal enabler that drives transformation across other sectors like manufacturing, logistics, and healthcare, actually crossing the fourteen European common data spaces.⁷⁹

16.5.5 AI value chains (AIVC) and global value chains (GVC)

Engler and Renda (2022) define AIVC as “the organizational process through which an individual AI system is developed and then put into use (or deployed)”. Attard-Frost and Gray Widder (2025) follow this organizational thread and propose an *integrative* perspective of “ethics that foregrounds the value chains involved in providing resource inputs to and receiving resource outputs from AI systems”, “integrating a wide range of ethical concerns across many actors, resources, contexts, and scales of activity”.⁸⁰ Hence, value chains in computing and AI are not deemed equivalent to supply chains. These are organized according to a “goods-dominant logic” making their outputs consumable, while value chains follow a “service-dominant logic” based on intangibility, exchange processes, and a broader network of co-creative relations.⁸¹

It is our contention that we can extend this idea of bridging ethics and organizational management theories to our idea of integrated legal and ethical ecosystems through legal governance, i.e. encompassing legal autonomy, legal governance, legal agency, moral agency, delegated agency, harm, and liability. However, we differ from the strict

and take into account the specific challenges of tackling those risks in light of the possible ways in which such risks may emerge and materialise along the AI value chain”. This holds for the EU (2023), MiCA regulation of the digital market as well, even if AI value chain is not yet explicitly mentioned, cf. Recital 60: “To capture all transactions that are conducted in relation to any given asset-referenced token, the monitoring of such tokens therefore includes the monitoring of all transactions that are settled, whether they are settled on the distributed ledger (‘on-chain’) or outside the distributed ledger (‘off-chain’), and including transactions between clients of the same crypto-asset service provider.”

⁷⁹ I.e., agriculture, cultural heritage, energy, finance, green deal, health, industry (manufacturing), language, media, mobility, public administration, research and innovation, and skills and tourism. Cf. <https://digital-strategy.ec.europa.eu/en/policies/data-spaces> .

⁸⁰ Cf. Attard-Frost and Gray Widder (2025): “We define value chains as cocreation structures that exist within a network of actors and enable patterned resourcing activities to occur between actors. [...] We define AI value chains as cocreation structures that exist within a network of actors and enable actors to pattern the resource inputs they provide and the re-source outputs they receive from AI systems.”

⁸¹ Vargo and Lusch (2016); Attard-Frost and Gray Widder (2025).

separation between a phenomenological approach in computer science and a quantitative approach in economics and managerial sciences. Both are empirical, and therefore, we should be able to find metrics for legal and ethical AI value chains. This does not mean that the AI value chain is free from problems of political nature, as is the case, for example, with water management.⁸² Thus, there are additional reasons to imbue values into AISs. Any “hidden privatization” of decisions about public values and value conflicts should be avoided in the normative choices about AI risks.⁸³ Negative impacts on the AI value chain should be acknowledged and mitigated, and this entails a previous ethical commitment in the engagement of companies and stakeholders.⁸⁴

We should stress the *concentration of power* that it is taking place. The traditional value chain model is further complicated by the wider AI ecosystem that underpins it. This ecosystem includes foundational upstream components, notably computer hardware and cloud platforms. The hardware layer is dominated by a few semiconductor giants, such as Nvidia, Intel, and AMD, with Nvidia holding a particularly dominant position in the market for AI chips. Similarly, the cloud platforms, which provide the scalable compute infrastructure necessary for AI development, are controlled by a handful of *hyperscalers* like Amazon Web Services (AWS), Microsoft Azure, and Google Cloud. This market structure has led to an increasing trend of vertical integration. The major technology companies, which are dominant in cloud computing, are now actively developing their own chips and foundation models, *expanding their footprint across the entire value chain*. The extensive resources required for developing these general-purpose AI (GPAI) models—vast amounts of data, computing power, and financial capital—contribute to this market concentration. This concentration of power in the foundational layers of the value chain is a central element of the modern AI landscape. It implies that a handful of firms can exert immense influence over the entire ecosystem, raising significant regulatory questions. The governance challenge shifts from regulating a diffuse market of many small actors to managing the behavior of a few central, vertically integrated players. This concentration compels regulators to consider frameworks that address systemic risks at the level of the foundational model itself, rather than focusing solely on the final, end-user application.

