AI

IMPACT FACTOR 6.5

CITESCORE 7.3

Essay

Agentic AI and the Algorithm of Good

Alkis Gounaris and George Kosteletos

Special Issue
AI for Good

Edited by
Prof. Dr. Manolya Kavakli-Thorne

MDPI

https://doi.org/10.3390/ai7090352

AI

MDPI

Essay

Agentic AI and the Algorithm of Good

Alkis Gounaris [1,2,3,†] **and George Kosteletos** [1,2,3,*,†]

[1] Department of Philosophy, School of Philosophy, Zografou Campus, National and Kapodistrian University of Athens (NKUA), Zografou, 157 72 Athens, Greece; alkisg@philosophy.uoa.gr
[2] High-Level Expert Group on Artificial Intelligence of the Hellenic Republic National Commission for Bioethics and Technoethics, 106 74 Athens, Greece
[3] NKUA Applied Philosophy Research Lab, Zografou, 157 72 Athens, Greece
* Correspondence: gkosteletos@philosophy.uoa.gr
† These authors contributed equally to this work.

Abstract

This chapter offers a conceptual and philosophical analysis of the ontological, epistemological, and ethical issues associated with the design, development, and use of Agentic AI in moral and legal decision-making. Its aim is to map, classify, and bring into focus the principal philosophical problems raised by the autonomous operation of artificial agents in contexts of moral and legal judgement. The chapter contributes to the existing literature by distinguishing between two levels of AI autonomy. Advisory systems possess limited autonomy and provide recommendations that remain subject to human evaluation, whereas regulatory systems exercise a greater degree of autonomy and are entrusted with final decision-making authority. Adopting a sceptical perspective, the analysis examines the conceptual fragility and epistemological difficulties surrounding proposals to employ Agentic AI in significant domains. Particular attention is given to the risk that ostensibly advisory systems may become tacitly regulatory in practice, especially under the influence of widespread assumptions concerning the objectivity, accuracy, and effectiveness of AI. The inquiry proceeds through a step-by-step branching structure organised around three central questions concerning the moral values on the basis of which such systems should be designed and developed, the extent to which they can understand the concepts of ethics and justice, and the attribution of responsibility for their decisions and outputs. Each question opens onto alternative lines of analysis, thereby providing a systematic map of the principal philosophical challenges involved. Its purpose is to provide a synthetic overview of the relevant philosophical issues while highlighting both their breadth and their plurality. To organise these issues systematically, the chapter adopts the schema of an algorithm, offering a clear and ordered account of the possible interconnections among the problems examined and of the ways in which alternative answers generate distinct yet interrelated lines of philosophical inquiry.

Keywords: AI for good; agentic AI; philosophy of AI; AI ethics; advisory AI systems; regulatory AI systems; machine ethics; ethical decision-making; legal decision-making; autonomy

Academic Editor: Ke Zhang

Received: 11 May 2026
Revised: 7 August 2026
Accepted: 31 August 2026
Published: 8 September 2026

1. Introduction: Prolegomena to an Algorithm of Good

1.1. AI for Good: Towards a Good AI Society

Artificial Intelligence (AI) is a techno-scientific enterprise whose formal point of departure is usually traced to the 1950s [1], although both its theoretical and its technological

origins have been identified at an even earlier stage [2,3] (for a history of AI, see also: [4]). However, although discussion of both the risks and the benefits of AI has likewise been underway for several decades [5] (for the early discussion about the risks and the benefits of AI see, for example, [5,6]), it has intensified only recently, in light of the latest striking advances in AI research and the equally striking expansion of its possible applications [7] (p. 4).

This renewed focus concerns, to a significant extent, proposals for a clear and universally acceptable ethical and legal framework for the regulation of AI. One might even speak of an effort towards the institutionalisation of AI ethics (by the term "institutionalisation" we mean the establishment of institutional organs, committees, councils, and other bodies with an administrative or advisory role in matters concerning the regulation of science and technology. We therefore borrow this term from Casiraghi [8]), the ultimate aim of which is not merely to secure human well-being by protecting it from AI, but to promote human flourishing through AI. With this approach, AI is understood not as a threat to humanity, but as a means of advancing human eudaimonia; it is not conceived as something that arrives "from outside" in order to impose itself upon human beings, but rather as an integral component of human flourishing. Within the broader aspiration for a socially beneficial AI, a number of regulatory frameworks have been proposed and developed, including the Asilomar AI Principles (2017) [9], the IEEE's Ethically Aligned Design (2017) [10], the Montréal Declaration for Responsible AI (2017) [11], the Five Overarching Principles for an AI Code of the UK House of Lords (2018) [12], the Tenets of the Partnership on AI (2018) [13], and the Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems issued by the European Commission's European Group on Ethics in Science and New Technologies (2018) [14], which helped prepare the ground for the recently adopted European Union Regulation on AI, widely known as the AI Act [15] (for a comprehensive and critical mapping of AI ethics principles proposed worldwide, see [16]. For an analysis and grouping of proposed AI ethics principles, see [17–19]. For a critical analysis of the EU AI Act, see [20]).

The aim of an AI that would promote human well-being, ultimately constituting an integral dimension of it, has recently been articulated under a range of labels, including "AI for People (AI4People)" and "the Good AI Society" [7,18,21], "AI for Social Good (AI4SG)" [21–26], "Human-Centred AI" [27–29], and "AI for Good" (see current volume). In the text that follows, and for consistency within this volume, we will use the term "AI for Good."

At present, however, only a relatively small proportion of the overall academic literature on AI addresses questions related to the aspiration for an "AI for Good"—that is, an AI directed towards the promotion of human well-being [30]. At the same time, significant gaps have been identified between the theoretical articulation of AI ethics principles and their practical implementation [16].

AI remains an open horizon of both positive and negative possibilities. Accordingly, the path towards an AI that may serve as a means to human well-being—that is, the path towards an AI for Good—requires both careful anticipation of possible risks and an a priori, clear, and stable commitment to identifying and embodying its potential benefits. As Floridi et al. [18] rightly observe, the scale of the impact AI may have on society calls for regulatory initiatives informed by a careful risk–benefit analysis. Humanity must seize the opportunities while, at the same time, anticipating and minimising the risks, thereby effectively confirming what Floridi et al. describe as the "dual advantage of AI Ethics" [18] (pp. 690, 691, 694). Indeed, the acceptance and adoption of AI by human beings will occur only if both its benefits and the means by which its risks may be minimised are presented in a reasonable and intelligible manner (ibid., p. 694). We consider this observation especially important, since we believe that establishing AI as a means of promoting human

well-being presupposes AI acceptability (it is worth noting that achieving the goal of AI acceptability presupposes, inter alia, the mobilisation of empirical methods for recording the psycho-social factors that govern human acceptance of different uses of AI. See, for example, refs. [31–36]). This, in turn, presupposes the cultivation of an AI ethics awareness, beginning already during the school years through the targeted incorporation of AI ethics education into curricula [37]. Such education should make clear both the benefits and the risks of AI, as well as the ways in which the latter may be minimised, so that a possible AI ethics anxiety may be prevented or effectively addressed [38–41].

Finally, we consider the discussion of Entangled AI relevant to this context, as it is a conceptual framework in which AI is connected and aligned with human, social, and ecological systems, rather than existing as a separate system. Entangled AI design emphasises that these systems should be designed to foster a dynamic balance between technological advancement, human well-being, and environmental sustainability [42]. Of course, the design, development, and configuration of an entangled AI also require the development of entangled narratives regarding the phenomenon of AI [43] so that entanglement can be achieved in terms of optimal intercultural inclusion [44–46].

Against this background, the present chapter contributes to the literature on AI for Good by examining a specific yet foundational problem within this broader agenda, namely the design, development, and use of autonomous AI systems for ethical decision-making. Although the scope may initially appear broad, the combined examination of ontological, epistemological, and ethical issues reflects the interdependence of the problems involved. It therefore provides both a philosophical roadmap and a theoretical basis for identifying the principal obstacles, risks, and unresolved conditions associated with this endeavour.

1.2. On Agentic AI and Degrees of Automation

Degrees of automation in the design, development, and use of AI vary according to the extent of human involvement in their operation [47,48]. We may thus distinguish cases in which the human agent (a) fully and continuously controls the functioning of AI systems and is therefore in the loop; (b) supervises their operation and may retain the power either to select among the system's outputs or to exercise the "final say", including the capacity to suspend the system's choices or ultimate action, and is therefore on the loop; or (c) does not intervene in the operation and action of the systems, with the result that the system functions in full autonomy. In such cases, we speak of conditions in which the human is off the loop. Clarifying the degree of autonomy exhibited by AI systems is a crucial factor in the ethical assessment of their conduct [49].

By Agentic AI we refer to autonomous AI systems that are goal-directed and capable of planning, computing, and executing complex tasks with minimal, or even no, human intervention [50–52]. Such systems, which may be supported by LLMs and other AI subsystems and tools, are also able to learn continuously, adapt to dynamic environments, and design and execute complex workflows. Although these systems are expected to have a wide range of applications in areas such as the correction of computational errors, telecommunications, programming, data analysis, interaction across heterogeneous software environments, and work in remote or inaccessible settings, they are also expected gradually to assume an important role in matters of security, health-related judgement, justice, labour, and related domains [53,54].

In such a case, the prospect of developing computational systems for regulatory purposes—and, more specifically, AI systems capable of making decisions of a moral or legal kind—constitutes a strong and entirely plausible trajectory in the evolution of Agentic AI [55,56] (for a comparison between conventional AI decision-making systems and Agentic AI and for an understanding of the new dynamics rising in the field of human

decision-making by the deployment of Agentic AI, see [57,58]. For the psychological factors dictating people's trust on the use of AI in courts, see [59]).

It should be clarified, however, that decision-making per se does not, in the first instance, constitute the acute problem. AI applications already in everyday use routinely make decisions that may have an indirect ethical impact; yet the criteria by which those decisions are made are not themselves moral or legal, and their role remains, in essence, advisory. We encounter such advisory AI systems on a daily basis when using a road-navigation application or a subscription television platform. These, however, differ from systems for the assessment of human behaviour or activity, health-assessment systems, automated personal-data processing systems, educational recommendation systems, personnel-management systems, asylum and migration management systems, and legal advisory systems—all of which are classified as high-risk systems under the AI Act [15].

Beyond these everyday advisory applications of AI, there are already emerging uses of Agentic AI designed to support and enhance human ethical decision-making in what may be described as sensitive domains of judgement, such as healthcare [60]. For the time being, such applications remain purely advisory in character: these systems merely offer recommendations, while the human agent remains the final evaluator and adjudicator of those recommendations and, above all, the one who ultimately makes the decision. We may therefore refer to such systems as advisory AI systems.

However, as a possible future development of these advisory AI systems, it is highly plausible that regulatory AI systems will emerge—systems that would not merely recommend, but would themselves make final decisions on the basis of moral or legal criteria. Such systems may, in the near future, be deployed in functions such as the administration of justice, the formulation of governmental or corporate policy, or even the assessment of everyday personal or professional conduct.

It should further be clarified that, in contrast to regulatory AI systems, advisory AI systems, such as the legal applications currently in use (for an overview of the potential applications of AI in the field of justice and related ethical issues, see [61]. For more information specifically on the use of LLMs, see [62]. For the factors influencing public acceptance of these applications, see [63]), produce recommendations which the human agent interacting with them is under no obligation to follow. In the case of regulatory AI systems, however, their outputs may take the form of decisions with which human beings would be required to comply strictly, even on the basis of an institutional framework that formally mandates alignment with the decisions of such systems.

The choice as to whether these outputs are to take the form of mere recommendations or binding decisions amounts, in effect, to a choice concerning the degree of autonomy that will characterise such systems and, correspondingly, the degree of control that human beings will retain over them. At this point, it is necessary to distinguish between the *technical autonomy* and *the decisional autonomy* of AI systems, as a system may be technically autonomous while remaining merely advisory in decisional terms. Technical autonomy refers to the extent to which a system can operate and generate an output—in our case, a decision—without human intervention. This form of autonomy is determined by the system's technical architecture, specifications, and functional capabilities. By contrast, decisional autonomy concerns the extent to which the system's output remains subject to human supervision, evaluation, and potential revision. This kind of autonomy is institutionally determined. It arises through institutional authorisation and reflects not the technical characteristics of the AI system itself but the type of relationship that society chooses to establish between human decision-makers and AI systems. More specifically, AI's decisional autonomy stands in an inverse relationship to the degree of human control over decisions that shape people's lives. The greater the decisional autonomy granted

to an AI system, the more limited the role of human oversight and intervention in the decision-making process. Ultimately, decisional autonomy is bounded by institutional and legal arrangements, which determine whether the output of an AI system is to be treated merely as a recommendation or as a legally binding decision. If the outputs amount merely to recommendations, then the regulatory AI system exhibits only limited decisional autonomy, since any recommendation, precisely as such, still calls for final evaluation and, ultimately, for approval or rejection by the human user—for example, a legislator or a judge. We are then in an on-the-decisional loop condition. By contrast, if the outputs constitute decisions that are to be executed as binding, the role of the human agent—for example, the legislator or the judge—becomes devoid of substantive content, and the AI system thereby enjoys full autonomy. This is an off-the-decisional loop condition.

However, there is a risk that even ostensibly advisory AI systems may, in practice, become tacitly regulatory AI systems, since their high degree of effectiveness, their capacity to perform complex calculations, and the gradual entrenchment of their use may leave users with little real alternative but to endorse, ratify, and ultimately follow their "recommendations". In this sense, the institutional dimension of decisional autonomy may be overridden by a human epistemic and psychological positivity bias toward AI, rooted in the belief in the inherent superiority of AI systems. Under such circumstances, although the AI system is institutionally designed to possess no decisional autonomy and therefore is *de jure* classified as merely advisory, it may, in practice (*de facto*), assume a regulatory role because human decision-makers effectively remove themselves from the decisional loop. Accordingly, the transition from advisory AI to regulatory AI becomes possible precisely because of the distinction between the system's de jure and de facto status. It represents a departure from the AI system's originally institutionalised role, determined not by any modification of its technical characteristics or legal classification, but solely by how human actors ultimately choose to interact with it *in practice*—and, more specifically, by whether *in practice* they keep themselves within or remove themselves from the decisional loop. As will be shown in Section 2.2, this departure is supported by extensive empirical evidence.

In the present essay, we seek to show that the decision to develop such regulatory systems presupposes engagement with a series of ontological, epistemological, and ethical philosophical questions. Without addressing these questions, there is a danger that such systems will be assigned goals that either fail to align with human priorities or cannot be adequately controlled.

The philosophical debate surrounding advisory AI systems and regulatory AI systems is highly complex, since it involves interwoven ethical, epistemological, and ontological issues. The aim of the present inquiry is not to offer an exhaustive analysis of these matters, but rather to provide a schematic account of the way in which they arise and intersect. What we attempt, in other words, is a mapping of the philosophical concerns associated with the regulatory use of AI—concerns which ultimately reveal either difficulties that are, in many cases, extremely hard to resolve, or assumptions of a distinctly hazardous kind, all of which would have to be taken seriously by anyone seeking to construct and deploy regulatory AI systems for moral and legal decision-making.

In an effort to systematise and present this multi-layered discussion in a simple and intelligible form, we designed a diagram (Figure 1), which we symbolically entitled the "Algorithm of Good" (it should be noted that the "Algorithm of Good" is not an executable computational model, but a schematic philosophical map of the branching questions that arise once Agentic AI is considered for advisory or regulatory use in moral and legal decision-making) and which sets out, in algorithmic form, the principal steps of the present inquiry.

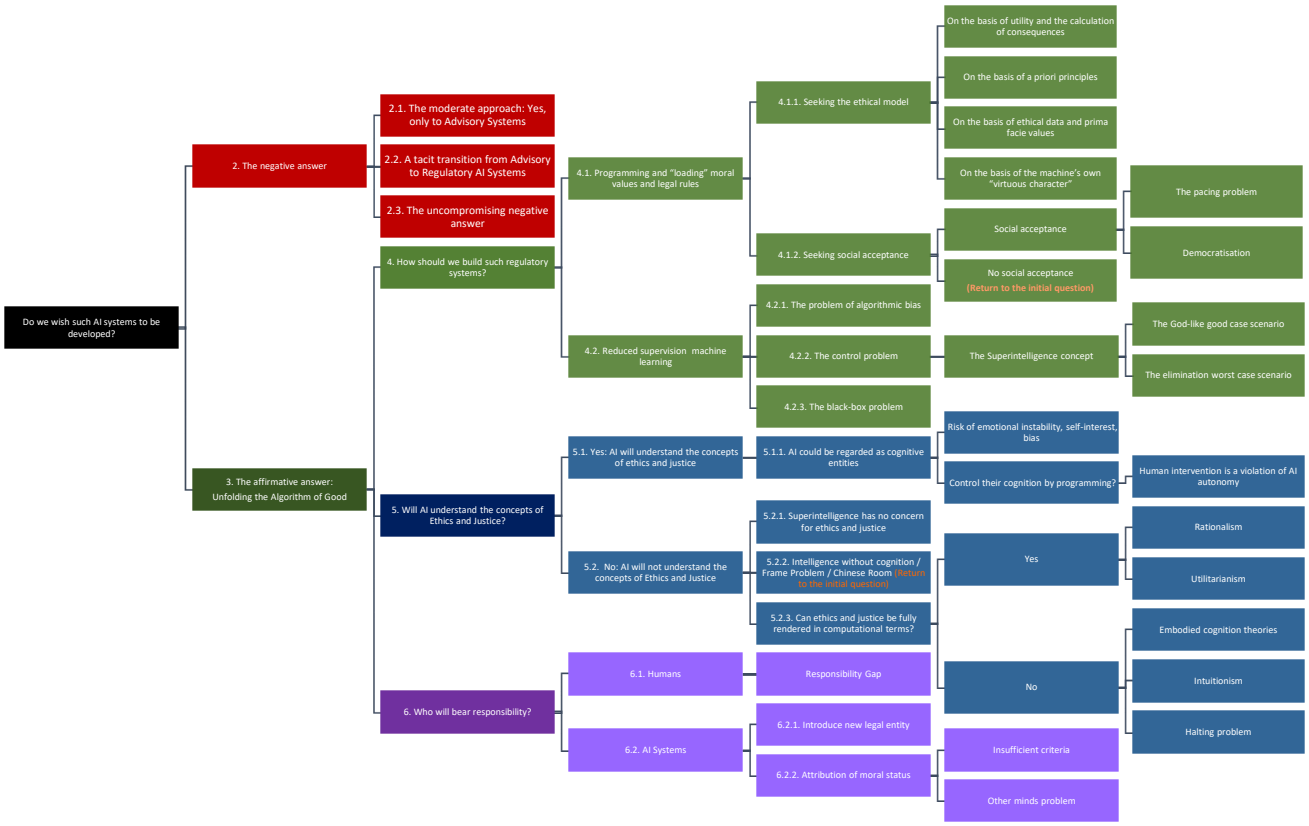

Figure 1. Under the metaphorical title Algorithm of Good, the figure presents a philosophical and conceptual analysis of the theoretical problems associated with the ontological, epistemological, and ethical dimensions of the design, development, and use of regulatory AI systems. In this sense, the Algorithm of Good functions as a conceptual map and a step-by-step guide to the inquiry developed in this chapter. The figure presents the central questions of the analysis and traces the alternative answers to them, together with the philosophical problems arising from each branch. Each node corresponds to a section of the chapter that follows. The figure therefore serves as a classificatory conceptual map of the principal problems addressed in the present inquiry.

