This is an Accepted Manuscript for *Calitatea Vieții*. Subject to change during the editing and production process.

THE LIMITS OF ETHICAL ARTIFICIAL INTELLIGENCE: MORAL AGENCY, NON-DELEGABLE RESPONSIBILITY AND SOCIOTECHNICAL GOVERNANCE

PAULA VALENTE
DANIEL MILL
DANIELA BARROS
ANA NOBRE
DIENE EIRE DE MELLO
ELIANE CRISTINA FLEXA DUARTE

This article critically examines the conceptual legitimacy of the expression "ethical artificial intelligence" by asking whether artificial systems can be regarded as moral agents in any robust normative sense. Adopting a conceptual and theoretically argumentative design, and drawing on a critical narrative review of the literature on AI ethics, moral agency, algorithmic formalisation and sociotechnical governance, the article argues that AI systems may generate normatively relevant outputs, simulate morally acceptable decisions and support human deliberation, but they cannot assume moral responsibility, provide normative justification for their decisions or answer to those affected by their operation. The analysis distinguishes between ethics as a property of moral agents and ethics as a quality of sociotechnical systems designed, deployed, and governed by human actors. From a responsibility-centred, discourse-ethical and pluralist perspective, the article contends that the impossibility of fully ethical machines does not invalidate AI ethics. Rather, it redefines its scope by relocating ethical responsibility from artificial systems themselves to the human and institutional arrangements that design,

Address of the corresponding authors: Paula Valente, Universidade Aberta, Rua da Escola Politécnica, No. 141, Lisbon, Portugal, e-mail: paula.silva@uab.pt (https://orcid.org/0000-0002-6022-6402); **Daniel Mill**, Universidade Federal de São Carlos, Rodovia Washington Luiz, 235, São Carlos, São Paulo, Brazil, e-mail: mill@ufscar.br (https://orcid.org/0000-0002-8336-3645); **Daniela Barros**, Universidade Aberta, Rua da Escola Politécnica, No. 141, Lisbon, Portugal, e-mail: daniela.barro@uab.pt (https://orcid.org/0000-0002-1412-2231); **Ana Nobre**, Universidade Aberta, Rua da Escola Politécnica, No. 141, Lisbon, Portugal, e-mail: ana.nobre@uab.pt (https://orcid.org/0000-0002-9902-1850); **Diene Eire de Mello**, Universidade Estadual de Londrina, Rodovia Celso Garcia Cid, km 380, Londrina, Paraná, Brazil, e-mail: diene.eire@uel.br (https://orcid.org/0000-0001-6048-8130); **Eliane Cristina Flexa Duarte**, Universidade Federal do Oeste do Pará, Av. Marechal Rondon, s/n, Caranazal, Santarém, Pará, Brazil, e-mail: eliane.duarte@ufopa.edu.br (https://orcid.org/0000-0003-2814-165X).

CALITATEA VIEȚII, XXXVII, nr. 3, 2026, pp. , https://doi.org/10.46841/RCV.2026.03.01

authorise, supervise, and contest their use. The article concludes that the central challenge is not to construct morally responsible machines, but to prevent automation from producing zones of opacity, normative displacement, and human irresponsibility in socially consequential decision-making.

Keywords: *Artificial Intelligence; AI Ethics; Moral Agency; Human Responsibility; Algorithmic Governance.*

INTRODUCTION: THE PROBLEM OF "ETHICAL AI"

Contemporary debates on the ethics of artificial intelligence (AI) have intensified as algorithmic systems have come to intervene increasingly in sensitive domains of social life, including healthcare, justice, education, and the circulation of information (Floridi 2013; Jobin *et al.* 2019). The diffusion of these technologies has been accompanied by a proliferation of normative principles, codes of conduct and regulatory frameworks designed to guide their development and use (UNESCO 2021; European Parliament and Council of the European Union 2024). Yet, despite this consolidation, a conceptually decisive question remains unresolved: in what sense can ethical status be attributed to artificial intelligence systems?