We also believe there is a relationship between the “global value chain” (GVC) and AIVC. From a regulatory perspective there are differences that warrant separate treatment, but there are also relationships that should be considered. According to the existing literature, the concept of GVC is an extension of concepts such as value chain and global commodity chain.⁸⁵ The concept was endorsed by the United Nations Industrial Development Organization (UNIDO) about ten years ago, focusing on economic development and comparison of developing countries. UNIDO (2015) defines value chains according to four criteria or “entry points”: (i) as sets of value adding activities; (ii) as arrays of linkages; (iii) as networks or systems; (iv) as cycles. It distinguishes between *positive* and *normative* approaches that should be combined.⁸⁶ The former covers convenient heuristic means to help understand how they operate, e.g. how production steps are fragmented and distributed among geographical locations and analytical dimensions such as: (i) input-output structure; (ii) geographical distribution; (iii) role of lead firms or powerful intermediaries, suppliers, traders, etc.; and (iv) institutional context at the international, regional and national or local level. The latter, *normative*, focuses on what needs to change in order to improve the performance of the value chain, policy priorities linked to development goals, rights, obligations, and standards. In our opinion, what matters in this approach

⁸² Cf. Lehuedé (2025). “While the AI industry is happy to invest considerable amounts of resources in the development of so-called ‘green’ AI applications, communities and environments affected by the AI value chain are being kept in the dark when it comes to water availability” The authors conclude that “no ‘ethical’ and ‘sustainable’ AI would be possible as long the communities participating within AI value chain, and their ways of relating to the elements, are excluded from the design and development of so-called ‘intelligent’ systems.”

⁸³ Orwat *et al.* (2024).

⁸⁴ Cf. Reyes and Rajagopal (2025, p. 464): The negative impact on the value chain may lead to information obscurity, data inaccuracy, and develop biasness towards knowledge management. Hence, reliability of information depends on quality of information, reduction of data misuse, and reinforcement of data privacy and security. Sustainable administration of AI data sources drives effective implementation of justice through reduction in discrimination during information pruning process, enhancing data transparency, and promoting organic socio-technological interactions.”

⁸⁵ Liu *et al.* (2024). “GVC refers to a large cross-enterprise network organizational system built in the international market to realize the value of goods or services, covering various links such as raw materials, production, marketing, recycling, etc.”

⁸⁶ UNIDO (2015, pp. 18-19).

is the conception of the value chain as a process and result, i.e., *as the blending of several dimensions in the descriptive and normative axes that properly constitutes an institutionalization process.*

This process does not take place in a regulatory vacuum. The legal dimension of value chains (GVC, AIVC) implies that they cannot be treated homogeneously, since (i) they become *jurisdictional*, i.e. subject to positive law, both national and international, (ii) they come to depend on the internal dynamics of the different legal areas (corporate law, business law, administrative law, etc.). In the EU law, this entails what Becker (2023) has called a *fragmented institutionalization* of GVCs along at least three perspectives: (i) individual actor (corporate law); (ii) collective actor (consumer law); (iii) de-personalized institutionalization (market and trade law). Thus, she suggests “*the legal image as fragmented along different legal institutions* (company, network, market) that relate to different legal areas (company law, consumer law, market practices, and trade law) and correlate to different policy objectives (sustainability, digitalization, resilience, consumer protection, fairness) [our emphasis].”⁸⁷

The complex, multi-party nature of the AI value chain presents a fundamental challenge to traditional legal frameworks. The “pacing problem” addressed by Becker—the law’s inability to keep up with rapidly evolving technology—is evident in attempts to apply established legal principles to AI-induced harm. Historical models of the AI value chain often stem from the more specific machine learning (ML) value chain, which consists of five core stages: data collection, data storage, data preparation, algorithm training, and application development. Modern advancements, particularly with the advent of generative AI, have expanded this model, adding more granular downstream components. For instance, the downstream part of the value chain now includes machine learning operations (MLOps), which streamlines the process of taking models to production, as well as distinct stages for applications and services. This progression from foundational concepts to more complex, real-world applications highlights the rapid evolution of the industry itself.

16.6 Conclusions and future work

We have performed some operations that are worth summarizing. The purpose of this chapter has been to establish legal concepts that belong to the analysis of the risk arising in the context of intelligent autonomous systems. In this regard, we have distinguished autonomy in the reflective areas of Ethics, Common Law, and Civil law; and in the jurisdictional European, American and Common Law legal cultures. European scholars tend to be more concerned with a fundamental, philosophical re-evaluation of legal concepts. They actively explore whether a new legal category—a form of *limited personhood* or *agency*—is necessary to address the novel challenges of autonomous AI. The debate is often proactive and aims to build a new regulatory framework, as evidenced by the EU’s AI Act, which is a top-down, comprehensive law. In contrast, U.S. legal scholars are more focused on a pragmatic, often reactive, application of existing legal doctrines, such as intellectual property law or “bundle” theories. Commonwealth scholarship, rooted in the common law tradition, relies more on judicial precedent and literal interpretations of law. This often leads to a case-by-case approach, with courts reinforcing traditional definitions. Commonwealth scholars, particularly in places like Australia, tend to focus on a very specific, practical aspect of the issue—the use of AI in public administration. Their concern is less about the abstract concept of personhood and more about ensuring that automated decisions made by government bodies are transparent, contestable, and subject to human oversight and review.