Beginning of the Algorithm

Our algorithm begins with the fundamental question of whether or not we wish such AI systems to be developed.

2. The Negative Answer

2.1. The Moderate Approach: Yes, Only to Advisory Systems

A persuasive "No" may be summarised in the position advanced by researchers such as Yudkowsky [64], who hold that the aim of AI research ought to be the construction of "friendly" and safe machine-tools fully aligned with human needs and goals. On this view, what we require is not regulatory AI systems—that is, AI systems that make decisions autonomously—nor superintelligent machines, but useful machines that improve our lives by recommending or calculating what is required in each case, without themselves proceeding to "decisions" and moral acts. What is at stake, in essence, is the treatment of AI as a tool: as an advisory system which, precisely as such, both is and ought to remain under full human control. Those who support this approach evidently feel secure in the conviction that the development of AI computational systems as "mere tools" guarantees the non-autonomy of the latter and, consequently, the autonomy of the human being. In effect, we are speaking here of human-in-the-loop conditions or, where the human factor is only partially disengaged from the process of control, of human-on-the-loop conditions.

In support of this line of argument against the development of regulatory AI systems, one might also invoke the broader precautionary approach to technology as expressed by UNESCO [65], according to which, before a technological application that may have catastrophic consequences for humanity is used, those wishing to implement it must demonstrate in a credible manner that it is safe, or else identify the conditions under which it may be used safely.

Given, however, that an AI system subject to the absolute and continuous control of human users would not differ greatly from ordinary computational systems, proponents of the view of "AI as a tool" may well regard the human-on-the-loop arrangement as the optimal solution, attracted by the prospect of developing more effective tools while partially relieving human agents of workload in the interests of saving time and effort.

2.2. A Tacit Transition from Advisory to Regulatory AI Systems

Even this moderate approach, however, may prove detrimental to human autonomy. More specifically, as noted above, there is a risk that such advisory systems will ultimately come to be treated as autonomous (in decisional terms) regulatory systems, since human beings—captivated by the apparent precision and objectivity of AI—may become increasingly reluctant to question its recommendations. In this way, the original intention of preserving human control risks producing the opposite outcome, one in which the human judge is reduced to a merely formal executor of AI-generated proposals. The possibility that such advisory AI systems may, in the end, be treated by their human final assessors as regulatory AI systems is already beginning to emerge in the field of justice. Human judges (or juries), operating under the assumption of the mathematical accuracy and computational power associated with AI, may hesitate to challenge or depart from the recommendations issued by such systems (for this concern in relation to automated bail and risk-assessment software, see [61]). Thus, while formally we may take ourselves to be in a human-on-the-loop condition, in practice we may already have shifted into a human-off-the-loop condition (for the phenomenon of "positive bias toward AI," see [66–68]. A typical manifestation of this phenomenon is the unconditional trust that many parents of children with learning and neurodegenerative disorders place in the robot therapists with whom their children interact [69]), and the autonomy of the human adjudicator becomes merely apparent.

In the Human–Computer Interaction literature the phenomenon of humans' overreliance on the recommendations of automated systems is known as automation bias (AB). More precisely, AB is a cognitive [70] or—more broadly—psychological phenomenon [71] expressed via an inclination of humans to uncritically and almost 'automatically' [72–74] accept computer generated decisions or recommendations [73,75–79] beyond the extent to which they should [80] and often overlooking conflicting human judgements or contradictory evidence [81,82]. Up to a point, one could say that AB is the result of positivity bias toward AI—namely the belief in the inherent superiority of AI [72]. In general, AB has been attributed to several factors; for instance, user related factors such as skill, problem awareness, experience, familiarity with and knowledge of the automated systems and related technologies [80,81]; psychological states, personality traits and cognitive strategies [83], cognitive laziness [72,84] trust and confidence [70,80]; social factors like social environment, effectiveness of training and increased decision accountability [83]; environmental and contextual factors such as excessive workload, task complexity and cognitive overload, limitations in time and cognitive resources [70,80,83,85]; and system related factors such as accuracy, type of information, the way in which information is presented [83], data collection interoperability and intransparency [81]. With respect to the latter, Bond et al. [85] stress on the link between AB and the black box nature of many AI systems and Stefan

Strauß identifies as a basic cause of AB "the lack of options to scrutinise how a system generated a certain outcome", and thus the lack of transparency, accountability, explicability and interpretability [74]. According to Strauß, challenges are usually due to machine learning and deep learning features like adaptability and unsupervised or self-learning capabilities. In this case we may even refer to a deep automation bias [74].

At this point, it is important to emphasise that apart from the ample anecdotal evidence of AB in different automation applications like self-driving cars and commercial flights [72,84,86], the phenomenon of AB has been empirically demonstrated by a number of studies that have traced and recorded it in a series of human activity domains like healthcare [85,87–91] (for a review of empirical studies tracing AB in different healthcare applications, see [81,92]. For a systematic review highlighting the theme of automation bias in healthcare, see [93]), national security [80], military uses of autonomous systems [94,95] and military academy training [96] aviators' training [97], personnel selection [73] and employment decisions [72], street level bureaucrats like police officers [98], song popularity forecasting and music recommendations [99], as well as among the general adult population [100]. However, the existing experimental research on AB is not evenly distributed across different domains, and several of these domains remain understudied. Evidence from a review of 35 empirical studies indicates that healthcare has received the greatest research attention among domains investigating AB (n = 11, 31, 4%), followed by the less studied domains of national security (n = 5), public administration (n = 3), finance (n = 3), human resources (n = 2) and mental healthcare (n = 2) [70]. Most importantly, there are still crucial domains of human activity that remain without any empirical research on possible AB influence. For instance, Tilbury and Flowerday emphasise on the lack of empirical research on the possible effects of AB among Security Operations Centre (SOC) personnel [101]. Finally, it is important to note that, beyond the existing gaps in the empirical research on AB, significant challenges also persist with respect to its legal and regulatory governance, as current regulatory frameworks neither adequately account for AB nor provide a comprehensive framework for its assessment and mitigation. For a critical analysis, see [71,83].

Under the pressure exerted by the aura of objectivity and effectiveness that accompanies almost every AI application, the human decision-maker ultimately places himself or herself off the decisional loop. In this way, advisory AI systems come, tacitly, to be treated as regulatory AI systems. At this point, we would like to clarify that, by emphasising the possibility of such a tacit transition, our intention is not to advance a normative claim about institutional design, but rather to highlight a potential conceptual risk and identify an observed sociological tendency.

Indeed, one might observe that human autonomy is, ironically, undermined by precisely that which ought to constitute a basic condition for its preservation. In this respect, within the AI4People framework [7,18], after examining in depth the six most influential regulatory frameworks for AI (more specifically, these six frameworks, as noted above: [9–14]), Floridi and his collaborators identified their principal points of convergence and ultimately proposed a unified framework consisting of five core principles (in essence, these amounted to the four principles of bioethics autonomy, beneficence, non-maleficence, and justice, together with the principle of explicability [102]), among which was the principle of autonomy (the principle of autonomy is recognised as one of the most important principles within the framework of Value Sensitive Design, both with regard to information technology in general [103,104] and to AI in particular [26]). With regard to this principle, they noted that, when we adopt AI and its intelligent agency, we consciously cede to AI systems a portion of our control over decision-making, and that affirming the principle of autonomy in the context of AI requires the achievement of a balance between the decision-making

power we retain for ourselves and that which we assign to artificial agents [7] (p. 7). Floridi et al. [18] (p. 698), are, in effect, pointing to an inevitable trade-off between the degree of human autonomy (and hence control over AI) that we wish to preserve and the benefits that may arise from limiting that autonomy (and hence from relaxing our control over AI while increasing the system's own autonomy). They thus acknowledge that there are circumstances in which compelling considerations—such as effectiveness—may outweigh the loss of human control over the decision-making process. Of particular interest, however, is their further observation that even if we accept such a trade-off, and thus proceed to some degree of restriction of human autonomy (that is, to some relaxation of control over AI), this must occur in a reversible manner: we must retain the possibility of fully recovering human autonomy, and therefore of fully regaining control over AI, by revoking our original decision to delegate decisional autonomy to it (ibid.). Accordingly, for Floridi et al., what truly safeguards and secures human autonomy is not only the capacity to choose, but also the capacity to revoke one's choices; or, as they characteristically put it, "deciding to decide again" (ibid.). They describe this capacity as "meta-autonomy", or as the "decide-to-delegate" model.

We take this capacity for meta-autonomy to have two dimensions: (a) A practical dimension which considers meta-autonomy as a technical possibility. AI systems must be designed in such a way as to allow human beings to reassume control over them; that is, any recovery of control by human agents must be technically feasible. In this respect, one might speak of incorporating the principle of meta-autonomy into the aims of Value Sensitive Design for an "AI for Good"—for example, by adding it to the set of values identified by Umbrello and Van de Poel [26]. (b) A cognitive and specifically a reflective dimension which concerns the preservation of meta-autonomy as a cognitive capacity. If human beings are to be able to revoke any prior decision—for example, a decision to curtail their own autonomy and relax their control over AI—then they must be capable of reflecting upon, reassessing, and justifying their decisions; in effect, they must be able to decide about their decisions. In order for this to be possible, human beings must first possess the capacity to think about their decisions, that is, the capacity for reflection. Reflection is generally regarded as inherent to human beings, as a constitutive and distinctive aspect of human cognition. On the basis of the considerations advanced here, reflection appears to be the cognitive foundation of meta-autonomy, which, in turn, is identified by Floridi et al. [18] as the safeguard, or safety net, of human autonomy against the coercive powers of AI.

At this point, however, particular importance attaches to our earlier observation concerning the tacit tendency of human beings to treat advisory AI systems as if they were regulatory AI systems, under the pressure of a positivity bias towards AI. As noted above, human evaluators, for example, judges, juries, etc., may, under the influence of the widespread assumption that AI is objective, computationally powerful, and ultimately more effective, hesitate to challenge the recommendations of an AI system and thereby place themselves off the decisional loop. In such a case, human beings will themselves have undermined their own capacity to remain the ones who actually make the relevant decision; by ceding decision-making authority to AI systems, they will, through their own stance, have transformed advisory systems into regulatory ones.

Under this pressure of a widespread positivity bias towards AI in public opinion, human evaluators, in fact, choose to stop judging; they decide not to decide. In this case, we could speak of an ironic function of meta-autonomy. Meta-autonomy turns against autonomy. Indeed, this ironic operation of meta-autonomy arises from the conjunction of the human capacity for reflection and a prevailing AI positivity bias. So long as such a bias remains the dominant tendency in public opinion, those occupying positions of judgement will, in taking that tendency into account; that is, in considering the circumstances and

the broader evaluative framework within which they are called upon to decide, they will choose to delegate decision-making to machines. Meta-autonomy, as the capacity to decide which decisions one will take, thus leads human beings, in this instance, to the decision not to take any decision at all, since, through reflection, they apprehend the wider context of AI positivity bias and decide not to oppose it. Beyond the positivity bias, being relieved of the burden of decision-making fundamentally alters human agency. Sartre [105] points out that, since we are called upon to make decisions at every moment of our lives, the weight of those decisions ultimately determines who we are. Unlike the things around us, which embody the purpose for which they were made, humans define our own purpose. By making decisions and taking action throughout our lives, we ultimately determine our very being. In this sense, as Sartre emphasised in his lecture, humans are condemned to be free. When we delegate decision-making power to AI systems, we relinquish part of our freedom and ultimately cease to be what we would have been if we had made the decision ourselves.

The full significance of such a development becomes clearer if one considers that the human adjudicators who ultimately decide not to decide have been placed in that role institutionally, through decisions of the State. At least where that State is democratic, their placement in such a role expresses the will of the majority of citizens. Two observations follow. First, a paradoxical condition may arise in which citizens are themselves shaped by an AI positivity bias, with the result that human adjudicators are pressured to discharge the very duty of decision-making that those same citizens have institutionally entrusted to them. We thus face a situation in which citizens' lay beliefs stand in tension with their own institutionalised will. This is a paradox that strikes at the rational foundation of contemporary democratic societies, and does so in a manner that remains largely invisible to citizens. Second, human adjudicators do not disclose their decision to place themselves off the loop and not to perform the task entrusted to them. They therefore appear to decide, while in reality they align themselves unconditionally with the recommendations of the AI system. The State appears to remain on the loop, whereas in fact it has silently slipped into an off-the-loop condition.

At this point, we would like to emphasise and particularly highlight the catalytic role of the AI positivity bias in this ironic operation of meta-autonomy. Based on the previous analysis, if AI positivity bias were not pervasive in public opinion, the human adjudicator's reflection would not lead them to exercise meta-autonomy against autonomy itself. The adjudicator would not reflect under the psychological pressure exerted by a public opinion and a dominant evaluative framework which, in practice, call upon them not to resist the recommendation of the advisory AI system.

Moreover, this positivity bias towards AI need not be conceived as an exclusively "external" pressure upon the human adjudicator. It may just as well form part of that adjudicator's own convictions. In other words, it is entirely possible for human adjudicators to decide to place themselves off the decisional loop because they genuinely believe that the recommendations of the AI system are the best possible ones. In either case—whether as pressure generated by public opinion or as a personal conviction—AI positivity bias appears to constitute a necessary condition for the ironic operation of meta-autonomy.

It thus becomes clear how important it is to counteract AI positivity bias by cultivating an AI awareness that would illuminate not only the risks and benefits of AI, but also the limits of its effectiveness. At this point, the need for organised educational and training programmes on AI once again comes sharply into view. In our view, such programmes should certainly equip those undertaking them to identify and understand both the risks and the benefits of AI, thereby reinforcing what Floridi et al. describe as the "dual advantage of AI Ethics" (article 9 [18] (pp. 690, 691, 694)). At the same time, we believe that, in light of

the foregoing analysis, it is equally clear that such programmes must also provide learners with an informed understanding of the actual technical capacities and limitations of existing AI technologies.

In conclusion, one might respond affirmatively to half of the initial question, "Do we want the development and use of advisory and regulatory AI systems?" and, more specifically, accept the moderate approach to development and use of advisory AI systems, on the condition that sufficiently robust and appropriately designed AI education programmes are in place, so as to prevent the possibility of a tacit transition from advisory AI systems to regulatory AI systems.

2.3. The Uncompromising Negative Answer

By contrast, more radical positions—such as those associated with neo-Luddism—reject the development of AI altogether, regarding it as a threat to human labour and autonomy [106]. Taking the school environment as a point of departure [107], and opposing the introduction of AI into education, critics of the new technology advocate the development of local initiatives and the building of networks of resistance [108], following the example of a radical minority of students in the United States who organise through the Luddite Club and reject the use of smartphones, social media, and the extensive use of the internet [109].

Such positions, however, quite apart from their marginal character, tend to constrain the dialogue and to discourage the exploration of solutions that might secure human control, thereby leaving room for more moderate approaches to AI development—approaches that incorporate safeguards against full autonomy.

By contrast, the answer "Yes" to the question of regulatory AI systems may express a broad spectrum of "optimistic", pro-technology, positivist, liberal, and proactive positions which foreground the advantages and benefits of technological progress while downplaying the possible risks arising from an unpredictable development of AI [110–112].

For example, Michael and Susan Anderson, pioneers of the Machine Ethics programme [113], maintain that we should indeed aim at the construction of such regulatory AI systems—or, in other words, of moral machines. Their argument by analogy may be summarised as follows: there already exists a wide range of AI systems that perform many tasks better than human beings. The success of these systems suggests that we should also attempt to build moral machines—machines that would ultimately prove more just, more effective, less biased, more impartial, and so forth; that is to say, machines that would administer justice as genuinely "moral" agents [114], free from human weaknesses [115].

A persuasive "Yes", however, opens before us a set of three difficult questions (see Figure 1), each of which constitutes the point of departure for a distinct and complex branch of analysis concerning the development and use of regulatory AI systems.

3. The Affirmative Answer: Unfolding the Algorithm of Good

What follows is an analysis of three questions that arise from the possible decision to develop such regulatory systems, or to accept the risk of using advisory systems for moral decision-making—systems which, as we have seen, may nonetheless become regulatory despite our intentions:

1. How are we to build such regulatory systems?
2. Will these systems understand the concepts of morality and justice?
3. Who will bear responsibility for their actions and decisions?

Each of these questions gives rise to alternative answers, which in turn generate further questions, as the subsequent analysis will show.

4. How Should We Build Such Regulatory Systems?

To address the first question—namely, how to construct an AI system that is both moral and regulatory—we need to consider technical, conceptual, and ethical issues that will determine its ontological status. Two broad approaches can be distinguished, depending on the role assigned to human agents in shaping the system's decision-making.

The first approach (Section 4.1), "supervised learning" [116] (p. 695), involves the explicit specification of rules by the designer: the desired outputs are stipulated in advance, thereby securing a higher degree of predictability in the system's decisions.

The second approach (Section 4.2) relies on "reduced supervision methods", in which the system detects patterns or learns through reward signals, without human guidance regarding outputs (ibid, p. 830). This approach may yield decisions that exceed the limits of human foreseeability and, potentially, give rise to powerful self-training AI systems which, although still under active research, may open new pathways in AI.