This question cannot be reduced to a technical problem, nor can it be resolved merely by improving system performance. It is not simply a matter of determining whether algorithms can follow rules, optimise decisions, or produce outcomes considered desirable. The issue is more demanding: can an artificial system be regarded as a moral agent? In other words, can AI be understood as an entity capable of responsibility, justification, and ethical judgement?

Much of the literature has addressed this challenge through two main strategies. The first seeks to embed ethical principles into artificial systems, either through explicit rules or through learning models capable of aligning behaviour with human values. The second emphasises the need for governance, supervision, and regulation mechanisms capable of ensuring the responsible use of these technologies. Although these approaches have contributed significantly to the development of the field, both tend, to varying degrees, to assume that ethics can be treated as a problem of alignment, compliance, or technical control.

This article proceeds from a different hypothesis. Rather than assuming that the central task is to make artificial intelligence ethical, it interrogates the conceptual legitimacy of the expression "ethical AI" itself. The analysis is grounded in the premise that any response to this problem requires prior clarification of what is meant by ethics and moral agency. If ethics entails imputable responsibility, normative justification and situated judgement within pluralistic contexts, then it is necessary to examine the extent to which such attributes can, or cannot, be ascribed to artificial systems.

The argument advanced here is that AI systems cannot be considered ethical in the full sense of the term when ethics is understood as a practice of responsible judgement, normative justification, and answerability before others. AI systems

may follow programmed parameters, optimise predefined objectives, and generate outputs that appear morally acceptable. They do not, however, possess moral consciousness, autonomous normative understanding, or the capacity to justify their decisions ethically. Although AI may support decision-making processes, it operates primarily based on data, patterns and probabilities, without human experience or contextual consciousness.

This limitation does not render ethics irrelevant in the field of AI. On the contrary, it displaces the problem. The ethics of artificial intelligence should not be conceived as a property of the machine, but as the responsibility of the human, institutional and sociotechnical systems that design, implement, and use it. The contribution of this article lies in advancing the debate through a double movement: first, by clarifying the notions of ethics, responsibility, moral agency, and formalisation; and second, by shifting the focus from the moralisation of the machine to the construction of institutional arrangements capable of ensuring responsibility, transparency, contestability and reparation in contexts mediated by artificial intelligence.

More specifically, the article's contribution lies in showing that the problem of "ethical AI" is not exhausted by the question of whether artificial systems can generate normatively acceptable outputs. The decisive issue is where ethical responsibility should be located when decisions are mediated by AI. From this perspective, the conceptual critique of "ethical AI" is inseparable from the claim that responsibility remains non-delegable and from an understanding of AI governance as a sociotechnical arrangement. Ethics is therefore not transferred to the artificial artefact, but remains located in the human and institutional practices through which AI systems are designed, authorised, supervised, contested and corrected (Mittelstadt 2019). This perspective explains why technical alignment, declarations of principle and regulatory compliance are insufficient when they are detached from identifiable actors and institutions capable of responding to those affected by AI-mediated decisions.

METHODOLOGICAL APPROACH

This article is designed as a conceptual study with a theoretically argumentative orientation, supported by a critical narrative review of the literature in the fields of artificial intelligence ethics, moral philosophy, and sociotechnical studies. Its central aim is to clarify concepts, reconstruct lines of argument and advance a normative position on the ethical status of artificial intelligence. In this respect, the article is aligned with Jaakkola's (2020) understanding of conceptual research, according to which scientific contribution may reside in the development of analytical frameworks and argumentative syntheses.

The choice of a conceptual methodological design follows from the nature of the problem under investigation. The article examines the attribution of ethical

status to artificial intelligence through normative categories such as responsibility, justification, imputability and moral judgement, none of which can be resolved through direct empirical inquiry alone. The problem therefore requires philosophical analysis, conceptual precision, and the critical articulation of different theoretical traditions.