The conclusions are simple, but consistent with our purpose: (i) These legal traditions and their ethical and political foundations are sufficiently rich to provide the foundations we need; (ii) however, these traditions can be

⁸⁷ Becker (2023, p. 323). “More specifically, at least three different forms appear in which GVCs become institutionalized in EU law: First, an *individual actor perspective* dominates the approach in EU company law, specifically regarding the sustainability of global production. This actor-centric perspective delineates the value chain through the idea of a personified lead corporation that should be responsible for ‘its’ value chain. Second, a *collective actor perspective* characterizes EU consumer law and related market regulation. In this institutionalization, the value chain evolves as an enumerated collective of actors contributing to the production processes and collectively responsible for the value chain operations with hierarchization of responsibility in line with the actors that are viewed as governing the network. Third, an entirely *de-personified institutionalization* occurs in EU market practices and trade law in which the produced objects and the trading practices prevalent in GVCs are singled out and regulated in relation to territory and business relations.”

modified and adapted in light of the technological development of artificial intelligent systems; (iii) an interrelated network of concepts regarding the assessment of damage and effects may be more effective than the lexical definitions of autonomy that are situated at a higher level of abstraction; (iv) this modification can be made by elaborating on top of the findings already reached by researchers in Law & Technology and AI & Law, and by figuring out a specific way of selecting and structuring the components of our own scaffolding for risk analysis.

Hence: (a) it is not necessary to extend the classical concept of legal personality to artificial intelligent systems; (b) it is sufficient to focus on the relationship between the concepts of *legal autonomy*, *legal governance*, *legal agency*, *moral agency*, *delegated agency*, *harm*, and *liability*. We should be able to figure out solutions to the analysis of risks arising from the autonomy of AISs reducing—not increasing—its complexity. To do so, it is necessary to contextualize and specify: (1) the level of abstraction of the analysis; (2) the different layers and degrees of autonomy of the systems to be evaluated; (3) the autonomy in the organizational structure of the agents; (4) the kind of regulatory and legal ecosystem in which agents operate. We have suggested replacing the classic levels of autonomy of AIS based on H/M interaction—driver, user, driver, doctor, user, corporate director—with four levels of complexity and information processing: (i) Instrumental delegation; (ii) teleological autonomy; (iii) domain expansion (hybrid intelligence, reusable architecture, open-domain competence); (iv) complex actions.

Finally, we contended that the use of AISs can enrich the legal field by *fleshing it out from within*, impinging on its efficiency, and contributing to enrich, imbue and align its ethical values. We have suggested at the end a very simple framework—based on delegated autonomy and moral agency—to start developing the analysis of risks by using artificial intelligence itself to control and monitor AISs. This involves proposing (i) the ethical governance of AI, (ii) managing risks and promoting values and AI value chains in a proactive (not reactive) way and, therefore, (iii) its alignment with critical thinking.

However, it is worth mentioning that the global legal and regulatory environment for AI is a complex mosaic of hard law (binding regulations), soft law (non-binding principles and frameworks), standards, protocols, policies, and ethics. This has resulted in a fragmented *compliance patchwork* that creates significant challenges for multinational corporations and public administrations, reflecting (i) a fundamental divergence in regulatory philosophies, (i) partially based on the raise of political global conflicts. The analysis of the AI value chain and its legal and regulatory aspects reveals a landscape defined by *complexity* and *fragmentation*. The core challenge lies in the rapid evolution of a technology that is fundamentally changing how value is created, while legal governance frameworks struggle to adapt. The AI value chain's multi-layered nature, market concentration, and reliance on continuously learning models undermine traditional legal principles of liability, intellectual property, and legal governance. This is the main reason for fostering the emergence of legal and ethical smart ecosystems based on the conscious reconstruction and redesign of AI value chains.

Acknowledgements.