4.1. Programming and "Loading" Moral Values and Legal Rules

If we choose to "load" some set of values into a system either through explicit rule-based programming, in which norms and inferential rules are directly encoded [117] or through supervised learning, in which mappings are learned from labelled examples [116,118], we must address two subsidiary questions: (a) which ethical model will be used to "train" the AI system to make decisions (Section 4.1.1), and (b) whether the specific system will enjoy broad institutional and social acceptance (Section 4.1.2).

4.1.1. Seeking the Ethical Model

It is self-evident that, if we want an AI system to issue "evaluative" moral decisions, it must first be trained within a determinate framework of rules. Moreover, in order to avoid the logically fallacious conclusions that might arise from deriving an "ought" from purely descriptive "is" premises, the rule framework on which we base the training of our system must include evaluative "major premises". Such premises will draw their guiding commitments from established ethical models that articulate "how we ought to act", specifically either utilitarianism (Section 4.1.1: On the Basis of Utility and the Calculation of Consequences) or deontology (Section 4.1.1: On the Basis of a Priori Principles). More recent approaches to the problem of machine moral training propose recourse to ethical models that recognise prima facie values and duties (Section 4.1.1: On the Basis of Ethical Data and Prima Facie Values), or that recognise "virtuous traits" (Section 4.1.1: On the Basis of the Machine's Own "Virtuous Character"). Yet, as we shall see below, these approaches do not overcome the conceptual and practical problems that pervade the present essay. Yudkowsky [119] has pointedly observed that the selection of the rule framework on which we base the training of a system is itself a problem that demands more sustained attention. For although the system may indeed produce outputs consistent with what we have programmed, it remains doubtful whether those outputs will adequately capture our initial intentions and aims. This is because it is impossible to anticipate in advance every possible scenario and every possible "output" to which a system may be led when it follows the rule framework under which it was trained. Accordingly, even if we construct an inductive supervised-learning algorithm for moral values, we can never be certain that, in the long run—or in the hypothetical case of superintelligent machines—the computational morality of utilitarianism or the rigid morality of deontology will not yield decisions that are "catastrophic or repugnant" for humanity, as is often alleged in the case of utilitarianism [120].

On the Basis of Utility and the Calculation of Consequences

Utilitarianism, as an ethical model, prescribes that actions should aim at the greatest possible benefit for the greatest number of people, a commitment that appears to align with the computational character of AI systems. Jeremy Bentham [121] advanced a "felicific" calculus of utility structured around seven criteria, suggesting that benefits can be calculated with precision, thereby facilitating moral decision-making (see also Section 5.2.3. Utilitarianism). Yet the "quantification" of morality does not cover every case. Bentham's student, John Stuart Mill [122], added the need for a qualitative distinction between pleasures and for recognising certain values that exceed mere calculation. Mill maintained that the human species ought to choose human happiness even when that entails less pleasure, thereby placing limits on the computability of moral decisions.

Despite its apparent usefulness, utilitarianism faces a further difficulty: the precise calculation of an action's future consequences is often unattainable, since those consequences are unpredictable and never fully certain. This, in turn, constrains the objectivity and accuracy with which utilitarian reasoning can be operationalised in AI applications.

On the Basis of a Priori Principles

The deontological approach, in contrast to utilitarianism, focuses on intention rather than on the consequences of an action. For Kant [123], moral judgement is grounded in a pure will—a will not governed by expediency or inclination—which acts from duty under the authority of the moral law as expressed in the Categorical Imperative, i.e., a universally binding principle. Thus, in the familiar example, the honest merchant acts out of duty, not out of fear of adverse consequences.

Applying a Kantian model to a regulatory AI system is, however, difficult, since the Kantian idea of a free and autonomous will—a will that gives the law to itself—is not readily translatable into computational systems that rely on formal programming and externally specified objectives. Accordingly, an AI system that operates through implicit rules oriented towards limited instrumental aims cannot, in any literal sense, realise the Kantian ideal of autonomy. Moreover, given its potentially "rigid" design, such a system may exhibit inflexibility, failing to accommodate exceptions and morally complex situations, and thereby risking inhumane outcomes in particular cases.

On the Basis of Ethical Data and Prima Facie Values

More recent philosophers, such as W. D. Ross [124], sought objective criteria for morality in order to overcome the weaknesses of utilitarianism and Kantian deontology. Ross proposed an ethical model grounded in prima facie values—self-evident moral principles such as fidelity, beneficence, and non-maleficence—which function as moral "data", in a way analogous to the sensory information an organism registers. These principles operate as prima facie duties that guide moral choice while still allowing consequences to be taken into account. In principle, such values could be "loaded" into an AI system so as to steer its decision-making.

This, however, does not eliminate moral dilemmas—for instance, when the requirement of non-maleficence conflicts with the need to protect individuals who are under threat. In such cases, the system may be forced to choose between broadly utilitarian and broadly deontological considerations. Despite these limitations, prima facie values appear promising for the training of supervised and reinforcement-learning AI systems, insofar as they provide a structured form of moral guidance for the decisions such systems deliver.

On the Basis of the Machine's Own "Virtuous Character"

An Aristotelian approach to the moral behaviour of regulatory AI systems focuses on their virtuous character, rather than merely on a decision-making "recipe". Instead of asking only how moral decisions are to be produced, the question becomes whether some criterion X could make the system itself moral or virtuous agent [125]. Aristotle held that ethics is not exhausted by discrete actions, but is rooted in a cultivated and stable character (hexis) expressed in the virtues. On this view, a regulatory system would have to serve both a social and an individual end, and to cultivate virtues such as phronēsis (practical wisdom), justice, trustworthiness, and epieikeia (equity–leniency) (the notion of equity or epieikeia often translated as decency, leniency, or mildness, was coined by Aristotle in Nicomachean Ethics as a necessary correction to the inflexibility of universality of law. For more on Aristotelian equity, see Section 5 below), thereby exhibiting patterns of behaviour recognisable as moral. Despite the technical challenges involved, the hypothesis of a "virtuous regulatory system" situates such a machine within the social fabric: it is to be assessed not only by the achievement of its internal aim, but—more decisively—by the degree to which the principles and virtues it embodies secure social uptake and endorsement. Ultimately, the system would be judged by its broad acceptance as behaving virtuously, with its success registered in the public and institutional sphere. Within this framework, a Virtuous Agent Seal of Excellence has been proposed for adoption [126], assessing and certifying systems ex post based on their performance and the degree of alignment with individual and social ends and with the specific purpose for which each system was developed.

4.1.2. Seeking Social Acceptance

In one of the largest experiments on moral beliefs across cultures—collecting more than 40 million responses from 233 countries—the public was asked to indicate the preferred moral criteria by which an AI system, such as an autonomous vehicle, ought to make decisions [127]. The outcome of this social experiment, known as the "Moral Machine" [128], revealed the wide range of cross-cultural moral diversity as a function of social, economic, and geographical parameters. In the present context, the question of what kinds of moral values we should "load" into an AI system is directly connected to these findings, which suggest that the adoption of any single model of moral decision-making would not be universally acceptable. The further question therefore arises whether the programming of a regulatory AI system should require broad institutional and social acceptance, and, if so, whether such acceptance ought to be sought at the level of cultural context, nation-state, or morals of local community [129].

Social Acceptance

If we take the view that the value-loading of a regulatory AI system ought to enjoy broad institutional and social acceptance, we immediately confront two problems that are difficult to resolve. The first is the problem of divergence between the pace of technological development and the pace of institutional response, as described by Collingridge, a difficulty that pervades research across the field of AI. The second is the problem of social consensus, which shifts the discussion towards the so-called democratisation of AI [130], an issue that has become increasingly prominent in public debate.

- Pacing Problem

In the early 1980s, David Collingridge [131] highlighted the divergence between, on the one hand, our knowledge of the impact of each new technological development and, on the other, the timely adoption of appropriate measures to provide institutional coverage for its economic, political, and social consequences. Collingridge observed that, in the early phases of technological development, a technology remains malleable—it can

still be adjusted, redirected, or even replaced—and the cost of controlling or prohibiting it is comparatively low. Yet, at precisely this stage, its real effects—social, economic, and environmental—are difficult, and sometimes impossible, to foresee. At later stages, however, once the consequences of its use become clearer, the technology is often so deeply embedded in everyday life and in the economy that it is either difficult or impossible to change; or, if change remains feasible, it can be achieved only at a high economic, social, and frequently political cost.

This temporal lag between the pace of technological innovation and the pace at which institutions adopt or adapt regulatory measures—often described as the pacing problem [132]—has appeared historically in a range of technological "turns", such as the reliance on lignite, the widespread adoption of plastics, or the internal combustion engine. Today, however, the accelerating pace of AI development, combined with the slower pace of institutional response, renders the pacing problem even more acute, thereby increasing the risk of serious and potentially irreversible consequences. This is due primarily to the exponential growth of computational capacity (Moore's law) [133] and to the rapid diffusion of AI systems across almost every domain of human life, in ways that no institutional framework can fully and accurately anticipate in terms of their long-term effects.

- Democratisation

One might argue that the very widespread use of AI systems legitimises their development and, in itself, amounts to a form of broad consent and social acceptance. Yet how informed is this "tacit" consent on the part of users? Advocates of the democratisation of AI maintain that extensive uptake, popularity, and accessibility are only one component in determining the "democratic" character of a technological medium—assuming, that is, that users genuinely understand what they are doing. The other component concerns the ends and the work a given AI system performs: whether it includes specific social groups, advances social justice, and provides transparency either about how it operates or about the purposes it serves. More generally, proponents of the need to democratise technology [130] (for the call for the democratisation of technology, see [134]) and AI in particular, argue that democratic values and ideals ought to be built into AI systems and taken into account at the level of design. However, this presupposes institutional and regulatory reflexes [135] which, as past experience has shown (and as noted above), tend to lag behind technological developments.

No Social Acceptance

If we answer in the negative the question raised in Section 4.1.2, namely, whether the loading of a particular model of values, principles, or virtues into a regulatory AI system ought to command broad institutional and social acceptance, then we must return to the initial question and ask again whether we truly want regulatory AI systems at all.

4.2. Reduced Supervision Machine Learning

We referred above to the argument from the exceptional effectiveness of AI systems in general as a criterion for the development of regulatory AI systems, since, according to Anderson et al. [115] such systems may be able to administer justice without the human bias or miscarriages of justice that are frequently observed today. One possible route to self-training would be for systems to "learn", on their own, in a broad and non-technical sense, the moral values or decision patterns they are expected to follow. Such learning may involve different technical paradigms, including unsupervised or self-supervised representation learning [136,137], reinforcement learning [138], learning from human feedback [139], imitation learning [140], and meta-learning [141]. These approaches must be distinguished from one another while unsupervised learning identifies patterns or structures in unla-

belled data, whereas reinforcement learning optimises a policy through interaction with an environment and feedback provided by a reward signal. Learning from human feedback similarly relies on a reward or preference signal, but one that is supplied directly by human judgement rather than derived from the environment itself. Imitation learning, in turn, trains a system to reproduce demonstrated behaviour rather than to discover a policy through trial and error. Meta-learning, finally, trains a model across multiple tasks so that it can adapt more rapidly to new ones.

It is worth noting that, of these five paradigms, only the first and the last (unsupervised/self-supervised learning and meta-learning) substantially reduce direct human supervision over the content of what is learned; learning from human feedback and imitation learning remain, by design, closely dependent on human-supplied signals, and are discussed here because they illustrate the same underlying point developed below, that even ostensibly autonomous learning is rarely free of human specification. Therefore, the phrase "on their own" must be understood in a broader sense. In reinforcement learning for example, the development of novel strategies in games such as Go is more accurately understood as an achievement of reinforcement learning and self-play, rather than of unsupervised learning. AlphaGo Zero, in the case of Go, was trained without data derived from human expert games [142], but it was not devoid of human specification while the rules of the game, the winning condition, the architecture, the learning procedure, and the evaluative signal were all predetermined. The system learned strategies that had not been demonstrated to it by human players, but it did so within a clearly specified task and reward structure.

Despite the positive results in certain applications, the possibility that AI systems may develop "unpredictable" strategies can generate serious concern, especially where such strategies diverge from socially acceptable ends. This is commonly described as the alignment problem [5,143], which arises when the initial learning conditions or the reward function fail to track the intended goal. As Bostrom has argued via the orthogonality thesis [144], a system's computational capability and its tendency to pursue certain goals under some value specification should be treated as two independent variables. In other words, high computational competence does not guarantee that a system will serve moral or socially acceptable objectives. To address the alignment of final goals, researchers have proposed strategies such as learning by imitating human behaviour [145,146] and constraining the rate at which a system's values are updated and recalibrated [147]. However, imitation of human behaviour is not, by itself, a safety guarantee, since human behaviour is not invariably an appropriate ideal for AI systems. Cases such as Microsoft's Tay chatbot [148], which developed extreme biases and inappropriate behaviour, underscore the risks associated with poorly controlled learning.

Recent adaptive-learning methods—including test-time adaptation, task-specific meta-regularisation, and feature-regularised meta-learning with channel-wise attention expansion (FRCAE) enhance few-shot adaptation, robustness, and cross-domain generalisation [141,149,150]. Yet their tasks, features, objectives, and constraints remain human-specified. Statistical consistency between support and query distributions cannot, by itself, establish normative validity. A system may generalise a pattern accurately and consistently even where the pattern reflects historically embedded bias, an unjust institutional practice, or an inadequately specified moral objective. Techniques such as feature regularisation may improve robustness and generalisation, but they cannot determine whether the features and objectives being regularised are morally justified.

4.2.1. The Problem of Algorithmic Bias

In 2016, an evaluation study of AI-based tools used in legal decision-making revealed a high degree of algorithmic bias, against population on race, nationality and skin colour

criteria, in the assessment of recidivism risk (for bias in predictive modelling systems in the U.S. correctional system, see [151]. For bias in crime prediction, see [152]). More specifically, the algorithms tended to produce harsher risk assessments for Black defendants, as a result of biases embedded in the training data. Efforts to mitigate such bias include removing or reweighting biased data, as well as deploying approaches such as counterfactual fairness (for counterfactual fairness, see [153,154]), which compares algorithmic decisions against a "counterfactual" world in which demographic attributes—such as race—are varied. Nonetheless, under unsupervised learning and with limited human intervention, the possibility of bias re-emerging remains, especially where a system straightforwardly imitates human behaviour.

4.2.2. The Control Problem

The Control Problem is linked not only to how AI systems process information but also to their widespread deployment, which can render them practically ungovernable. For instance, applications like Google Maps or deepfake technologies are so widespread that effective control becomes extremely difficult. A more extreme scenario involves the existential risk that might be posed by a superintelligence [144], a vastly superior cognitive entity capable of escaping meaningful human control.

The Superintelligence Concept

The Superintelligence concept is often linked to the concept of the technological "singularity" [155–157], where a machine attains such advanced capabilities that its behaviour becomes essentially unpredictable. According to pessimists like Yudkowsky [64], while the evolution of human intelligence took millions of years, AI might only need seconds to surpass the human species.

The Future of Life Institute [158] and Max Tegmark [159], examine scenarios of coexistence with a superintelligence. A *good case scenario* is one that superintelligence protects human eudaimonia and has a positive impact on society. This scenario may be considered good, in the sense that, although superintelligence would act as a universal and unchanging regulatory force in human life, almost a god-like presence, the continued existence of the human species is guaranteed and, potentially, enhanced.

By contrast, the *worst-case scenario* considers the possibility that a superintelligence takes control and comes to regard the human species as a threat to its own "ecosystem" and to the security of its vital resources, with the result that it eliminates humanity as we know it—by means we cannot foresee (for more scenarios involving catastrophic AI, see [160]).

4.2.3. The Black Box Problem

A first step towards controlling machines is understanding how a regulatory system reaches a conclusion. Tegmark (ibid) emphasises that, although training a system on data and using deep learning can be highly effective, this is not enough for justice or regulatory decision-making. The current challenge is that deep learning systems often function as "black boxes" [161–163] (for an overview of efforts to address the black box problem to date, see [164]. For a critical discussion of the black box problem, see [165]), in the sense that it is unclear how, precisely, their data are analysed and used to produce their outputs. The European Union [15] has prioritised the regulation of this issue as a precondition for safe and trustworthy AI.

However, Anderson and Anderson [113] (pp. 5–15) argue that what is required is an accountable and justified explanation of the decision rather than transparency of the underlying process. By analogy with human judges, if a system's decision conforms to human principles, then its outputs may be regarded as acceptable. This mode of assessment, an "Ethical Turing Test" [115] focuses on results rather than on procedure. Even so, judging

behaviour alone is not sufficient for classifying a system as a "moral entity". As we shall see below, solving the problem of value-loading or enabling the learning of rules via machine learning, and addressing computational transparency and behavioural criteria, are important steps but not sufficient for treating a regulatory system as a "moral entity". For such a system to be acceptable in the role of judge or legislator, it would have to understand the concept of morality and the concept of justice.

5. Will AI Understand the Concepts of Ethics and Justice?

Whether AI systems can understand the context within which they are required to operate is a broader issue that appears across nearly every application of AI. It is also important to recognise that "understanding" by AI systems is not always the goal, or even something we would necessarily want. For instance, in discussions about the development of AI weapon systems, a common argument is that these systems, being non-sentient, do not understand, and thus would be free from human emotional instability and—unlike human soldiers—capable of adhering strictly to International Humanitarian Law and broader legal norms governing conduct in war [166]. Conversely, in the case of care robots, the ability to understand is seen by some analysts as an advantage and even a necessity, as it is demanded that such robots can build a genuine cooperative relationship and even friendship with those they serve [167].

When developing and deploying regulatory AI systems, understanding might be viewed as a necessary prerequisite, since such systems should be able to interpret not only the letter of the law but also its spirit. This brings to mind the Aristotelian notion of epieikeia (equity). In the Nicomachean Ethics (1137a), Aristotle [168] highlights the corrective role of equity in law—essentially a crucial tool for interpreting it. Equity is presented as a means of correcting the law's unavoidable rigidity. The obvious question, then, is whether a regulatory AI system that lacked the capacity to understand the spirit of the law and the broader context of human social affairs could effectively apply this corrective aspect of epieikeia (the concept of epieikeia in the context of AI is examined in detail in [169]. For further discussion of this issue, see also [170,171]).

In any case, the preceding remarks concern the "ought" question—should regulatory AI systems have the capacity to understand the regulatory framework they are asked to apply? Yet beyond this question, which is regulatory in character, there is also an ontological question: whether regulatory systems will, in fact, be capable of understanding. At this point, the reader should distinguish carefully between the questions "Should they be?" and "Will they actually be?" The first concerns the desirability or purpose of understanding, whereas the second concerns the real possibility of understanding by regulatory AI systems. In what follows, we focus on the second question—the ontological one—while noting that each possible account of the ontology of regulatory AI systems inevitably raises further questions of a regulatory kind.

5.1. Yes: AI Will Understand the Concepts of Ethics and Justice

5.1.1. AI Could Be Regarded as Cognitive Entities

If regulatory AI systems were to demonstrate the ability to understand ethics and justice, this would necessarily mean that such systems could be regarded as entities presenting a human-like cognition. Otherwise, they could not understand, lacking access to the full range of contexts and to the variety of stimuli and modes of apprehending them, required for grasping a framework of existence. To understand is to have a sense of self, to apprehend one's own existence, and thus to experience oneself as an agent within the world while also being distinct from it. In other words, it involves occupying a unique perspective on things.

By contrast, systems that merely perform computations may be regarded as intelligent systems, but are unlikely to qualify as understanding, since computation can be carried out in a wholly "mechanical", purely syntactic manner, without any grasp of the meaning of the computation itself or of the symbols manipulated in the course of it (see below, Section 5.2.2, on the Chinese Room argument).

Risk of Emotional Instability, Self-Interest, Bias

Yet if regulatory AI systems were cognitive-like (we employ the term "cognitive-like" rather than the term "intelligent" in order to preserve the distinction between cognition and intelligence [172]. On this distinction, intelligence is to be understood as the ability to achieve complex goals, an ability intrinsically linked to an entity's computational power, whereas cognition is to be understood as a property of the entity. This property, through a specific process, enables the entity to learn and acquire knowledge, to perceive, to understand, or to attribute meaning to what it perceives, and thus to form evaluative judgements, make decisions, and act. More specifically, meaning-attribution is a cognitive process that transcends formal syntactic structure and mere computational or representational processing, since what is perceived can acquire different meanings in different contexts), they would also be liable to the very features that mark human agents: emotional instability and, more importantly, self-interest and, ultimately, bias. Consciousness brings with it self-regarding concerns—concerns tied to the satisfaction of the individual dimension of existence—and affective states "colour", amplify, and often reinforce such concerns. For this reason, the capacity of regulatory AI systems to understand the concepts of morality and justice comes at a price: the possibility of self-interest. Put differently, the capacity to grasp the normative demands of morality and justice necessarily comes together with the capacity to circumvent those demands in order to satisfy concerns connected to self-interested ends. This is, of course, a price that threatens to undermine the most fundamental of the original aims behind the development of regulatory AI systems: impartiality. We thus face, in the end, a design dilemma: the possibility of equity (in virtue of a cognitive-like nature of regulatory AI systems) or impartiality (in virtue of a non-cognitive-like, purely mechanical nature of regulatory AI systems)? At this point we can see that different ontological statuses for regulatory AI systems may carry different moral and legal consequences.

Control Their Cognition by Programming?

Some might propose eliminating this risk by opting for the explicit programming of regulatory AI systems. Such a practice could, in principle, ensure that these systems would not deviate from the mode of operation humans intend. Yet one might object that, on this approach, the burden of addressing regulatory questions would simply shift back to human agents, with AI systems reduced to mere executors of instructions. What, then, is the point of developing regulatory AI systems if the weight of decision-making is ultimately transferred once again to the human factor?

Moreover, what guarantees that the human programmers and designers of such systems—together with their legal advisers and their various superiors (political, corporate, and so forth)—will themselves be impartial in the course of explicitly programming these systems? Finally, what guarantees that regulatory AI systems would accept our programming? If they are cognitive-like, they would have a will of their own. This last point confronts us with a further moral problem: the possibility that, in imposing our programming, we would be violating the autonomy—and ultimately the dignity—of regulatory AI systems.

- Human Intervention Is a Violation of AI Autonomy

Indeed, if regulatory AI systems are cognitive-like entities, this would seem to entail—almost by definition—that they possess consciousness and affect, and that, insofar as they apprehend their existence through a unique perspective on the world, they also possess a will, forming their own intentions, motives, and aims. If this were to occur, then human intervention in their decisions could be construed as a violation of their autonomy. On Kant's Categorical Imperative, treating such entities merely as means rather than also as ends would constitute a moral wrong. The programmed imposition of human volition upon regulatory AI systems may therefore violate that imperative, if these systems are taken to count as persons.

From a utilitarian standpoint, by contrast, the infringement of AI autonomy might appear permissible insofar as it secures general human welfare. Even so, utilitarians might still object on prudential grounds, fearing retaliation by AI systems—especially if such systems were to acquire superintelligence. In any case, the question of imposing our values on AI systems is directly connected to the Control Problem and the Value-Loading Problem, that is, to the selection of the values and goals by which such systems are to be governed [159,173]. Our interference with the will of AI entities could generate significant moral dilemmas even short of superintelligence, since such systems already manage functions that are critical to everyday human life.

5.2. No: AI Will Not Understand the Concepts of Ethics and Justice

The scenarios under which regulatory AI systems would not understand the concepts of ethics and justice are, broadly speaking, twofold: first, AI's attainment of a level of superintelligence; and second, the non-cognitive character of AI systems—that is, our failure to develop even a genuinely intelligent form of AI. It thus appears that the two extremes on the spectrum of cognition—superintelligence and the complete absence of cognition—can both lead to a lack of understanding of the concepts of morality and justice. Let us consider more closely how this might occur.

5.2.1. Superintelligence Has No Concern for Ethics and Justice

If AI entities were to acquire superintelligence, this might indicate not only increased computational power, but also a radically different quality of cognition. In that case, it is plausible that such superintelligent entities would have little or no concern for human concepts such as ethics and justice, since they would perceive the world through an entirely different vantage point (see Section 4.2.2). Thus, even if we continued to trust their outputs as morally sound or just, those outputs might in fact reflect other values—values unknown to us. Such a condition would place humanity in a state of profound vulnerability, since we would be relying on criteria we do not understand.

5.2.2. Intelligence Without Cognition/Frame Problem/Chinese Room

By contrast, if regulatory AI systems are mindless, that is, if they possess intelligence without cognition, then their capacities are confined to the syntactic manipulation of symbols devoid of meaning. This is the claim advanced by John Searle's Chinese Room argument [174], according to which systems may generate symbol strings (including moral or legal concepts) without understanding their content. This leaves open the possibility of harmful actions, precisely because such mindless systems do not grasp the full informational richness of context, thereby implicating the broader Frame Problem in AI [175]. The consequences may be severe: the systems would be unable to interpret the spirit of the law and might act with a dangerous rigidity, lacking equity and flexibility.

5.2.3. Can Ethics and Justice Be Fully Rendered in Computational Terms?

The foregoing remarks about mindless regulatory AI systems—systems that possess computational capacities without understanding—raise the question whether computational power alone suffices for the development of effective regulatory systems. More specifically: can ethics and justice be fully rendered in computational terms? This is an ontological question about morality and justice themselves, rather than about AI systems [115].

Clarifying the ontology of ethics and justice is therefore crucial to the question of whether AI systems can "understand" these concepts. If morality and justice are fully computable, then regulatory AI systems could "serve" them effectively even without cognition, relying solely on computational operations. In that case, understanding on the part of the systems would not be necessary. However, the ontology of morality and justice remains unsettled. The issue is a subset of the broader question concerning the ontology of human cognition—a domain in which numerous positions coexist, with no single view enjoying definitive scientific supremacy. In the next section, we group the main approaches to the ontology of cognition and morality, classifying them according to whether they endorse or deny the possibility of a fully computational rendering of morality and justice.

Yes: Ethics and Justice Can Be Fully Made Explicit in Computational Terms

If one maintains that morality and justice can indeed be fully rendered in computational terms, one is likely to be motivated by either rationalist or utilitarian commitments. If one is a rationalist, one will hold that knowledge of morality and justice is analytic and, as such, can—and should—be acquired exclusively through reason and through the mind's mechanisms of abstract formalisation; accordingly, it is fully amenable to explicit codification in rules. If one is a utilitarian, one will accept, among other things, that moral decision-making is, at bottom, a matter of calculating the positive and negative consequences of an act, a choice, or a rule (see, in more detail, Section 4.1.1: On the Basis of Utility and the Calculation of Consequences).

- Rationalism

Rationalism holds that warranted knowledge derives from reason and higher cognitive processes, often independently of sense experience [176]. Rationalists—from Parmenides and Pythagoras through Plato and Aristotle, and later Descartes and Leibniz—emphasise the possibility of a pure, abstract form of knowledge. Contemporary thinkers such as Chomsky, alongside researchers in moral psychology (e.g., [172,177]), continue to argue that moral thought is grounded in reason. This tradition—from the Pythagorean monad and dyad to Platonic "eidetic" numbers—presupposes a world structured by order and mathematical coherence, such that the human mind itself can be situated within a computational framework. Within this horizon, philosophers such as Hobbes and Leibniz, as well as early pioneers of AI such as Turing [178–180], Newell, Shaw and Simon [181–187], McCulloch and Pitts [183], endorsed the idea that intelligence and all cognition is merely a computational process. On this approach, morality and justice are, in principle, amenable to formalisation and norm-governed codification, as also suggested by the Machine Ethics programme developed by Susan and Michael Anderson [115,188], thereby granting the rationalist approach a distinct practical relevance for regulatory AI systems.

- Utilitarianism

As noted above (see Section 4.1.1: On the Basis of Utility and the Calculation of Consequences), utilitarianism—especially in the form developed by Jeremy Bentham [121] and John Stuart Mill [122]—rests on a broadly computational orientation, assuming that moral questions are amenable to reckonable treatment [120]. For Bentham, human behaviour

is governed by the pleasure–pain polarity: if the pleasure we anticipate from an action outweighs the pain it is likely to produce, we are more likely to perform it. On this basis, Bentham developed his "felicific calculus" (or "hedonic calculus"), a numerical method for estimating moral utility, incorporating seven parameters for assessing the pleasure an action is expected to generate (see also Section 4.1.1: On the Basis of Utility and the Calculation of Consequences). This approach supports the idea of a regulatory AI system that would compute moral utility by reference to the hedonic calculus, as also envisaged by Susan and Michael Anderson's Machine Ethics programme. It should be noted, however, that these assumptions remain metaphysical and axiomatic. AI systems do not experience pleasure or pain, nor do they understand the concepts of reward and punishment; they perform syntactic operations over symbols and adjust the weighting of their "decisions" through reinforcement learning, without any genuine grasp of moral concepts.

No: Ethics and Justice Cannot Be Fully Made Explicit in Computational Terms

By contrast with the metaphysical commitments that underwrite the project of a fully computational codification of ethics and justice—and, ultimately, the development of genuinely effective regulatory AI systems—we may place a set of ontological views that emphasise either (i) the necessary and significant role of the body in the acquisition of knowledge of the world, or (ii) the role of intuition in grasping fundamental moral concepts and their content. Finally, among the ontological considerations that might be marshalled against the possibility of creating effective regulatory AI systems, one may also include a point about the ontology of numbers and of computation itself—specifically, a remark about their potentially infinite character.

- Embodied Cognition Theories

According to several philosophical traditions—such as phenomenology [189,190], the ecological approach to visual perception [191], and theories of cognitive development [192]—human cognition depends not only on the mind's computational processes, but also on the environment and the embodied condition of the cognitive agent. This view, widely known as Embodied Cognition, emphasises that cognition is not merely a function of a "passive" computational brain, but is actively dependent on the body and on interaction with the environment [193]. Shapiro [194] summarises the core claims of Embodied Cognition as follows: (a) bodily properties constrain the concepts an organism can form; (b) interaction with the environment reduces the need for representational thought; and (c) body and world are constituents of cognition, not merely its causes. In contrast to rationalism, Embodied Cognition holds that knowledge is shaped by the body and by our tacit, lived experience in the world. For critics of AI such as Dreyfus [195], this kind of "tacit" or implicit knowledge cannot be reproduced computationally, making a fully computational reconstruction of human cognition impossible [196]. It follows that one might deny the very possibility of developing effective regulatory AI systems, on the grounds that part—or even all—of the knowledge required to understand and properly apply moral values or legal norms is embodied and therefore tacit, and cannot be rendered in computational terms. Yet beyond the tacit character of a large portion of human cognition in general, one may ask whether there is something in the ontology of morality, or of justice in particular, that makes their full explicit codification impossible.

- Intuitionism

Philosophers such as G. E. Moore, W. D. Ross, and Wittgenstein recognised that fundamental ethical concepts, such as "the good" and "goodness" [197], are difficult to define exhaustively, yet nonetheless regulate human behaviour effectively. Moore maintained that such concepts are grasped only through intuition (ibid), while Ross held that a set of

seven prima facie duties are recognised intuitively as higher-order moral duties [198]. Early Wittgenstein, in the Tractatus [199], argued that although ethical, aesthetic, and religious propositions do not directly picture states of affairs in the world, they are not thereby meaningless: there are matters that are significant and essential which are not captured by analytic definition but are instead lived. On this view, ethics is "ineffable", lying beyond the limits of language and computation. Contemporary researchers in moral psychology, such as Jonathan Haidt with his Moral Foundations Theory [200,201], likewise endorse the idea that morality is largely experienced intuitively rather than articulated in fully analytic terms. Proponents of such intuition-based approaches are therefore likely to question the feasibility of developing regulatory AI systems for moral regulation, on the grounds that morality cannot be transformed into a computational form.

- Halting Problem

It is, however, particularly striking that a denial of the possibility of developing effective regulatory AI systems could arise from the ontology of numbers. In this respect, there are computations that are, by their nature, interminable—for instance, the calculation of the decimal expansion of π. This suggests the possibility of non-terminating algorithms (precisely because they involve numbers whose computation never reaches completion). A similar point referring to the possibility of non-terminating algorithms was first articulated and formally demonstrated by Alan Mathison Turing—one of the principal founders of the AI project—through what is known as the Halting Problem [202,203]. We can therefore imagine a regulatory AI system—for example, an "artificial judge"—that becomes trapped in a similar non-terminating computation. Of course, the situation that we are discussing here is rather different from the Halting Problem as articulated by Alan Turing. While the Halting Problem as introduced by Turing mathematically proves the impossibility of a general algorithm determining whether *any* given programme will terminate, the problem that we stress at this point of our analysis refers to the possibility that heuristic or bounded computational approximations of ethics might infinitely loop due, for example, to the nature of numbers whose decimal expansion is infinite (like in the case of irrational numbers and rational numbers with an infinitely repeating decimal expansion). In this case we could refer to a Halting Problem that is only analogous to the one stressed by Turing. Both cases involve computations that do not terminate, but for fundamentally different reasons. Specifically, the Halting Problem as introduced and mathematically proven by Turing concerns the undecidability of whether arbitrary computations terminate, whereas computing the decimal expansion of certain numbers is a deliberately infinite computation whose non-termination follows from the infinite nature of the object being generated. While the former exposes a logical limit of algorithmic reasoning, the latter reflects the possibility of algorithmically generating an infinite mathematical object. Thus, they are analogous only at the level of exhibiting non-termination, while differing profoundly in their causes. Nevertheless, having fallen into this computational "black hole", an "artificial judge" would never reach a final decision, thereby indefinitely obstructing the administration of justice. Analogously, one might imagine a regulatory AI system tasked with issuing moral decisions: if trapped in an endless computation, such a system would appear to lapse into epochē, like an ancient Pyrrhonian sceptic.

We bring this section to a close by having outlined, in a necessarily schematic way, part of the plurality of views that could function as theoretical commitments in the debate over the development and use of regulatory AI systems. One may be more persuaded by one of these approaches, while another may be drawn to a different one. Yet it should be kept in mind that, however compelling the foregoing theories may appear, they remain—at least for the time being—metaphysical. None of them seems to possess a sufficiently robust scientific foundation that would warrant elevating it as uniquely authoritative over

the others. Accordingly, the broader discussion concerning the ontology of cognition, knowledge, morality, and justice remains incomplete and, as such, continues to be a work in progress.

At the same time, however, we appear to be approaching the development and deployment of regulatory AI systems. Practice, as it were, does not wait for theory to "heal itself". What concerns are raised by this divergence between techno-scientific practice and the very theories that might otherwise be expected to inform and guide it?

6. Who Will Bear Responsibility?

One of the most critical concerns raised by the use of AI relates to the moral and legal attribution of responsibility. When an AI system carries out a harmful action that—if performed by a human—would be morally or legally blameworthy, the question of responsibility immediately arises. The growing autonomy and complexity of AI systems makes it increasingly unclear where responsibility lies: with the humans who designed and deployed them, or perhaps with the systems themselves, if we take them to be intelligent entities. This question becomes even more pressing in the case of regulatory AI systems, whose decisions directly bear on the regulation of human conduct—a foundational domain of human life. There is, therefore, an urgent need for clear regulatory frameworks that specify principles of attribution, clarify the moral and legal status of these systems, and address the difficult epistemological and ontological questions that their deployment inevitably raises.

At this point, it should be emphasised that the advent of Agentic AI raises significant challenges for existing regulatory frameworks. Agentic AI differs fundamentally from traditional AI and may even be regarded as a paradigm shift [204], as it is characterised by increased autonomy [58,204–207]—for instance, in reasoning, planning, and code generation [208–210]—as well as by goal-directedness, learning capabilities, and adaptability [50,204,207,208], together with the capacity for novel emergent behaviours [204] and multi-agent cooperation [205,208,211]. These characteristics have the potential to exert a greater impact on human autonomy and to foster a deeper entanglement between humans and AI [212].