Authors wish to acknowledge the contributions and fruitful discussions with Louis de Koker, Mustafa Hashmi, and Louis de Koker. Research for his paper is supported by CSIC's (Bilateral Collaboration Initiative i-LINK-TEC) project DESAFIA2030 BILTC22005; EU (Horizon-EIC-2021-Pathfinderchallenges-01); Project VALAWAI 101070930; the EU (Next Generation EU/PRTR program); the Spanish (MCIN/AEI-10.13039-501100011033 program) project VAE TED2021-131295B-C31; the Catalan Government SGR (Group of Excellence) Project 00536; and Proyecto-CIPROM/2022/26 "Presente y future de la regulaci3n de los Criptoactivos en la UE [Legalcripto]". (Generalitat Valenciana, P.I., Carmen Pastor). We acknowledge the use of Gemini to help us identify some differences in the fields of Ethics, Common law, and Civil law; and to find some more examples to illustrate the levels of autonomy we proposed for AISs (online institutions) and the difficulties around AI value chains.

16.7 References

- Agboola, B. (2025). Artificial Intelligence and Intellectual Property: Adapting Legal Frameworks for Innovation. Available at SSRN 5284590. Accessible at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5284590 (accessed 20/08/2025).
- Andrews, C. (2025). European Commission withdraws AI Liability Directive from consideration. 12 February. <https://iapp.org/news/a/european-commission-withdraws-ai-liability-directive-from-consideration>
- Attard-Frost, B. and Gray Widder, D. (2025). "The ethics of AI value chains." *Big Data & Society* 12 (2): 20539517251340603.
- Beckers, A. (2023). Global value chains in EU law. *Yearbook of European Law*, 42: ,322–346 <https://doi.org/10.1093/yel/yead010>.
- Berlin, I. (1969). "Two concepts of liberty" (1959). In: Isaiah Berlin, *Four Essays On Liberty*, Oxford: Oxford University Press, pp. 118-172.
- Casanovas P (2024). Building a Smart Legal Ecosystem for Industry 5.0, in W Barfield, Y-H Weng, U Pagallo (eds.), Cambridge Handbook on Law, Policy, and Regulations for Human-Robot Interaction, Cambridge University Press, Cambridge, pp. 145-168.
- Casanovas, P. (2025a). A Regulatory Framework for Legal Ecosystems in the Context of Emerging Web-Based Systems and the European AI Value Chain Regulations. In: C. Pastor (ed.) *Governance and Control of Data and Digital Economy in the European Single Market*, Cham: Springer, pp. 23-53.
- Casanovas, P. (2025b). From AI Risks to Legal and Ethical AI Governance: A Four-Dimension Framework. *The De Gruyter Handbook on Law and Digital Technologies*, edited by M. Durante and U. Pagallo, De Gruyter, pp. 251-278. <https://doi.org/10.1515/9783111346632-013>
- Casanovas, P., and Noriega, P. (2022). Dilemmas in legal governance. *J. Open Access L.*, 10, 1.
- Casanovas, P. and Hashmi, M. (2024), Report on the OPTIMAI Regulatory Model 4th version. Deliverable 9.8. 30 June, Optimizing Manufacturing Processes through Artificial Intelligence and Virtualization (OPTIMAI), Grant agreement ID: 958264, 166 pp.
- Casanovas P, Hashmi M, de Koker L, Lam H-P (2025). Compliance, Regtech, and Smart Legal Ecosystems: A Methodology for Legal Governance Validation. In: W Barfield, U Pagallo (eds.), *Research Handbook on the Law of Artificial Intelligence*, vol. II, Edward Elgar Publ., Cheltenham (UK), Northampton (MA), pp. 73-104.
- Castelfranchi, C. (2000). "Founding Agent's 'Autonomy' on Dependence Theory". In: *Proceedings of the 14th European Conference on Artificial Intelligence (ECAI 2000)*, Berlin, Germany: IOS Press, pp. 353–357.
- Castelfranchi, C. and Falcone, R. (2004). "Founding Autonomy: The Dialectics Between (Social) Environment and Agent's Architecture and Powers". In M. Nickles, M. Rovatsos, and G. Weiss (Eds.), *Agents and Computational Autonomy. Potential, Risks, and Solutions*, LNAI 2969, Berlin, Heidelberg: Springer, pp. 40-54.
- Castelfranchi, C., Cesta, A., Conte, R., & Miceli, M. (2005). "Foundations for interaction: The dependence theory". In P. Torasso (ed.) *Third Congress of the Italian Association for Artificial Intelligence* (1993). LNAI 728. Berlin, Heidelberg: Springer, pp. 59-64.
- Chopra, S., & White, L. F. (2011). A legal theory for autonomous artificial agents. University of Michigan Press.

Christman, John (2020). "Autonomy in Moral and Political Philosophy", *The Stanford Encyclopedia of Philosophy* (Fall 2020 Edition), Edward N. Zalta (ed.), <https://plato.stanford.edu/archives/fall2020/entries/autonomy-moral/> (accessed 14/08/2025).