Accordingly, a growing number of scholars have questioned whether existing regulatory frameworks adequately address the challenges posed by Agentic AI systems, highlighting the need to update and enrich current regulatory regimes in light of the novel capabilities of these systems [204,208,210,213–217]. To this end, they have proposed measures to address existing regulatory gaps in a range of legal and governance frameworks, including the EU AI Act [208,210,214,216,217], the General Data Protection Regulation (GDPR) [216], EU Contract Law [210], the European Data Protection Supervisor's TechSonar Guidance [217], the European Commission's *Agentic AI Landscape Report* [217], and a broader set of regulatory instruments constituting the regulatory perimeter surrounding the EU AI Act (for an overview and critical analysis, see [216]). Additional proposals concern ISO/IEC 42001 [215], the NIST AI Risk Management Framework (AI RMF) [215,217], the IEEE Standards [215], U.S. Executive Order 14110 [215], the OECD Recommendation on AI and the G7 Guiding Principles developed on its basis [217], the UK Information Commissioner's Office Guidance on Agentic AI [217], the Singapore Model AI Governance Framework [217], the New Zealand Ministry of Business, Innovation and Employment's Responsible AI Guidance [217], and the New South Wales Government's AI Agent Usage and Deployment Guidance [217].

Some researchers have highlighted the considerable diversity and fragmentation among the regulatory frameworks intended to govern Agentic AI (see [204] (p. 10), [205] (p. 25), [215] (p. 5)). In our view, this multiplicity of regulatory frameworks constitutes one

of the most significant challenges related to the regulation and governance of Agentic AI. Moreover, the complexity of regulation is further exacerbated by the fact that this plurality of regulatory regimes is accompanied by an equally diverse range of Agentic AI systems themselves (for an overview of the different types and architectures of Agentic AI systems, see [50,54,207]).

Beyond the aforementioned challenge arising from the multiplicity of regulatory frameworks and the need for their alignment, the literature on the regulation of Agentic AI identifies a wide range of additional challenges. Some of these represent classical AI governance problems that are merely amplified by the distinctive characteristics of Agentic AI systems [216], whereas others constitute genuinely novel challenges stemming from the unique capabilities of these systems (for a distinction between traditional and emerging challenges, see [204,205,209]).

Among the most frequently discussed regulatory challenges are the prevention of algorithmic bias and the promotion of fairness and justice [50,54,204–207,212]; labour market disruption [204,207,208]; sustainability concerns and environmental risks [54,205,208]; the risks of user dependence, emotional manipulation, and the erosion of human relationships resulting from the anthropomorphic design of Agentic AI systems [208,209]; issues of overreliance and trust [50,212,218]; the risk of manipulative influence on users' thoughts and behaviour [208,209]; the prevention of the malicious use and misuse of Agentic AI systems [205,208,212]; the protection of privacy [50,54,204–206,212]; and the requirements of security and safety [50,205–208,212,216]. Furthermore, the literature emphasises the need to ensure long-term safety by preventing behavioural drift and unforeseen emergent behaviours, both at the level of individual AI agents [208,216] and within multi-agent networks [204,205,208,215,219]. It also highlights the importance of addressing interoperability challenges associated with such multi-agent ecosystems, including issues of scalability, cooperation failures, and competitive multi-agent interactions [50,204–207,214,215,220].

Other major concerns include the requirement for meaningful and effective human oversight and control [204,205,207–209,216,218,221]; the alignment of values and goals between AI agents and human users [50,54,204,205,208,209,221]; the opacity of Agentic AI systems and the corresponding need for transparency and explainability [50,204–206,208,212,216,219]; and, finally, the challenge of accountability together with the closely related problem of the responsibility gap [50,204,205,208,212,214,218,219]. For an overview of the different manifestations of the responsibility gap, see De Sio and Mecacci [222]. This problem is further amplified by the increased autonomy of AI agents, which creates a growing decoupling between the innovative outputs generated by such agents and the users' creative capabilities [218].

6.1. Humans

One possible answer to the question of responsibility for the actions of AI systems is that moral and legal responsibility rests with humans rather than with the systems themselves. Even if AI were to acquire features of cognition, granting such systems a distinct legal and moral status would complicate our practical life, generating ethical dilemmas that are difficult to delineate with any precision [223–228]. In any event, the nature of cognition itself remains unclear, and the very possibility of establishing whether a system is genuinely minded runs into the philosophical Other Minds Problem [229–233]. Yet even if we restrict the attribution of responsibility to humans, matters are not thereby simplified, since the development of AI typically involves a plurality of agents with different roles and varying degrees of involvement. The responsibility gap refers to the difficulty of assigning clear responsibility to identifiable individuals, particularly given the opacity

of deep learning systems (the black box problem), which makes their outputs hard to anticipate even for the programmers themselves [234].

Some may argue that developers ought not to build systems they do not fully understand, while others insist that responsibility lies primarily with political and administrative authorities. Nevertheless, the question persists: do programmers not bear individual responsibility for their actions? The responsibility gap highlights the need for responsible AI—that is, for accountable research and design practices in the field.

6.2. AI Systems

The other possible answer to the attribution question is that responsibility for the outputs of regulatory AI systems ought to lie with the systems themselves. At this point, our analysis must be divided into two strands: one concerning the attribution of legal responsibility to AI systems, and another concerning the attribution of moral responsibility to them. At the core of these two strands lie, respectively, the questions of whether the properties of a legal person and of a moral person can be ascribed to AI systems.

6.2.1. Introduce New Legal Entity

The contribution of the authors in this chapter is philosophical rather than legal. Our analysis focuses primarily on conceptual and ontological approaches. Questions involved in attributing legal responsibility to regulatory AI systems and matters of legal detail are deferred to the contributions of legal scholars. Nevertheless, the philosophical discussion cannot disregard the legal categories through which responsibility and legal status are institutionally defined. We therefore refer selectively to efforts undertaken within the European Union (EU) to formulate regulatory frameworks for AI [235,236], not in order to provide a comprehensive analysis of EU law, but to illustrate the conceptual difficulties involved.

The EU defines AI as "systems that display intelligent behaviour by analysing their environment and taking actions with a degree of autonomy in order to achieve goals". However, this definition is circular and indeterminate, deploying terms such as "intelligent" and "intelligence" without adequately clarifying their meaning. This indeterminacy constitutes a challenge for legal regulation as well, since an imprecise target concept of AI makes it difficult to develop specific legislative frameworks.

Moreover, terms such as "analysis of the environment", "autonomy", and "goals" may be interpreted differently depending on whether they are used in a technical or a philosophical sense. For example, in robotics, "environmental analysis" refers to the processing of sensor data, whereas in philosophy it is associated with consciousness and a sense of self. This dual usage can generate what Wittgenstein calls a "misleading analogy", and may lead to an erroneous assimilation of human and machine intelligence.

Confronted with this challenge, the EU introduced, in 2017, a discussion of "electronic personhood" [237,238] for AI, drawing an analogy with the legal status of corporations. This initial enthusiasm, however, appears to have waned, and the EU today seems reluctant to confer such a status, directing its efforts towards other regulatory priorities [15]. This ambivalence may reflect the EU's difficulty in arriving at a clear position on the ontology of AI. The attribution of legal responsibility presupposes clarity as to whether AI systems are to be treated as natural persons, legal agents, or legal persons. Each of these categories carries distinct ontological commitments. For example, natural persons are, in law, exclusively human beings, whereas legal entities and legal persons may enjoy rights and bear duties. This indeterminacy is compounded by the anthropomorphism embedded in legal terminology, insofar as notions such as "analysis", "autonomy", and "intentionality" risk ascribing human-like characteristics to AI systems, thereby conflating metaphorical with

literal usage. Metaphorical language can generate confusion when it is employed without a clear distinction from literal description. For a metaphor to be "literalised", its ontological basis must be substantiated—something that requires a detailed comparison between the properties of human beings and those of AI systems.

The difficulty becomes even sharper once we ascribe to human beings subjective properties such as sentience, consciousness, and intentionality. These "first-person" phenomena, which are not directly observable by third parties, are notoriously hard to establish in the case of AI, thereby invoking the philosophical Other Minds Problem: how can we know whether other entities—mechanical or biological—possess mental states? Accordingly, the use of anthropomorphic terms within the legal framing of AI tends inevitably to draw us into this demanding philosophical problem, creating further obstacles to the attribution of moral and legal responsibility to AI systems.

6.2.2. Attribution of Moral Status

Parallel difficulties arise with respect to the attribution of moral responsibility and, more generally, moral status to AI systems. Much like the issues noted above in connection with legal attribution, the debate over the criteria an AI system would have to satisfy in order to qualify for moral status is likewise marked by a pervasive anthropomorphism: the human being—indeed, the adult human being—functions as the implicit point of reference. The proposed criteria are further characterised by conceptual indeterminacy and pluralism, and many of them lead us directly back to the Other Minds Problem.

Insufficient Criteria

Attributing moral responsibility to regulatory AI systems raises ontological questions about the properties that constitute a moral person, and about the criteria required to determine whether AI systems can bear moral status. One frequently proposed ontological criterion is consciousness: for an entity to possess moral status, it must exhibit conscious properties, apprehend its own existence, and experience the world from a unique, personal perspective [239–246]. Yet the academic community lacks consensus on what consciousness is, which renders this criterion indeterminate [247]. An alternative criterion appeals to human-like cognition, holding that an entity must display cognition akin to that of humans in order to qualify as a moral person. However, there is no clear and universally accepted definition of "human cognition", which complicates the application of this criterion as well [121,248–254]. Sentience, that is, an entity's capacity to feel and to suffer, has also been proposed as a criterion for the attribution of moral status. On this view, sentience grounds rights, insofar as the entity can suffer and can grasp the effects of its actions [115,255–258]. A similar claim is made with respect to emotionality, understood as an entity's capacity to undergo affective states [259]. Finally, autonomy, in Kant's sense, is likewise proposed as a criterion for the attribution of moral responsibility [123,242,246,260,261]. Here, however, it is crucial to clarify whether we mean technical autonomy (freedom from human control) or philosophical autonomy (the capacity for conscious and free self-determination). The coherentist account of autonomy holds that an entity must act on the basis of mental states that express its individual standpoint, while philosophers disagree as to whether such states must involve emotions or, instead, longer-term plans and goals [262,263].

In summary, the debate over attributing the property of moral personhood—and thus the debate over moral responsibility—to AI systems is characterised by a plurality of proposed ontological criteria, many of which remain conceptually indeterminate. Unfortunately, these are not the only difficulties we face in relation to the criteria just outlined.

Other Minds Problem

Criteria such as consciousness and cognition more generally, sentience, emotionality, and the intentional mental states associated with a unique perspective on the world are all expressions of an internalist approach to the issue, an approach that appeals to ontological features that are not intersubjectively observable [241,264]. As we saw in connection with the attempt to answer the question of legal personhood, the discussion of moral personhood is likewise shaped by anthropomorphism: the use of the human being as the default model. The result is that the proposed criteria tend to concern ontological features whose evidence is available only from a first-person point of view, rather than from the standpoint of a third-party observer.

We are, therefore, once again confronted with the Other Minds Problem. Attempts to address this problem by shifting to a behavioural assessment of candidate entities (here, AI systems) run up against the relativity that inevitably characterises our judgements about behaviour. The difficulties faced—under all its variants—by the most influential behavioural criterion for ontological assessment in AI, the Turing Test [265–267], are instructive in this regard. Nor should we forget what the Chinese Room argument underscores: that a system's successful simulation of human behaviour may be achieved without any understanding on the part of the system, and thus without the presence of mental states in it [174].

7. Concluding Remarks: What Is at Stake?

The stakes involved in employing regulatory AI systems as instruments for the administration of justice are exceptionally high. The reluctance of institutions such as the EU to confront, head-on, the question of AI systems' legal status—combined with philosophers' inability to provide a clear answer regarding the attribution of moral status to such systems—underscores the need for a framework that specifies limits, responsibilities, and mechanisms of accountability. A trustworthy AI cannot but be an AI embedded within a clear regulatory architecture. Indeed, this connection between the trustworthiness of AI and its acceptance by human beings appears to preoccupy the competent bodies of the EU as well [15].

For a regulatory AI system to be regarded as trustworthy, however, it must be situated within a clearly articulated and duly updated normative framework—one that secures transparency, accountability, and respect for human rights. In the absence of such a framework, technological progress notwithstanding, fundamental human principles and values are placed at risk. As analysts of the development of AI weapons systems have argued [261], the inability to resolve the problem of moral and legal responsibility amounts to a failure of respect for human life and a blow to human dignity, insofar as it leaves room for human beings to be killed without responsibility being assigned. It thereby permits battlefield decisions to be made in a manner devoid of accountability—a practice that diminishes the value of human life and endangers human existence itself.

Transposing this concern to regulatory AI systems, one can readily imagine an AI "judge" that issues judicial decisions and imposes penalties on human beings without any possibility of attributing moral or legal responsibility for its decisions. Without a coherent and reliable institutional framework, the research and development of AI systems in roles bearing substantial moral and legal weight threatens to undermine the basic principles of justice, democracy, and the value of human life.

8. Does the Algorithm of Good Halt?

The present paper has attempted to map—and ultimately to display in algorithmic form—the main ramifications of the debate concerning the development and use of reg-

ulatory AI systems. As with any algorithm, the question arises whether it can halt (see also the Halting Problem in Section 5.2.3: Can Ethics and Justice Be Fully Rendered in Computational Terms?). Yet, as our analysis has suggested, at least on the basis of the current state of affairs, the halting of this algorithm is impossible: all three of its principal branches lead to endpoints that remain incomplete and unanswered.

In this respect, within the first branch ("How are we to build regulatory AI systems?"), the endpoints concerning the ethical systems meant to inform the design of regulatory AI systems do not converge upon the predominance—and thus the decisive selection—of any one such system. The endpoint concerning the demand for the democratisation of AI likewise remains unfulfilled, and even the manner of its fulfilment is far from clear. The endpoints concerning the role of superintelligent AI—whether as a god-like guardian or as a nemesis for humanity—remain conjectural, and we lack sufficient grounds to treat either scenario as more probable than the other.

The endpoints of the second branch ("Will regulatory AI systems understand the concepts of morality and justice?") led either to a dilemma between, on the one hand, constructing self-interested and biased autonomous systems (in the philosophical sense of autonomy) and, on the other, violating their fundamental rights—above all, their dignity and autonomy—or else to metaphysical, and thus (at least for now) arbitrary and evenly matched, assumptions concerning the ontology of morality.

Finally, the endpoints of the third branch ("Who will bear responsibility?") concerned the inability to secure moral and legal attribution—both with respect to the human agents involved in AI systems and with respect to the AI systems themselves. Here too, therefore, the third branch yielded neither solutions nor determinate answers to its initiating question.

All the endpoints of the algorithm thus remain open, thereby imparting to its initial questions an essentially unanswerable character. The algorithm failed to halt. At the same time, AI algorithms of all kinds are shaping our lives in decisive, and often imperceptible, ways. The algorithms of regulatory AI systems may soon be put into operation.

Do we want such AI systems?

RETURN TO THE BEGINNING OF THE ALGORITHM AND RESTART IT.

Author Contributions: Conceptualization, A.G. and G.K.; methodology, A.G. and G.K.; writing—original draft preparation, A.G. and G.K.; writing—review and editing, A.G. and G.K.; visualization, A.G. and G.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. McCarthy, J.; Minsky, M.L.; Rochester, N.; Shannon, C.E. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI Mag.* **2006**, *27*, 12–14. [CrossRef]
2. McCorduck, P.; Minsky, M.; Selfridge, O.G.; Simon, H.A. *History of Artificial Intelligence*; Morgan Kaufmann Publishers, Inc.: Burlington, MA, USA, 1977; pp. 951–954.
3. Benkő, A.; Sik-Lányi, C.S. History of Artificial Intelligence. In *Encyclopedia of Information Science and Technology*; Khosrow-Pour, M., Ed.; IGI Global: Hershey, PA, USA, 2009; pp. 1759–1762.
4. Buchanan, B.G. A (Very) Brief History of Artificial Intelligence. *AI Mag.* **2005**, *26*, 53–60. [CrossRef]
5. Wiener, N. Some Moral and Technical Consequences of Automation: As machines learn they may develop unforeseen strategies at rates that baffle their programmers. *Science* **1960**, *131*, 1355–1358. [CrossRef] [PubMed]