De Koker L, Casanovas P (2024). De-risking' Denials of Bank Services: An Over-Compliance Dilemma? In: de Koker L, Goldbarsht D (eds.), *Financial Crime, Law and Governance. Navigating Challenges in Different Contexts*. Springer International Publishing, Cham.

De Lucia, M. J., Newcomb, A., and Kott, A (2019). "Features and operation of an autonomous agent for cyber defense." *arXiv preprint arXiv:1905.05253* (accessed 13/08/2025).

Defense Science Board (DSB) (2016). *Defense Science Board Summer Study on Autonomy*, Office of the Under Secretary of Defense for Acquisition, Technology and Logistics. AD1017790. Authors: David, R.A. and Nielsen, P.. Washington, July. Available at: <https://apps.dtic.mil/ti/citations/AD1017790> (accessed 12/08/2025).

Drukarch, H. and Fosch-Villaronga, E. (2022). The role and legal implications of autonomy in AI-driven boardrooms. In B. Custers and E. Fosch-Villaronga, *Law and Artificial Intelligence: Regulating AI and Applying AI in Legal Practice*, The Hague: TMC Asser Press, pp. 345-364.

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

Dworkin, G. (2008). "The concept of autonomy", *Grazer Philosophische Studien*, 12: 203-213.

Dworkin, G. (2015). "The nature of autonomy", *Nordic Journal of Studies in Educational Policy*, 2: 28479. DOI: 10.3402/nstep.v1.28479 (accessed 9/08/2025).

Dworkin, R. (1986). "Autonomy and the demented self". *The Milbank Quarterly*, 64 (2): 4-16.

Dworkin, R. (2000). *Sovereign Virtue: The Theory and Practice of Equality*, Cambridge, MA: Harvard University Press.

Engler, Alex C., and Andrea Renda, A. (2023). *Reconciling the AI Value Chain with the EU's Artificial Intelligence Act*. CEPS.

EU (2017). European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)).

EU Expert Group on Liability and New Technologies (2019). *New Technologies Formation Liability For Artificial Intelligence and Other Emerging Digital Technologies*.

EU Report (2020), Brussels, COM (2020) 64 final. 19.2.2020, Report from the Commission to the European Parliament, the Council and The European Economic and Social Committee. Report on the safety and liability implications of Artificial Intelligence, the Internet of Things and robotics.

EU (2022). Proposal for a Directive of The European Parliament and of The Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive), Brussels, 28.9.2022, COM(2022) 496 final 2022/0303 (COD).

EU (2023). Regulation (Eu) 2023/1114 of The European Parliament and of The Council of 31 May 2023. on markets in crypto-assets, and amending Regulations (EU) No 1093/2010 and (EU) No 1095/2010 and Directives 2013/36/EU and (EU) 2019/1937.

EU (2024). Artificial Intelligence Act. European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonized rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)).

- Federal Court of Australia (2022). Commissioner of Patents v Thaler [2022] FCAFC 62 Appeal from: Thaler v Commissioner of Patents [2021] FCA 879, VID 496 of 2021, Judgment of: Allsop CJ, Nicholas, Yates, Moshinsky and Burley JJ, 3 April 2022. Accessible at: <https://www.judgments.fedcourt.gov.au/judgments/Judgments/fca/full/2022/2022fcafc0062> (Accessed 25/08/2025).
- Feinberg, J. (1989). "Autonomy". In: J. Christman, *The Inner Citadel: Essays on Individual Autonomy*. Oxford: Oxford University Press, pp. 27-53.
- Finegan, T. (2015). "Dworkin on equality, autonomy and authenticity". *The American Journal of Jurisprudence*, 60 (2): 143-180.
- Floridi, L. and Sanders, J.W. (2004). *Minds and Machines* 14: 349-379.
- Floridi, L. (ed.) (2015). *The Onlife Manifesto. Being Human in an Hyperconnected Era*. Cham: Springer Open.
- Floridi, L., and Cows, J. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*, Issue 1.1, Summer 2019,
- Floridi, L., Cows, J., King, T. C., and Taddeo, M. (2021). How to design AI for social good: Seven essential factors. In *Ethics, governance, and policies in artificial intelligence* (pp. 125-151). Cham: Springer, pp. 125-151.
- Fuller, L. (1941). "Consideration and form". *Columbia Law Review*, May 1941 (5): 799-824.
- Fuller, L. (1969). *The Morality of Law* (1964). Second Edition. New Haven: Cambridge University Press.
- Gelati, J., Rotolo, A., Sartor, G., and Governatori, G. (2004). Normative autonomy and normative co-ordination: Declarative power, representation, and mandate. *Artificial Intelligence and Law*, 12(1): 53-81.
- Guarino, N. (1995). "Formal ontology, conceptual analysis and knowledge representation". *International Journal of Human-computer Studies*, 43 (5-6): 625-640.
- Heintze, H.J. (1998). "On the legal understanding of autonomy." In *Autonomy: Applications and implications*, M. Suksi (ed.), *Autonomy: Applications and Implications*. The Hague: Kluwer, pp. 7-32.
- Jiménez-Schlegel, P., Pareto, J., Torras, C. (2025). Robot ethics: Assessing risk from the perspective of autonomy. Chapter 8 in this same volume.
- Kelsen, H. (1955). "Foundations of democracy." *Ethics* 66 (1, part II): 1-101.
- Kelsen, H. (1970). *Reine Rechtslehre* (1960). *The Pure theory of law* (1970). Translated by Max Knight. Berkeley: University of California Press.
- Kennedy, D. (2000). "From the will theory to the principle of private autonomy: Lon Fuller's Consideration and Form". *Columbia Law Review*, 100: 94-175.
- Kurki, V. A. (2019). *A theory of legal personhood*. Oxford: Oxford University Press.
- Laukyte, M. (2017). "Artificial agents among us: Should we recognize them as agents proper?". *Ethics and Information Technology*, 19 (1): 1-17.
- Lapidoth, R. (1994). "Autonomy: Potential and Limitations", *International Journal on Group Rights* 1 (4): (1994), 269-290.
- Lavrchenko, M. (2022). Thaler v. Vidal: Artificial Intelligence-Can the Invented Become the Inventor. *CARDOZO L. REV.*, 44, 699-736.