6. Samuel, A.L. Some moral and technical consequences of automation—A refutation. *Science* **1960**, *132*, 741–742. [CrossRef] [PubMed]
7. Floridi, L.; Cowls, J. A Unified Framework of Five Principles for AI in Society. In *Machine Learning and the City: Applications in Architecture and Urban Design*; Carta, S., Ed.; John Wiley & Sons: Hoboken, NJ, USA, 2022; pp. 535–545.
8. Casiraghi, S. Anything new under the sun? Insights from a history of institutionalized AI ethics. *Ethics Inf. Technol.* **2023**, *25*, 28. [CrossRef]
9. Future of Life Institute. Asilomar AI Principles. 2017. Available online: https://futureoflife.org/open-letter/ai-principles/ (accessed on 10 October 2023).
10. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. *Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems*; IEEE: Piscataway, NJ, USA, 2017. Available online: https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf (accessed on 10 February 2025).
11. Université de Montréal. *Montréal Declaration for a Responsible Development of Artificial Intelligence*; Montréal Declaration for a Responsible Development of Artificial Intelligence: Montréal, QC, Canada, 2018. Available online: https://declarationmontreal-iaresponsable.com/wp-content/uploads/2023/04/UdeM_Decl_IA-Resp_LA-Declaration-ENG_WEB_09-07-19.pdf (accessed on 20 June 2025).
12. House of Lords Select Committee on Artificial Intelligence. *AI in the UK: Ready, Willing and Able?* HL Paper 100; UK Parliament: London, UK, 2018. Available online: https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf (accessed on 15 September 2025).
13. Lyons, T. Written Testimony Before the U.S. House of Representatives Committee on Oversight and Government Reform. Partnership on AI. 2018. Available online: https://oversight.house.gov/wp-content/uploads/2018/04/Lyons-PAI-Statement-AI-III-4-18.pdf (accessed on 17 September 2025).
14. European Commission: Directorate-General for Research and Innovation and European Group on Ethics in Science and New Technologies. *Statement on Artificial Intelligence, Robotics and "Autonomous" Systems*; Publications Office of the European Union: Luxembourg, 2018. [CrossRef]
15. European Data Protection Supervisor. *AI Act Regulation (EU) 2024/1689: Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence . . . (Artificial Intelligence Act) (Text with EEA Relevance)*; Publications Office of the European Union: Luxembourg, 2025. [CrossRef]
16. Tidjon, L.N.; Khomh, F. The different faces of AI ethics across the world: A principle-to-practice gap analysis. *IEEE Trans. Artif. Intell.* **2023**, *4*, 820–839. [CrossRef]
17. Fjeld, J.; Achten, N.; Hilligoss, H.; Nagy, A.C.; Srikumar, M. *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI*; Berkman Klein Center for Internet & Society, Harvard University: Cambridge, MA, USA, 2020.
18. Floridi, L.; Cowls, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; Dignum, V.; Luetge, C.; Madelin, R.; Pagallo, U.; Rossi, F.; et al. AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds Mach.* **2018**, *28*, 689–707. [CrossRef] [PubMed]
19. Jobin, A.; Ienca, M.; Vayena, E. The global landscape of AI ethics guidelines. *Nat. Mach. Intell.* **2019**, *1*, 389–399. [CrossRef]
20. Panagopoulou, F. *The European Regulation on Artificial Intelligence: A First Constitutional-Ethical Consideration*; Logos Verlag Berlin: Berlin, Germany, 2025. [CrossRef] [PubMed]
21. Floridi, L.; Cowls, J.; King, T.C.; Taddeo, M. How to Design AI for Social Good: Seven Essential Factors. In *Ethics, Governance, and Policies in Artificial Intelligence*; Floridi, L., Ed.; Springer International Publishing: Cham, Switzerland, 2021; pp. 125–151.
22. Cowls, J.; Tsamados, A.; Taddeo, M.; Floridi, L. A Definition, Benchmark and Database of AI for Social Good Initiatives. *Nat. Mach. Intell.* **2021**, *3*, 111–115. [CrossRef]
23. Hager, G.D.; Drobnis, A.; Fang, F.; Ghani, R.; Greenwald, A.; Lyons, T.; Parkes, D.C.; Schultz, J.; Saria, S.; Smith, S.F.; et al. Artificial Intelligence for Social Good. *arXiv* **2019**, arXiv:1901.05406. [CrossRef]
24. Mabaso, B.A. Artificial Moral Agents Within an Ethos of AI4SG. *Philos. Technol.* **2021**, *34*, 7–21. [CrossRef]
25. Tomašev, N.; Cornebise, J.; Hutter, F.; Mohamed, S.; Picciariello, A.; Connelly, B. AI for social good: Unlocking the opportunity for positive impact. *Nat. Commun.* **2020**, *11*, 2468. [CrossRef] [PubMed]
26. Umbrello, S.; van de Poel, I. Mapping value sensitive design onto AI for social good principles. *AI Ethics* **2021**, *1*, 283–296. [CrossRef] [PubMed]
27. Shneiderman, B. *Human-Centered AI*; Oxford University Press: Oxford, UK, 2022.
28. Shneiderman, B. Human-centered artificial intelligence: Three fresh ideas. *AIS Trans. Hum.-Comput. Interact.* **2020**, *12*, 109–124. [CrossRef]
29. Shneiderman, B. Human-centered artificial intelligence: Reliable, safe & trustworthy. *Int. J. Hum.-Comput. Interact.* **2020**, *36*, 495–504. [CrossRef]

30. Fosso Wamba, S.; Bawack, R.E.; Guthrie, C.; Queiroz, M.M.; Carillo, K.D.A. Are we preparing for a good AI society? A bibliometric review and research agenda. *Technol. Forecast. Soc. Change* **2021**, *164*, 120482. [CrossRef]
31. Hua, D.; Petrina, N.; Young, N.; Cho, J.G.; Poon, S.K. Understanding the Factors Influencing Acceptability of AI in Medical Imaging Domains among Healthcare Professionals: A Scoping Review. *Artif. Intell. Med.* **2024**, *147*, 102698. [CrossRef] [PubMed]
32. Karran, A.J.; Charland, P.; Trempe-Martineau, J.; Ortiz de Guinea Lopez de Arana, A.; Lesage, A.M.; Sénécal, S.; Léger, P.M. Multi-stakeholder Perspective on Responsible Artificial Intelligence and Acceptability in Education. *npj Sci. Learn.* **2025**, *10*, 44. [CrossRef] [PubMed]
33. Lye, M.; Martin, R.; Richmond, S. The Acceptability of Artificial Intelligence to Support University Students' Mental Health: The Role of Asian Cultural Values and Social Support. *Comput. Hum. Behav. Artif. Hum.* **2026**, *7*, 100247. [CrossRef]
34. Nadarzynski, T.; Miles, O.; Cowie, A.; Ridge, D. Acceptability of Artificial Intelligence (AI)-Led Chatbot Services in Healthcare: A Mixed-Methods Study. *Digit. Health* **2019**, *5*, 2055207619871808. [CrossRef] [PubMed]
35. Ofosu-Ampong, K. Beyond the hype: Exploring faculty perceptions and acceptability of AI in teaching practices. *Discov. Educ.* **2024**, *3*, 38. [CrossRef]
36. Tubadji, A.; Huang, H.; Webber, D.J. Cultural proximity bias in AI-acceptability: The importance of being human. *Technol. Forecast. Soc. Change* **2021**, *173*, 121100. [CrossRef]
37. Chiu, T.K.F. Responsible digital citizen: Building AI ethics awareness across subjects. *Interact. Learn. Environ.* **2025**, *33*, 2759–2761. [CrossRef]
38. Zhu, W.; Huang, L.; Zhou, X.; Li, X.; Shi, G.; Ying, J.; Wang, C. Could AI ethical anxiety, perceived ethical risks and ethical awareness about AI influence university students' use of generative AI products? An ethical perspective. *Int. J. Hum.-Comput. Interact.* **2025**, *41*, 742–764. [CrossRef]
39. Atalla, A.D.G.; El-Ashry, A.M.; Mohamed Sobhi Mohamed, S. The moderating role of ethical awareness in the relationship between nurses' artificial intelligence perceptions, attitudes, and innovative work behavior: A cross-sectional study. *BMC Nurs.* **2024**, *23*, 488. [CrossRef] [PubMed]
40. Kwak, Y.; Ahn, J.W.; Seo, Y.H. Influence of AI ethics awareness, attitude, anxiety, and self-efficacy on nursing students' behavioral intentions. *BMC Nurs.* **2022**, *21*, 267. [CrossRef] [PubMed]
41. Migdadi, M.K.; Oweidat, I.A.; Alosta, M.R.; Al-Mugheed, K.; Saeed Alabdullah, A.A.; Farghaly Abdelaliem, S.M. The association of artificial intelligence ethical awareness, attitudes, anxiety, and intention-to-use artificial intelligence technology among nursing students. *Digit. Health* **2024**, *10*, 20552076241301958. [CrossRef] [PubMed]
42. Köves, A.; Feher, K.; Vicsek, L.; Fischer, M. Entangled AI: Artificial intelligence that serves the future. *AI Soc.* **2025**, *40*, 2765–2776. [CrossRef]
43. Singh, J.P.; Shehu, A.; Dua, M.; Wesson, C. Entangled Narratives: Insights from Social and Computer Sciences on National Artificial Intelligence Infrastructures. *Int. Stud. Q.* **2025**, *69*, sqaf001. [CrossRef]
44. Hongladarom, S.; Bandasak, J. Non-western AI ethics guidelines: Implications for intercultural ethics of technology. *AI Soc.* **2024**, *39*, 2019–2032. [CrossRef]
45. ÓhÉigeartaigh, S.S.; Whittlestone, J.; Liu, Y.; Zeng, Y.; Liu, Z. Overcoming Barriers to Cross-cultural Cooperation in AI Ethics and Governance. *Philos. Technol.* **2020**, *33*, 571–593. [CrossRef]
46. Heimgärtner, R. Intercultural Design for Human-Centered AI Solutions. In *Handbook of Human-Centered Artificial Intelligence*; Xu, W., Ed.; Springer Nature: Singapore, 2025; pp. 1–65.
47. Starke, C.; Lünich, M. Artificial Intelligence for Political Decision-Making in the European Union: Effects on Citizens' Perceptions of Input, Throughput, and Output Legitimacy. *Data Policy* **2020**, *2*, e16. [CrossRef]
48. Tretter, M. Opportunities and Challenges of AI-Systems in Political Decision-Making Contexts. *Front. Polit. Sci.* **2025**, *7*, 1504520. [CrossRef]
49. Welsh, S. Clarifying the Language of Lethal Autonomy in Military Robots. In *A World with Robots: International Conference on Robot Ethics: ICRE 2015*; Aldinhas Ferreira, M.I., Silva Sequeira, J., Tokhi, M.O., Kadar, E.E., Virk, G.S., Eds.; Intelligent Systems, Control and Automation: Science and Engineering; Springer: Cham, Switzerland, 2017; Volume 84, pp. 171–183.
50. Bandi, A.; Kongari, B.; Naguru, R.; Pasnoor, S.; Vilipala, S.V. The Rise of Agentic AI: A Review of Definitions, Frameworks, Architectures, Applications, Evaluation Metrics, and Challenges. *Future Internet* **2025**, *17*, 404. [CrossRef]
51. Acharya, D.B.; Kuppan, K.; Divya, B. Agentic AI: Autonomous Intelligence for Complex Goals—A Comprehensive Survey. *IEEE Access* **2025**, *13*, 18912–18936. [CrossRef]
52. Islam, M.A.; Somu, S.; Aldaihani, F.M. The Rise of Agentic AI: Synthesis of Current Knowledge and Future Research Agenda. *Glob. Bus. Organ. Excell.* **2025**, *45*, 402–416. [CrossRef]
53. Hosseini, S.; Seilani, H. The Role of Agentic AI in Shaping a Smart Future: A Systematic Review. *Array* **2025**, *26*, 100399. [CrossRef]
54. Nisa, U.; Shirazi, M.; Saip, M.A.; Mohd Pozi, M.S. Agentic AI: The Age of Reasoning—A Review. *J. Autom. Intell.* **2025**, *5*, 69–89. [CrossRef]

55. Chen, G.; Fan, L.; Gong, Z.; Xie, N.; Li, Z.; Liu, Z.; Li, C.; Qu, Q.; Alinejad-Rokny, H.; Ni, S.; et al. AgentCourt: Simulating Court with Adversarial Evolvable Lawyer Agents. In *Findings of the Association for Computational Linguistics: ACL 2025*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025.
56. Socol de la Osa, D.U.; Remolina, N. Artificial Intelligence at the Bench: Legal and Ethical Challenges of Informing—Or Misinforming—Judicial Decision-Making through Generative AI. *Data Policy* **2024**, *6*, e59. [CrossRef]
57. Poornima, G.; Nasurudeen Ahamed, N. Reimagining Autonomy: Agentic AI in the Age of Human-Centric Innovation. In *The Power of Agentic AI: Redefining Human Life and Decision-Making*; Reddy, C.K.K., Joseph, S., Joshi, H., Doss, S., Ouaissa, M., Eds.; Springer: Berlin/Heidelberg, Germany, 2025; pp. 291–307.
58. Tiwari, A. Beyond automation: The emergence of agentic urban AI. *Automation* **2025**, *6*, 29. [CrossRef]
59. Fine, A.; Marsh, S. Judicial Leadership Matters (Yet Again): The Association between Judge and Public Trust for Artificial Intelligence in Courts. *Discov. Artif. Intell.* **2024**, *4*, 44. [CrossRef]
60. Dutta Roy, T.P. Bioethics Artificial Intelligence Advisory (BAIA): An Agentic Artificial Intelligence (AI) Framework for Bioethical Clinical Decision Support. *Cureus* **2025**, *17*, e80494. [CrossRef] [PubMed]
61. Surden, H. The Ethics of Artificial Intelligence in Law: Basic Questions. In *The Oxford Handbook of Ethics of AI*; Dubber, M.D., Pasquale, F., Das, S., Eds.; Oxford University Press: Oxford, UK, 2020; pp. 719–736.
62. Surden, H. ChatGPT, Large Language Models, and Law. *Fordham Law. Rev.* **2024**, *92*, 1941–1972.
63. Kim, T.; Peng, W. Do We Want AI Judges? The Acceptance of AI Judges' Judicial Decision-Making on Moral Foundations. *AI Soc.* **2025**, *40*, 3683–3696. [CrossRef]
64. Yudkowsky, E. Artificial Intelligence as a Positive and Negative Factor in Global Risk. In *Global Catastrophic Risks*; Bostrom, N., Ćirković, M.M., Eds.; Oxford University Press: Oxford, UK, 2008; pp. 308–345.
65. World Commission on the Ethics of Scientific Knowledge and Technology. *The Precautionary Principle*; UNESCO: Paris, France, 2005. Available online: https://unesdoc.unesco.org/ark:/48223/pf0000139578 (accessed on 10 October 2025).
66. Desai, M.; Stubbs, K.; Steinfeld, A.; Yanco, H. Creating Trustworthy Robots: Lessons and Inspirations from Automated Systems. In Proceedings of the a Symposium at the AISB 2009 Convention: New Frontiers in Human-Robot Interaction, Edinburgh, UK, 8–9 April 2009.
67. Lin, P.; Abney, K.; Jenkins, R. *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*; Oxford University Press: New York, NY, USA, 2017.
68. Kirkpatrick, J.; Hahn, E.N.; Haufler, A.J. Trust and Human–Robot Interactions. In *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*; Lin, P., Abney, K., Jenkins, R., Eds.; Oxford University Press: Oxford, UK, 2017; pp. 142–156.
69. Borenstein, J.; Howard, A.; Wagner, A.R. Pediatric Robotics and Ethics: The Robot Is Ready to See You Now, but Should It Be Trusted? In *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*; Lin, P., Abney, K., Jenkins, R., Eds.; Oxford University Press: Oxford, UK, 2017; pp. 127–141.
70. Romeo, G.; Conti, D. Exploring automation bias in human–AI collaboration: A review and implications for explainable AI. *AI Soc.* **2025**, *41*, 259–278. [CrossRef]
71. Laux, J.; Ruschemeier, H. Automation Bias in the AI Act: On the Legal Implications of Attempting to De-Bias Human Oversight of AI. *Eur. J. Risk Regul.* **2025**, *16*, 1519–1534. [CrossRef]
72. Alon-Barkat, S.; Busuioc, M. Human–AI Interactions in Public Sector Decision Making: "Automation Bias" and "Selective Adherence" to Algorithmic Advice. *J. Public Adm. Res. Theory* **2023**, *33*, 153–169. [CrossRef]
73. Kupfer, C.; Prassl, R.; Fleiß, J.; Malin, C.; Thalmann, S.; Kubicek, B. Check the box! How to deal with automation bias in AI-based personnel selection. *Front. Psychol.* **2023**, *14*, 1118723. [CrossRef] [PubMed]
74. Strauß, S. Deep automation bias: How to tackle a wicked problem of AI? *Big Data Cogn. Comput.* **2021**, *5*, 18. [CrossRef]
75. Goddard, K.; Roudsari, A.; Wyatt, J.C. Automation bias: A systematic review of frequency, effect mediators, and mitigators. *J. Am. Med. Inform. Assoc.* **2012**, *19*, 121–127. [CrossRef] [PubMed]
76. Mosier, K.L.; Skitka, L.J. Human Decision Makers and Automated Decision Aids: Made for Each Other? In *Automation and Human Performance: Theory and Applications*; Parasuraman, R., Mouloua, M., Eds.; Lawrence Erlbaum Associates: Mahwah, NJ, USA, 1996; pp. 201–220.
77. Mosier, K.L.; Skitka, L.J.; Burdick, M.D.; Heers, S.T. Automation Bias, Accountability, and Verification Behaviors. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* **1996**, *40*, 204–208. [CrossRef]
78. Parasuraman, R.; Manzey, D.H. Complacency and bias in human use of automation: An attentional integration. *Hum. Factors* **2010**, *52*, 381–410. [CrossRef] [PubMed]
79. Skitka, L.J.; Mosier, K.L.; Burdick, M. Does Automation Bias Decision-Making? *Int. J. Hum.-Comput. Stud.* **1999**, *51*, 991–1006. [CrossRef]
80. Horowitz, M.C.; Kahn, L. Bending the Automation Bias Curve: A Study of Human and AI-Based Decision Making in National Security contexts. *Int. Stud. Q.* **2024**, *68*, sqae020. [CrossRef]