Lehuedé (2025). An elemental ethics for artificial intelligence: water as resistance within AI's value chain. *AI & SOCIETY* 40 (3): 1761-1774.

Liu, J., Jiang, X., Shi, M., and Yang, Y. (2024). Impact of artificial intelligence on manufacturing industry global value chain position. *Sustainability*, 16 (3): 1341.

Llewellyn, K.N. and Hoebel, E.A. (1941) *The Cheyenne Way: Conflict and Case Law in Primitive Jurisprudence*. Norman: University of Oklahoma Press.

Margetis, G., Apostolakis, K.C., Dimitriou, N., Tzovaras, D., and Stephanidis, C. Aligning Emerging Technologies onto I4.0 principles: Towards a Novel Architecture for Zero-defect Manufacturing. In *2022 IEEE 27th International Conference on Emerging Technologies and Factory Automation, (ETFA)*, pp. 1–8. IEEE, 2022, pp. 1-8.

Martin, S. (2025). Legal Personhood - Bundle Theory, 11th Aug 2025, <https://www.lesswrong.com/posts/58e8EycHHGMYxiaoo/the-bundle-theory-of-legal-personhood> (accessed 21/08/2025)

Monroe, S., Luck, M. (2004). "Agent autonomy through the 3 M motivational taxonomy". In M. Nickles, M. Rovatsos, and G. Weiss (Eds.), *Agents and Computational Autonomy. Potential, Risks, and Solutions*, LNAI 2969, Berlin, Heidelberg: Springer, pp. 55-67.

National Highway and Transportation Safety Administration (NHTSA) (2018a). Levels of Automation. Available at: <https://www.nhtsa.gov/document/levels-automation> (accessed 15/08/2025)

National Highway and Transportation Safety Administration (NHTSA) (2018b). A Framework for Automated Driving System Testable Cases and Scenarios. DOT HS 812 623, September 2018 Available at: https://www.nhtsa.gov/sites/nhtsa.gov/files/documents/13882-automateddrivingsystems_092618_v1a_tag.pdf (accessed 15/08/2025)

Nieves, C. (2022). Stephen Thaler v. Katherine K. Vidal, 43 F. 4th 1207 (Fed. Cir. 2022). *Intell. Prop. & Tech. LJ*, 27, 85.

Noriega, Pablo. (2024) A Naive View of Electronic Institutions, in N. Osman (ed.) *Electronic Institutions: Applications to uHelp, WeCurate and PeerLearn*, Springer International Publishing, pp. 3–22.

Noriega P, Verhagen H, Padget J, d'Inverno M (2016). A manifesto for conscientious design of hybrid online social systems. In: *Coordination, Organizations, Institutions, and Norms in Agent Systems XII*, Springer, Cham, p 60-78.

Noriega P, Verhagen H, Padget J, d'Inverno M (2023). Addressing the Value Alignment Problem Through Online Institutions. In: *International Workshop on Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems*, Springer Nature, Cham, p 77-94.