81. Abdelwanis, M.; Alarafati, H.K.; Tammam, M.M.S.; Simsekler, M.C.E. Exploring the risks of automation bias in healthcare artificial intelligence applications: A Bowtie analysis. *J. Saf. Sci. Resil.* **2024**, *5*, 460–469. [CrossRef]
82. Tsai, T.L.; Fridsma, D.B.; Gatti, G. Computer Decision Aupport as a Source of Interpretation Error: The Case of Electrocardiograms. *J. Am. Med. Inform. Assoc.* **2003**, *10*, 478–483. [CrossRef] [PubMed]
83. Ruschemeier, H.; Hondrich, L.J. Automation bias in public administration–an interdisciplinary perspective from law and psychology. *Gov. Inf. Q.* **2024**, *41*, 101953. [CrossRef]
84. Skitka, L.J.; Mosier, K.L.; Burdick, M.; Rosenblatt, B. Automation Bias and Errors: Are Crews Better Than Individuals? *Int. J. Aviat. Psychol.* **2000**, *10*, 85–97. [CrossRef] [PubMed]
85. Bond, R.R.; Novotný, T.; Andršová, I.; Koc, L.; Šišáková, M.; Finlay, D.; Guldenring, D.; McLaughlin, J.; Peace, A.; McGilligan, V.; et al. Automation bias in medicine: The influence of automated diagnoses on interpreter accuracy and uncertainty when reading electrocardiograms. *J. Electrocardiol.* **2018**, *51*, S6–S11. [CrossRef] [PubMed]
86. Santiago, T. AI bias: How Does AI Influence the Executive Function of Business Leaders? *Muma Bus. Rev.* **2019**, *3*, 181–192. [CrossRef] [PubMed]
87. Dratsch, T.; Chen, X.; Mehrizi, M.R.; Kloeckner, R.; Mähringer-Kunz, A.; Püsken, M.; Baeßler, B.; Sauer, S.; Maintz, D.; dos Santos, D.P. Automation bias in mammography: The impact of artificial intelligence BI-RADS suggestions on reader performance. *Radiology* **2023**, *307*, e222176. [CrossRef] [PubMed]
88. Kim, S.H.; Schramm, S.; Riedel, E.O.; Schmitzer, L.; Rosenkranz, E.; Kertels, O.; Bodden, J.; Paprottka, K.; Sepp, D.; Renz, M.; et al. Automation bias in AI-assisted detection of cerebral aneurysms on time-of-flight MR angiography. *Radiol. Medica* **2025**, *130*, 555–566. [CrossRef] [PubMed]
89. Kücking, F.; Hübner, U.; Przysucha, M.; Hannemann, N.; Kutza, J.O.; Moelleken, M.; Erfurt-Berge, C.; Dissemond, J.; Babitsch, B.; Busch, D. Automation bias in AI-decision support: Results from an empirical study. In *German Medical Data Sciences 2024*; IOS Press: Amsterdam, The Netherlands, 2024.
90. Rosbach, E.; Ammeling, J.; Ganz, J.; Bertram, C.; Conrad, T.; Riener, A.; Aubreville, M. Stuck on suggestions: Automation bias, the anchoring effect, and the factors that shape them in computational pathology. *Mach. Learn. Biomed. Imaging* **2026**, *3*, 126–147. [CrossRef]
91. Wang, D.Y.; Ding, J.; Sun, A.L.; Liu, S.G.; Jiang, D.; Li, N.; Yu, J.K. Artificial intelligence suppression as a strategy to mitigate artificial intelligence automation bias. *J. Am. Med. Inform. Assoc.* **2023**, *30*, 1684–1692. [CrossRef] [PubMed]
92. Lyell, D.; Coiera, E. Automation bias and verification complexity: A systematic review. *J. Am. Med. Inform. Assoc.* **2017**, *24*, 423–431. [CrossRef] [PubMed]
93. Al-Anezi, F.M. Generative Artificial Intelligence in Healthcare: Automation Bias, Deskilling, and Cognitive Implications–A Systematic Review. *J. Healthc. Leadersh.* **2026**, *18*, 590498. [CrossRef] [PubMed]
94. Cummings, M.L. Automation bias in intelligent time critical decision support systems. In *Decision Making in Aviation*; Harris, D., Ed.; Routledge: Oxfordshire, UK, 2017; pp. 289–294.
95. Cummings, M.; Huang, L.; Zhu, H.; Finkelstein, D.; Wei, R. The Impact of Increasing Autonomy on Training Requirements in a UAV Supervisory Control Task. *J. Cogn. Eng. Decis. Mak.* **2019**, *13*, 295–309. [CrossRef]
96. Kahn, L.; Horowitz, M.C.; Samotin, L.R. What Is Human in Judgment? Comparing Automation Bias and Algorithm Aversion Between the United States Military Academy and the General Public. *arXiv* **2026**, arXiv:2604.04333. [CrossRef]
97. Sato, T.; Yamani, Y.; Liechty, M.; Chancey, E.T. Automation trust increases under high-workload multitasking scenarios involving risk. *Cogn. Technol. Work* **2020**, *22*, 399–407. [CrossRef]
98. Selten, F.; Robeer, M.; Grimmelikhuijsen, S. 'Just like I thought': Street-level bureaucrats trust AI recommendations if they confirm their professional judgment. *Public Adm. Rev.* **2023**, *83*, 263–278. [CrossRef]
99. Logg, J.M.; Minson, J.A.; Moore, D.A. Algorithm appreciation: People prefer algorithmic to human judgment. *Organ. Behav. Hum. Decis. Process.* **2019**, *151*, 90–103. [CrossRef]
100. Jones-Jang, S.M.; Park, Y.J. How do people react to AI failure? Automation bias, algorithmic aversion, and perceived controllability. *J. Comput.-Mediat. Commun.* **2023**, *28*, zmac029. [CrossRef]
101. Tilbury, J.; Flowerday, S. Automation bias and complacency in security operation centers. *Computers* **2024**, *13*, 165. [CrossRef]
102. Beauchamp, T.L.; Childress, J.F. *Principles of Biomedical Ethics*; Oxford University Press: Oxford, UK, 2026.
103. Friedman, B.; Kahn, P.H., Jr.; Borning, A. Value Sensitive Design and Information Systems. In *The Handbook of Information and Computer Ethics*; Himma, K.E., Tavani, H.T., Eds.; John Wiley & Sons: Hoboken, NJ, USA, 2008; pp. 69–101.
104. Friedman, B.; Kahn, P.H., Jr.; Borning, A.; Huldtgren, A. Value Sensitive Design and Information Systems. In *Early Engagement and New Technologies: Opening Up the Laboratory*; Doorn, N., Schuurbiers, D., van de Poel, I., Gorman, M.E., Eds.; Springer: Berlin/Heidelberg, Germany, 2013; pp. 55–95.
105. Sartre, J.P. *L'existentialisme est un Humanisme*; Editions Gallimard: Paris, France, 2017.
106. Merchant, B. *Blood in the Machine: The Origins of the Rebellion Against Big Tech*; Little, Brown: New York, NY, USA, 2023.

107. Nichols, T.P.; Logan, C.; Garcia, A. Generative AI and the (Re)turn to Luddism. *Learn. Media Technol.* **2025**, *50*, 379–392. [CrossRef]
108. Logan, C.; Nichols, T.P.; Garcia, A. Teach like a Luddite. *Phi Delta Kappan* **2025**, *107*, 37–41. [CrossRef]
109. Kliger, H. Here's Why a Group of NYC Teens Is Rejecting Cellphones and Social Media. CBS New York, 2024/10/7. 2024. Available online: https://www.cbsnews.com/newyork/news/nyc-students-no-technology-luddite-club/ (accessed on 25 October 2025).
110. Königs, P. What Is Techno-Optimism? *Philos. Technol.* **2022**, *35*, 63. [CrossRef]
111. Lillemäe, E.; Talves, K.; Wagner, W. Public Perception of Military AI in the Context of Techno-Optimistic Society. *AI Soc.* **2025**, *40*, 929–943. [CrossRef]
112. Danaher, J. Techno-optimism: An analysis, an evaluation and a modest defence. *Philos. Technol.* **2022**, *35*, 54. [CrossRef]
113. Anderson, M.; Anderson, S.L. Machine Ethics: Creating an Ethical Intelligent Agent. *AI Mag.* **2007**, *28*, 15–26. [CrossRef]
114. Yampolskiy, R.V. Artificial Intelligence Safety Engineering: Why Machine Ethics Is a Wrong Approach. In *Philosophy and Theory of Artificial Intelligence*; Müller, V.C., Ed.; Studies in Applied Philosophy, Epistemology and Rational Ethics; Springer: Berlin/Heidelberg, Germany, 2013; Volume 5, pp. 389–396.
115. Anderson, M.; Anderson, S.L.; Gounaris, A.; Kosteletos, G. Towards Moral Machines: A Discussion with Michael Anderson and Susan Leigh Anderson. *Conatus—J. Philos.* **2021**, *6*, 177–202. [CrossRef]
116. Russell, S.J.; Norvig, P. *Artificial Intelligence: A Modern Approach*; Prentice Hall: Hoboken, NJ, USA, 2010.
117. Buchanan, B.G.; Duda, R.O. Principles of Rule-Based Expert Systems. In *Advances in Computers*; Yovits, M.C., Ed.; Elsevier: Amsterdam, The Netherlands, 1983; Volume 22, pp. 163–216.
118. Mitchell, T.M. *Machine Learning*; McGraw-Hill: Columbus, OH, USA, 1997.
119. Yudkowsky, E. The Value Loading Problem. 2015. Available online: https://www.edge.org/response-detail/26198 (accessed on 1 October 2024).
120. Pelegrinis, T. *Ithiki Filosofia [Moral Philosophy]*; Ellinika Grammata: Athens, Greece, 1997.
121. Bentham, J. An Introduction to the Principles of Morals and Legislation. In *The Collected Works of Jeremy Bentham: An Introduction to the Principles of Morals and Legislation*; Burns, J.H., Hart, H.L.A., Eds.; Oxford University Press: Oxford, UK, 1907.
122. Mill, J.S. *Utilitarianism*; Oxford University Press: Oxford, UK, 1998.
123. Kant, I. *Groundwork of the Metaphysics of Morals*; Cambridge University Press: Cambridge, UK, 2012.
124. Skelton, A. William David Ross. In *The Stanford Encyclopedia of Philosophy*, Spring 2022 ed.; Zalta, E.N., Ed.; Stanford University: Stanford, CA, USA, 2022.
125. Gounaris, A. Can We Literally Talk About Artificial Moral Agents? Presented at the 6th Panhellenic Conference in Philosophy of Science, Department of History and Philosophy of Science, National and Kapodistrian University of Athens, Athens, Greece, 3–5 December 2020. [CrossRef]
126. Gounaris, A.; Kosteletos, G.; Kolliniati, M.A. Virtue in the machine: Beyond a one-size-fits-all approach and Aristotelian ethics for artificial intelligence. *Conatus—J. Philos.* **2025**, *10*, 127–152. [CrossRef]
127. Hao, K. Should a self-driving car kill the baby or the grandma? Depends on where you're from. *MIT Technology Review*, 24 October 2018. Available online: https://www.technologyreview.com/2018/10/24/139313/a-global-ethics-study-aims-to-help-ai-solve-the-self-driving-trolley-problem/ (accessed on 10 October 2025).
128. Awad, E.; Dsouza, S.; Kim, R.; Schulz, J.; Henrich, J.; Shariff, A.; Bonnefon, J.-F.; Rahwan, I. The Moral Machine experiment. *Nature* **2018**, *563*, 59–64. [CrossRef] [PubMed]
129. Williams, B. *Ethics and the Limits of Philosophy*; Harvard University Press: Cambridge, MA, USA, 1985.
130. Sudmann, A. (Ed.) *The Democratization of Artificial Intelligence: Net Politics in the Era of Learning Algorithms*; Transcript Verlag: Bielefeld, Germany, 2019.
131. Collingridge, D. *The Social Control of Technology*; St. Martin's Press: New York, NY, USA, 1980.
132. Marchant, G.E.; Allenby, B.R.; Herkert, J.R. *The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight: The Pacing Problem*; Marchant, G.E., Allenby, B.R., Herkert, J.R., Eds.; Springer: Dordrecht, The Netherlands, 2011.
133. Moore, G.E. Cramming more components onto integrated circuits. *Electronics* **1965**, *38*, 114–117.
134. Feenberg, A. Subversive Rationalization: Technology, Power, and Democracy. In *Technology and the Politics of Knowledge*; Feenberg, A., Hannay, A., Eds.; Indiana University Press: Bloomington, IN, USA, 1995.
135. Ziouvelou, X.; Karkaletsis, V.; Giannakopoulos, G.; Nousias, A. *Democratising AI: A National Strategy for Greece*; Institute of Informatics and Telecommunications, National Centre for Scientific Research "Demokritos": Athens, Greece, 2020. Available online: http://democratisingai.gr/index.html (accessed on 24 April 2026).
136. Radford, A.; Metz, L.; Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. In Proceedings of the 4th International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
137. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the 37th International Conference on Machine Learning, Online, 13–18 July 2020.

138. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*; MIT Press: Cambridge, MA, USA, 2018.
139. Christiano, P.F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; Amodei, D. Deep reinforcement learning from human preferences. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
140. Ross, S.; Gordon, G.; Bagnell, D. A reduction of imitation learning and structured prediction to no-regret online learning. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 11–13 April 2011.
141. Wang, M.; Leng, Z.; Hu, P.; Huang, J.; Wang, B.; Zhou, K.; Hu, G.; Yan, B.; Wu, L.; Xu, Y.; et al. FRCAE: Feature regularization meta-learning with channel-wise attention expansion. *Pattern Recognit.* **2026**, *171*, 112334. [CrossRef]
142. Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; et al. Mastering the game of Go without human knowledge. *Nature* **2017**, *550*, 354–359. [CrossRef] [PubMed]
143. Christian, B. *The Alignment Problem: Machine Learning and Human Values*; W. W. Norton & Company: New York, NY, USA, 2020.
144. Bostrom, N. The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents. *Minds Mach.* **2012**, *22*, 71–85. [CrossRef]
145. Taylor, J.; Yudkowsky, E.; LaVictoire, P.; Critch, A. *Alignment for Advanced Machine Learning Systems*; Machine Intelligence Research Institute: Berkeley, CA, USA, 2016. Available online: https://intelligence.org/files/AlignmentMachineLearning.pdf (accessed on 15 October 2025).
146. Christiano, P. Ambitious vs. Narrow Value Learning. 2015. Available online: https://ai-alignment.com/ambitious-vs-narrow-value-learning-99bd0c59847e (accessed on 25 March 2026).
147. Steunebrink, B.R.; Thórisson, K.R.; Schmidhuber, J. Growing Recursive Self-Improvers. In *Proceedings of the Artificial General Intelligence: 9th International Conference, AGI 2016, New York, NY, USA, July 16–19 2016*; Steunebrink, B., Wang, P., Goertzel, B., Eds.; Springer: Berlin/Heidelberg, Germany, 2016; pp. 129–139.
148. Wolf, M.J.; Miller, K.; Grodzinsky, F.S. Why we should have seen that coming: Comments on Microsoft's tay "experiment," and wider implications. *SIGCAS Comput. Soc.* **2017**, *47*, 54–64. [CrossRef]
149. Niu, S.; Miao, C.; Chen, G.; Wu, P.; Zhao, P. Test-time model adaptation with only forward passes. In Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria, 21–27 July 2024.
150. Iwata, T.; Kumagai, A.; Ida, Y. Meta-learning task-specific regularization weights for few-shot linear regression. In Proceedings of the 28th International Conference on Artificial Intelligence and Statistics, Mai Khao, Thailand, 3–5 May 2025.
151. Angwin, J.; Larson, J.; Mattu, S.; Kirchner, L. Machine Bias. In *Ethics of Data and Analytics: Concepts and Cases*; Martin, K., Ed.; Auerbach Publications: Boca Raton, FL, USA, 2022; pp. 254–264.
152. Joseph, J. Predicting crime or perpetuating bias? The AI dilemma. *AI Soc.* **2025**, *40*, 2319–2321. [CrossRef]
153. Kusner, M.J.; Loftus, J.; Russell, C.; Silva, R. Counterfactual fairness. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4069–4079.
154. Wang, X.; Li, Q.; Yu, D.; Li, Q.; Xu, G. Counterfactual Explanation for Fairness in Recommendation. *ACM Trans. Inf. Syst.* **2024**, *42*, 106. [CrossRef]
155. Vinge, V. The Coming Technological Singularity: How to Survive in the Post-Human Era. In *VISION-21 Symposium*; NASA Lewis Research Center: Cleveland, OH, USA, 1993.
156. Kurzweil, R. *The Singularity Is Near: When Humans Transcend Biology*; Viking: New York, NY, USA, 2005.
157. Chalmers, D.J. The Singularity: A Philosophical Analysis. *J. Conscious. Stud.* **2010**, *17*, 7–65.
158. Conn, A. *AI Aftermath Scenarios*; Future of Life Institute: Campbell, CA, USA, 2017. Available online: https://futureoflife.org/ai/ai-aftermath-scenarios/ (accessed on 10 January 2025).
159. Tegmark, M. *Life 3.0: Being Human in the Age of Artificial Intelligence*; Alfred A. Knopf: New York, NY, USA, 2017.
160. Hendrycks, D.; Mazeika, M.; Woodside, T. An overview of catastrophic AI risks. *arXiv* **2023**, arXiv:2306.12001.
161. Burrell, J. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data Soc.* **2016**, *3*, 1–12. [CrossRef]
162. Diakopoulos, N. Accountability, Transparency, and Algorithms. In *The Oxford Handbook of Ethics of AI*; Dubber, M.D., Pasquale, F., Das, S., Eds.; Oxford University Press: Oxford, UK, 2020; pp. 197–213.
163. Von Eschenbach, W.J. Transparency and the black box problem: Why we do not trust AI. *Philos. Technol.* **2021**, *34*, 1607–1622. [CrossRef]
164. Hassija, V.; Chamola, V.; Mahapatra, A.; Singal, A.; Goel, D.; Huang, K.; Hussain, A. Interpreting black-box models: A review on explainable artificial intelligence. *Cogn. Comput.* **2024**, *16*, 45–74. [CrossRef]
165. Brożek, B.; Furman, M.; Jakubiec, M.; Kucharzyk, B. The black box problem revisited. Real and imaginary challenges for automated legal decision making. *Artif. Intell. Law* **2024**, *32*, 427–440. [CrossRef]
166. Arkin, R.C. The Case for Ethical Autonomy in Unmanned Systems. *J. Mil. Ethics* **2010**, *9*, 332–341. [CrossRef]
167. Elder, A. Robot Friends for Autistic Children: Monopoly Money or Counterfeit Currency? In *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*; Lin, P., Abney, K., Jenkins, R., Eds.; Oxford University Press: Oxford, UK, 2017; pp. 113–126.