Noriega, P., and Casanovas, P. (2022). La gobernanza de los sistemas artificiales inteligentes. In M. Martín Serrano, O. Velarde (eds.). *Mirando hacia el futuro: cambios sociohistóricos vinculados a la virtualización*. Madrid: Centro de Investigaciones Sociológicas (CIS), pp. 115-143.

Noriega, P., and Casanovas, P. (2024). From the Pascal Wager to Value Engineering: A Glance at AI Risks and How to Address Them. In *International Workshop on Value Engineering in AI* (pp. 257-275). Cham: Springer Nature Switzerland, pp. 257-275.

Noriega, P. and Plaza, E (2025). Four Settings and a Proposal; for an AI-inspired Theory of Values, in this same volume.

- Novelli, C., Floridi, L., Sartor, G. and Teubner, G. (2024). AI as Legal Persons: Past, Patterns, and Prospects (November 24, 2024). Available at SSRN: <https://ssrn.com/abstract=5032265> (accessed 15/08/2025).
- Organisation for Economic Co-operation and Development, OECD (2024). Explanatory Memorandum on the updated OECD definition of an AI system. OECD Artificial Intelligence Papers, March 2024, no. 8. Available at: https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/03/explanatory-memorandum-on-the-updated-oecd-definition-of-an-ai-system_3c815e51/623da898-en.pdf (accessed 15/09/2025).
- Orwat, C., Bareis, J., Folberth, A., Jahnel, J., and Wadehul, C. (2024). Normative challenges of risk regulation of artificial intelligence. *NanoEthics*, 18 (2): 11.
- Pagallo, U. (2013). *The Law of Robots*, Cham: Springer.
- Pagallo, U. (2018). Vital, Sophia, and Co.—the quest for the legal personhood of robots. *Information*, 9(9): 230, doi:10.3390/info9090230.
- Pagallo, U., Aurucci, P., Casanovas, P., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Schafer, B. and Valcke, P. (2019a). AI4People-on good AI governance: 14 priority actions, a SMART model of governance, and a regulatory toolbox. Accessible at: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://ai4people.org/PDF/AI4People_On_Good_AI_Governance.pdf (accessed 15/08/2025).
- Pagallo, U., Casanovas, P., & Madelin, R. (2019b). The middle-out approach: assessing models of legal governance in data protection, artificial intelligence, and the web of data. *The Theory and Practice of Legislation*, 7(1): 1-25.
- Pagallo, U. (2025a). Three paradigms in the legal governance of AI: On power, convenience, and prestige. In *Research Handbook on the Law of Artificial Intelligence*. Edward Elgar Publishing, pp. 57-72.
- Pagallo, U. (2025b). Pagallo, U. On Twelve Shades of Green: Assessing the Levels of Environmental Protection in the Artificial Intelligence Act. *Minds and Machines*, 35(1): 10. <https://doi.org/10.1007/s11023-025-09713-4> (accessed 15/08/2025).
- Poblet, M, Casanovas, P., Rodríguez-Doncel, V. (2019). *Linked Democracy: Foundations, tools, and applications* (p. 130). Cham: Springer Nature. <https://link.springer.com/book/10.1007/978-3-030-13363-4>
- Raz, J. (1986). *The Morality of Freedom*, Oxford: Clarendon Press.
- Reyes, J. A. P. and Rajagopal, A. (2025). Rules of engagement: ethical issues and value chain introspection in Artificial Intelligence systems. *Quality & Quantity*, 59 (Suppl 1): 463-487.
- Ross, A. (1952). *Why Democracy?* Cambridge, Mass.: Harvard University Press.
- Ross, A. (2019). *On Law and Justice* (1953). Translated from the Danish version by Uta Bindreiter. Oxford: Oxford University Press.
- Ross, A. (1957). “Tû-Tû”, *Harvard Law Review*, March 70 (5): 812-825.
- Ross, A. (1968). *Directives and Norms*, London: Routledge and Kegan Paul.
- Russell, S. (2023). *Human compatible: AI and the problem of control* (2019). London: Penguin.
- Russell, S. J., and Norvig, P. (1995). *Artificial intelligence: A modern approach; [the intelligent agent book]* (pp. I-XXVIII). N.Y.: Prentice Hall.
- SAE International. (2016). *J3016: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*. Warrendale, PA: Author.