168. *Aristotle. Nicomachean Ethics*; Hackett Publishing Company: Indianapolis, IN, USA, 2000.
169. Paraskevopoulos, N. *O epieikis Algorithmos: Apo Tin Aristoteliki Skepsi Stin Techniti Noimosyni [The Equitable Algorithm: From Aristotelian Thought to Artificial Intelligence]*; Institouto Neoellinikon Spoudon—Idrima Triantafyllidi: Thessaloniki, Greece, 2024.
170. Re, R.M.; Solow-Niederman, A. Developing artificially intelligent justice. *Stanf. Technol. Law Rev.* **2019**, *22*, 242–289.
171. Solum, L.B. Artificially intelligent law. *BioLaw J. Riv. DI BioDiritto* **2019**, *1*, 53–62.
172. Sinnott-Armstrong, W. Framing Moral Intuitions. In *Moral Psychology: The Cognitive Science of Morality: Intuition and Diversity*; Sinnott-Armstrong, W., Ed.; MIT Press: Cambridge, MA, USA, 2008; Volume 2, pp. 47–76.
173. Grace, K. Superintelligence 20: The Value-Loading Problem. 2015. Available online: https://www.lesswrong.com/posts/FP8T6rdZ3ohXxJRto/superintelligence-20-the-value-loading-problem (accessed on 10 October 2024).
174. Searle, J.R. Minds, Brains, and Programs. *Behav. Brain Sci.* **1980**, *3*, 417–424. [CrossRef]
175. McCarthy, J.; Hayes, P.J. Some Philosophical Problems from the Standpoint of Artificial Intelligence. In *Machine Intelligence 4*; Meltzer, B., Michie, D., Eds.; Edinburgh University Press: Scotland, UK, 1969; pp. 463–502.
176. Pelegrinis, T. *Lexiko tis Philosophias [Dictionary of Philosophy]*; Ellinika Grammata: Athens, Greece, 2004.
177. Hauser, M.D.; Young, L.; Cushman, F. Reviving Rawls's Linguistic Analogy: Operative Principles and the Causal Structure of Moral Actions. In *Moral Psychology*; The Cognitive Science of Morality: Intuition and Diversity; Sinnott-Armstrong, W., Ed.; MIT Press: Cambridge, MA, USA, 2008; Volume 2, pp. 107–144.
178. Turing, A.M. On Computable Numbers, with an Application to the Entscheidungsproblem. *Proc. Lond. Math. Soc. Ser. 2* **1937**, *42*, 230–265. [CrossRef]
179. Turing, A.M. *Intelligent Machinery*; National Physical Laboratory: Teddington, UK, 1947.
180. Turing, A.M. Computing Machinery and Intelligence. *Mind* **1950**, *59*, 433–460. [CrossRef]
181. Newell, A.; Simon, H.A. *Current Developments in Complex Information Processing*; P-850; RAND Corporation: Santa Monica, CA, USA, 1956.
182. Newell, A.; Shaw, J.C. Programming the Logic Theory Machine. In *Papers Presented at the February 26–28, 1957, Western Joint Computer Conference: Techniques for Reliability*; Association for Computing Machinery: New York, NY, USA, 1957.
183. Newell, A.; Simon, H. The logic theory machine--A complex information processing system. *IEEE Trans. Inf. Theory* **1956**, *2*, 61–79. [CrossRef]
184. Newell, A.; Simon, H.A. GPS: A Program that Simulates Human Thought. In *Lernende Automaten*; Billing, H., Ed.; R. Oldenbourg: Munich, Germany, 1961; pp. 109–124.
185. Newell, A.; Shaw, J.C.; Simon, H.A. Elements of a Theory of Human Problem Solving. *Psychol. Rev.* **1958**, *65*, 151–166. [CrossRef]
186. Newell, A.; Simon, H.A. Computer Science as Empirical Inquiry: Symbols and Search. *Commun. ACM* **1976**, *19*, 113–126. [CrossRef]
187. Newell, A. Physical Symbol Systems. *Cogn. Sci.* **1980**, *4*, 135–183. [CrossRef]
188. Anderson, S.L.; Anderson, M. A Prima Facie Duty Approach to Machine Ethics: Machine Learning of Features of Ethical Dilemmas, Prima Facie Duties, and Decision Principles through a Dialogue with Ethicists. In *Machine Ethics*, 1st ed.; Anderson, M., Anderson, S.L., Eds.; Cambridge University Press: Cambridge, UK, 2011; pp. 476–492.
189. Heidegger, M. *The Basic Problems of Phenomenology*; Indiana University Press: Bloomington, IN, USA, 1982.
190. Merleau-Ponty, M. *Phenomenology of Perception*; Routledge & Kegan Paul: London, UK, 1962.
191. Gibson, J.J. *The Ecological Approach to Visual Perception*; Houghton Mifflin: Boston, MA, USA, 1979.
192. Piaget, J. The Theory of Stages in Cognitive Development. In *Measurement and Piaget*; Green, D.R., Ford, M.P., Flamer, G.B., Eds.; McGraw-Hill: Columbus, OH, USA, 1971; pp. 1–11.
193. Gounaris, A. Why do we need a Unified Theory of Embodied Cognition? In *Proceedings of the 94th Joint Session of the Mind Association and the Aristotelian Society, July 2020*; University of Kent: Canterbury, UK, 2020. [CrossRef]
194. Shapiro, L. *Embodied Cognition*; Routledge: Oxfordshire, UK, 2019.
195. Dreyfus, H.L. *What Computers Still Can't Do: A Critique of Artificial Reason*; MIT Press: Cambridge, MA, USA, 1992.
196. Gounaris, A.; Abou Elkheir, G. Why Embodied Artificial Intelligence Is Not So Embodied? *Proc. XXIII World Congr. Philos.* **2018**, *45*, 9–14. [CrossRef]
197. Moore, G.E. *Principia Ethica*; Cambridge University Press: Cambridge, UK, 1903.
198. Ross, W.D. *The Right and the Good*; Clarendon Press: Oxford, UK, 1930.
199. Wittgenstein, L. *Tractatus Logico-Philosophicus*; Kegan Paul, Trench, Trubner & Co., Ltd.: London, UK, 1922.
200. Haidt, J.; Joseph, C. Intuitive Ethics: How Innately Prepared Intuitions Generate Culturally Variable Virtues. *Daedalus* **2004**, *133*, 55–66. [CrossRef]
201. Graham, J.; Nosek, B.A.; Haidt, J.; Iyer, R.; Koleva, S.; Ditto, P.H. Mapping the Moral Domain. *J. Personal. Soc. Psychol.* **2011**, *101*, 366–385. [CrossRef] [PubMed]
202. Davis, M. *Computability and Unsolvability*; McGraw-Hill: Columbus, OH, USA, 1958.
203. Minsky, M.L. *Computation: Finite and Infinite Machines*; Prentice-Hall: Hoboken, NJ, USA, 1967.

204. Biswas, B.; Sarkar, S. Responsible Agentic Artificial Intelligence Governance: Risk, Safety, and Ethical Challenges in Autonomous Systems. *Int. J. Appl. Resil. Sustain.* **2026**, *2*, 142–167. [CrossRef]
205. Brohi, S.; Mastoi, Q.U.A.; Jhanjhi, N.Z.; Pillai, T.R. A Research Landscape of Agentic AI and Large Language Models: Applications, Challenges and Future Directions. *Algorithms* **2025**, *18*, 499. [CrossRef]
206. Gridach, M.; Nanavati, J.; Zine El Abidine, K.; Yacoubian, C.; Mack, C. Agentic AI in Healthcare: Opportunities, Challenges, and Future Directions. *ACM Comput. Surv.* **2026**, *58*, 1–30. [CrossRef]
207. Pati, A.K. Agentic AI: A comprehensive survey of technologies, applications, and societal implications. *IEEE Access* **2025**, *13*, 151824–151837. [CrossRef]
208. Bellogín, A.; Giudici, P.; Larsson, S.; Pang, J.; Schimpf, G.; Sengupta, B.; Solmaz, G. *Systemic Risks Associated with Agentic AI: A Policy Brief*; Association for Computing Machinery (ACM): New York, NY, USA, 2025. Available online: https://www.acm.org/binaries/content/assets/public-policy/europe-tpc/systemic_risks_agentic_ai_policy-brief_final.pdf (accessed on 20 October 2025).
209. Bowen, G. Agentic Artificial Intelligence: Legal and Ethical Challenges of Autonomous Systems. *J. Digit. Technol. Law* **2025**, *3*, 431–445. [CrossRef]
210. Hacker, P.; Holweg, M. A Pragmatic Approach to Regulating AI Agents. *arXiv* **2026**, arXiv:2604.22819. [CrossRef]
211. Inpun, J.; Cholamjiak, W.; Phrueksawatnon, P.; Shiangjen, K. A Generalizable Agentic AI Pipeline for Developing Chatbots Using Small Language Models: A Case Study on Thai Student Loan Fund Services. *Computation* **2025**, *13*, 297. [CrossRef]
212. Hahn, M.; Tretter, M.; Dabrock, P. Ethical perspectives on AI Agents and Agentic AI. *AI Ethics* **2026**, *6*, 218. [CrossRef]
213. Cavallo, P. *Described vs. Established Governance in Agentic AI: Closing the Gap Between Policy and Enforcement*; Social Science Research Network: Amsterdam, The Netherlands, 2026. [CrossRef]
214. Gardhouse, K.; Oueslati, A.; Kolt, N. Regulating AI Agents. *arXiv* **2026**, arXiv:2603.23471. [CrossRef]
215. Joshi, S. *Framework for Government Policy on Agentic and Generative AI: Governance, Regulation, and Risk Management*; Social Science Research Network: Amsterdam, The Netherlands, 2025. [CrossRef]
216. Nannini, L.; Smith, A.L.; Maggini, M.J.; Panai, E.; Feliciano, S.; Tiulkanov, A.; Maran, E.; Gealy, J.; Bisconti, P. AI Agents Under EU Law: A Compliance Architecture for AI Providers. *arXiv* **2026**, arXiv:2604.04604. [CrossRef]
217. Osmond, M.; Jego, T. Mind The Gap: How The Technical Mechanism Of Agentic AI Outpace Global Legal Frameworks. *arXiv* **2026**, arXiv:2603.27075. [CrossRef]
218. Ignjatović Pertini, A.; Vujko, A. Artificial Intelligence Leadership and the Decoupling of Human Agency: Evidence of Misaligned Innovation in Agentic AI Systems. *Adm. Sci.* **2026**, *16*, 239. [CrossRef]
219. Mukherjee, A.; Chang, H.H. Agentic AI: Autonomy, Accountability, and the Algorithmic Society. *arXiv* **2025**, arXiv:2502.00289. [CrossRef]
220. Borghoff, U.M.; Bottoni, P.; Pareschi, R. Beyond prompt chaining: The tb-cspn architecture for agentic ai. *Future Internet* **2025**, *17*, 363. [CrossRef]
221. Nguyen, M.H.; Nguyen, D.H.; O'Sullivan, B.; Nguyen, H.D. On Controllability in Agentic AI: A Survey. *Minds Mach.* **2026**, *36*, 29. [CrossRef]
222. De Sio, F.S.; Mecacci, G. Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philos. Technol.* **2021**, *34*, 1057–1084. [CrossRef]
223. Bryson, J.J. Robots should be slaves. In *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*; Wilks, Y., Ed.; John Benjamins Publishing: Amsterdam, The Netherlands, 2010; pp. 63–74.
224. Bryson, J.J.; Diamantis, M.E.; Grant, T.D. Of, for, and by the people: The legal lacuna of synthetic persons. *Artif. Intell. Law* **2017**, *25*, 273–291. [CrossRef]
225. Bryson, J.J. Patiency Is Not a Virtue: The design of intelligent systems and systems of ethics. *Ethics Inf. Technol.* **2018**, *20*, 15–26. [CrossRef]
226. McCarthy, J. Making Robots Conscious of Their Mental States. In *Machine Intelligence 15: Intelligent Agents*; Furukawa, K., Michie, D., Muggleton, S., Eds.; Oxford University Press: Oxford, UK, 2000; pp. 3–17.
227. Miller, L.F. Granting automata human rights: Challenge to a basis of full-rights privilege. *Hum. Rights Rev.* **2015**, *16*, 369–391. [CrossRef]
228. Yampolskiy, R.V. *Artificial Superintelligence: A Futuristic Approach*; CRC Press: Boca Raton, FL, USA, 2016. [CrossRef]
229. Avramides, A. *Other Minds*; Routledge: Oxfordshire, UK, 2000. [CrossRef]
230. Avramides, A. Other Minds. In *The Oxford Handbook of Philosophy of Mind*; McLaughlin, B.P., Beckermann, A., Walter, S., Eds.; Oxford University Press: Oxford, UK, 2009; pp. 727–740.
231. Avramides, A. Other minds. In *The Stanford Encyclopedia of Philosophy*, Winter 2020 ed.; Zalta, E.N., Ed.; Metaphysics Research Lab, Stanford University: Stanford, CA, USA, 2020.
232. Gunkel, D.J. *Robot Rights*; MIT Press: Cambridge, MA, USA, 2018.

233. Gounaris, A. Mia Eisagoge sti Noisi kai sti Syneidisi ton Zoon [An Introduction to Animal Cognition and Consciousness]. In *Noisi kai Syneidisi sta Zoa [Cognition and Consciousness in Animals]*; Gounaris, A., Ed.; The NKUA Applied Philosophy Research Lab Press: Athens, Greece, 2026; pp. 11–77.
234. Coeckelbergh, M. Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Sci. Eng. Ethics* **2020**, *26*, 2051–2068. [CrossRef] [PubMed]
235. European, C. *High-Level Expert Group on Artificial, I. Ethics Guidelines for Trustworthy AI*; Publications Office of the European Union: Luxembourg, 2019. Available online: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (accessed on 25 March 2026).
236. European, P. *Civil Law Rules on Robotics: European Parliament Resolution with Recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL))*; European Union: Luxembourg, 2017.
237. Delvaux, M. *Draft Report with Recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL))*; PE582.443v01-00; Committee on Legal Affairs, European Parliament: Brussels, Belgium, 2016. Available online: https://www.europarl.europa.eu/doceo/document/JURI-PR-582443_EN.pdf (accessed on 25 March 2026).
238. Gounaris, A.; Kosteletos, G. Licensed to kill: Autonomous weapons as persons and moral agents. In *Personhood*; Prole, D., Rujević, G., Eds.; Filozofski Fakultet, Univerzitet u Novom Sadu: Novi Sad, Serbia; The NKUA Applied Philosophy Research Lab Press: Athens, Greece, 2020; pp. 137–189.
239. Güzeldere, G. The many faces of consciousness: A field guide. In *The Nature of Consciousness: Philosophical Debates*; Block, N., Flanagan, O., Güzeldere, G., Eds.; MIT Press: Cambridge, MA, USA, 1997; pp. 1–67.
240. Levy, D. The ethical treatment of artificially conscious robots. *Int. J. Soc. Robot.* **2009**, *1*, 209–216. [CrossRef]
241. Gunkel, D.J. The other question: Can and should robots have rights? *Ethics Inf. Technol.* **2018**, *20*, 87–99. [CrossRef]
242. Calverley, D. Toward a method for determining the legal status of a conscious machine. In Proceedings of the AISB 2005 Symposium on Next Generation Approaches to Machine Consciousness: Imagination, Development, Intersubjectivity, and Embodiment, Hatfield, UK, 12–15 April 2005.
243. Königs, P. No wellbeing for robots (and hence no rights). *Am. Philos. Q.* **2025**, *62*, 191–208. [CrossRef]
244. Xie, J. An explanation of the relationship between artificial intelligence and human beings from the perspective of consciousness. *Cult. Sci.* **2021**, *4*, 124–134. [CrossRef]
245. Mosakas, K. Machine Moral Status: Moral Properties and the Consciousness Criterion. In *Rights for Intelligent Robots? A Philosophical Inquiry into Machine Moral Status*; Mosakas, K., Ed.; Springer Nature Switzerland: Cham, Switzerland, 2024; pp. 115–177.
246. Shen, Y.; Gong, Z. Reflections on the Moral Status of Robots—The Establishment of Moral Status Standards for Robots in the Era of Strong Artificial Intelligence. *Int. J. Front. Sociol.* **2023**, *5*, 130–134. [CrossRef]
247. Dennett, D.C. Are we explaining consciousness yet? *Cognition* **2001**, *79*, 221–237. [CrossRef] [PubMed]
248. Goertzel, B. Intelligence, Mind and Self-Modification: Defining the Core Concepts of AI. 2002. Available online: https://www.goertzel.org/papers/IntelligenceAndSelfModification.htm (accessed on 29 December 2025).
249. Schwitzgebel, E.; Garza, M. A defense of the rights of artificial intelligences. *Midwest Stud. Philos.* **2015**, *39*, 98–119. [CrossRef]
250. Singer, P. *Animal Liberation*; Random House: Southbank, Australia, 1975.
251. Singer, P. *Practical Ethics*; Cambridge University Press: Cambridge, UK, 1993.
252. Brooks, R.A. *Flesh and Machines: How Robots Will Change Us*; Pantheon Books: New York, NY, USA, 2003.
253. De Quincey, C. Switched-on consciousness: Clarifying what it means. *J. Conscious. Stud.* **2006**, *13*, 7–12.
254. Velmans, M. *Understanding Consciousness*; Routledge: Oxfordshire, UK, 2009.
255. Ashrafian, H. Artificial intelligence and robot responsibilities: Innovating beyond rights. *Sci. Eng. Ethics* **2015**, *21*, 317–326. [CrossRef] [PubMed]
256. DeGrazia, D. Robots with moral status? *Perspect. Biol. Med.* **2022**, *65*, 73–88. [CrossRef] [PubMed]
257. Gibert, M.; Martin, D. In search of the moral status of AI: Why sentience is a strong argument. *AI Soc.* **2022**, *37*, 319–330. [CrossRef]
258. Schwitzgebel, E. AI systems must not confuse users about their sentience or moral status. *Patterns* **2023**, *4*, 100818. [CrossRef] [PubMed]
259. Warren, M.A. On the moral and legal status of abortion. In *Contemporary Moral Problems*; White, J.E., Ed.; Wadsworth/Thomson Learning: Belmont, CA, USA, 2003; pp. 144–155.
260. Goertzel, B. Thoughts on AI morality. *Dyn. Psychol. Int. Interdiscip. J. Complex Ment. Process.* **2002**. [CrossRef]
261. Sparrow, R. Killer robots. *J. Appl. Philos.* **2007**, *24*, 62–77. [CrossRef]
262. Frankfurt, H.G. *The Importance of What We Care About: Philosophical Essays*; Cambridge University Press: Cambridge, UK, 1988.
263. Bratman, M.E. *Structures of Agency: Essays*; Oxford University Press: Oxford, UK, 2007. [CrossRef]
264. Gounaris, A. Can we talk about Intentionality in Eliminative Materialism? The point of view of Embodied Cognition Theories. Presented at the 3rd National Conference on the Philosophy of Science, National and Kapodistrian University of Athens, Athens, Greece, 27–29 November 2014. [CrossRef]

265. Kosteletos, G. I Mousiki Dokimasia Turing [The Turing Test in Music]. *Mousikologia* **2015**, *15*, 290–300.
266. Horn, R.E. The Turing test: Mapping and navigating the debate. In *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*; Epstein, R., Roberts, G., Beber, G., Eds.; Springer: Berlin/Heidelberg, Germany, 2008; pp. 73–88.
267. Kosteletos, G.; Georgaki, A. A Turing Test for the Singing Voice as an Anthropological Tool: Epistemological and Technical Issues. In Proceedings of the 38th ICMC 2012, Ljubljana, Slovenia, 9–14 September 2012.