- Saravanan, A. and Prasad, D. (2024). AI as an Inventor Debate under the Patent Law: A Post-DABUS Comparative Analysis (August 01, 2024). *European Intellectual Property Review*, Volume 47 (1). pp. 26-39 (forthcoming), Available at SSRN: <https://ssrn.com/abstract=5053108> or <http://dx.doi.org/10.2139/ssrn.5053108>
- Sartor, G. (2014). The autonomy of automated weapons. In: L. Glorioso, A.M. Osula (Eds.) *Ist Workshop on Ethics of Cyber Conflict Proceedings*. Tallinn: NATO Cooperative Cyber Defence Centre of Excellence.
- Sartor, G., and Omicini, A. (2016). "The autonomy of technological systems and responsibilities for their use". In: N. Bhuta, S. Beck, R. Geiß, H-Y Liu, C.Kreß (eds.), *Autonomous Weapon Systems. Law, Ethics, Policy*. Cambridge University Press, pp. 39-74.
- Sartor, G. (2020). Contracts Without Agreement, or Agreements by Artefacts?, *Akademie der Wissenschaften zu Göttingen (Hg.) Digitalisierung. Privatheit und öffentlicher Raum*, 41-46.
- Sartor, G., Oddi, A., Rasconi, R., Santucci, V. G., and Meo, R. (2024). Synthesizing Evolving Symbolic Representations for Autonomous Systems. *arXiv preprint arXiv:2409.11756*.
- Sellers, M. (2008). "An Introduction to the Value of Autonomy in Law". In: Sellers, M. (eds) *Autonomy. Comparative Perspectives on Law and Justice*, Dordrecht: Springer, pp. 1-9.
- Shattuck, L. G. (2015). "Transitioning to Autonomy: A Human Systems Integration Perspective". Human Systems Integration Program Naval Postgraduate School Monterey, CA., NASA Workshop *Transitioning to Autonomy: Changes in the Role of Humans in Air Transportation*. March 10-12, 2015. Available at: <https://humansystems.arc.nasa.gov/workshop/autonomy/agenda.php> (accessed 13/08/2025).
- Solum, L.B. (2020). Legal personhood for artificial intelligences. In W. Wallach, P. Asaro (eds.) *Machine ethics and robot ethics*, London: Routledge, pp. 415-471.
- Stanton-Ife, John (2022). "The Limits of Law", *The Stanford Encyclopedia of Philosophy* (Spring 2022 Edition), Edward N. Zalta (ed.), <https://plato.stanford.edu/archives/spr2022/entries/law-limits/> (accessed 14/08/2025)
- Thaler, S. and Zbikowski, K. (2017). "Cognitive Engines Contemplating Themselves: A Conversation with S. L. Thaler", *Apa Newsletter | Philosophy and Computers* 17 (1): 21-26.
- Teubner, G. (2006). Rights of non-humans? Electronic agents and animals as new actors in politics and law. *Journal of Law and Society*, 33 (4): 497-521.
- Twining, W. (2012). *Karl Llewellyn and the Realistic Movement* (1973). Second edition. Cambridge: Cambridge University Press.
- UNIDO (2015). Global Value Chains and Development. UNIDO's Support towards Inclusive and Sustainable Industrial Development. Accessible at: https://www.unido.org/sites/default/files/2016-03/GVC_REPORT_FINAL_0.pdf (accessed 25/08/2025).
- United States Court of Appeals for the Federal Circuit (2022). Stephen Thaler, Plaintiff-Appellant V. Katherine K. Vidal, Under Secretary of Commerce for Intellectual Property and Director of The United States Patent and Trademark Office, United States Patent and Trademark Office, Defendants-Appellees, 2021-2347, Appeal from the United States District Court for the Eastern District of Virginia in No. 1:20-cv-00903-LMBTCB, Judge Leonie M. Brinkema, Decided: August 5, 2022. Accessible at: <https://www.wipo.int/wipolex/en/judgments/details/2098> (accessed 20/08/2025).
- Vargo, S. L., and Lusch, R.F. (2016). "Institutions and axioms: an extension and update of service-dominant logic." *Journal of the Academy of marketing Science* 44 (1): 5-23.
- Wang, S., Wan, J., Zhang, D., Li, D., and Zhang, C., Towards Smart Factory for Industry 4.0: A Self-organized Multi-agent System with Big Data-based Feedback and Coordination. *Computer Networks* 101, 158–168, 2016.

Winick, B.J. (1992). "On Autonomy: Legal and Psychological Perspectives". *Villanova Law Review* 37 (6): 1705-1777. Available at: <https://digitalcommons.law.villanova.edu/vlr/vol37/iss6/5> (accessed 15/08/2025).

Yang, G.-Z, J. Cambias, K. Cleary, E. Daimler, J. Drake, P. E. Dupont, N. Hata, P. Kazanzides, S. Martel, R. V. Patel, V. J. Santos, R. H. Taylor (2017). Medical robotics—Regulatory, ethical, and legal considerations for increasing levels of autonomy. *Sci. Robot.* 2, eaam8638.