The study is guided by the following central research question: if ethics entails responsibility, normative justification, contextualisation and imputability, in what sense can artificial intelligence be considered ethical? This question unfolds into three interdependent analytical axes: the possibility of attributing moral agency to artificial systems; the limits of formalising ethics into decision functions; and the implications of this analysis for the construction of an ethics of AI centred on human responsibility and sociotechnical governance.

The literature review was conducted based on conceptual relevance, theoretical authority, currency, and analytical pertinence. Rather than pursuing exhaustiveness, the review privileged the articulation of different traditions and domains. Four main bodies of literature were mobilised: philosophical literature on ethics, responsibility, moral pluralism and practical judgement; foundational and contemporary work on machine ethics and artificial moral agency, including Moor (2006) and Wallach and Allen (2009); recent studies on AI principles, challenges and governance, with particular attention to Floridi (2013) and Floridi *et al.* (2018); and international normative documents, including the Recommendation on the Ethics of Artificial Intelligence (UNESCO 2021), the European Artificial Intelligence Act (European Parliament and Council of the European Union 2024) and the European Commission guidelines on the responsible use of generative AI in research (European Commission 2026).

The literature was selected from three related areas. The first concerned moral agency, responsibility, justification and practical judgement. The second concerned machine ethics and artificial moral agency, especially texts that distinguish norm-conforming behaviour from moral agency in a stronger sense. The third concerned AI governance, algorithmic accountability and institutional guidance, particularly texts that link ethical principles to oversight, contestability and responsibility in organisational settings. Sources were excluded when they approached AI ethics mainly as a matter of technical performance, general innovation policy or regulatory description, without engaging the conceptual problem of agency, responsibility or justification.

The philosophical traditions mobilised in the article were selected for their analytical role, not for historical or bibliometric representativeness. Discourse ethics, theories of responsibility, moral pluralism and accounts of situated judgement are used because they help to clarify how ethical claims can be justified, contested and attributed to agents. The review does not claim to cover all empirical, legal or intercultural debates on AI ethics. This is a limitation of a critical narrative review. At the same time, this selective approach allows the

article to clarify concepts and reconstruct a normative argument that can later be specified, examined or challenged in sectoral studies.

The analysis proceeds through three complementary methodological procedures (Holmes *et al.* 2019). The first consists of conceptual reconstruction, aimed at distinguishing categories that are often used interchangeably, such as ethics, morality, normativity, moral agency, explanation, and justification. The second consists of the critical comparison of positions, through which approaches that defend the possibility of ethical machines are examined alongside arguments that reject the attribution of moral agency to artificial systems. The third consists of a normative synthesis, connecting the results of the conceptual analysis with an ethical framing of artificial intelligence centred on human responsibility and institutional governance.

In this article, the term artificial intelligence is used to designate computational systems based on rules, machine learning or statistical models that are capable of classifying, recommending, predicting or generating outputs from data, with varying degrees of operational autonomy, but without presupposing consciousness, intentionality or moral agency.

THE CONTEMPORARY DEBATE ON AI ETHICS: PRINCIPLES, LIMITS AND GOVERNANCE

The contemporary debate on AI ethics has crystallised around three closely interrelated developments: the formulation of principles, the recognition of their limitations and the movement towards models of sociotechnical governance. The literature points to a broad convergence around values such as transparency, fairness, responsibility, privacy and beneficence (Floridi 2013; Jobin *et al.* 2019), while also drawing attention to the difficulty of translating these principles into effective mechanisms of implementation, accountability and monitoring (Khan *et al.* 2022; Corrêa *et al.* 2023). Recent studies have therefore increasingly shifted attention from abstract normative principles to governance structures capable of addressing risks, safety, data protection and the concrete social effects of AI (Farisco *et al.* 2020; Larsson 2019; Korobenko *et al.* 2024).

This reconfiguration signals a deeper transformation: AI ethics is moving from the level of normative formulation to that of institutional and sociotechnical infrastructure. Rather than asking only which principles should guide AI, it is now necessary to ask how such principles can be translated into effective practices, who is responsible for their implementation and which mechanisms allow automated decisions to be evaluated, contested and corrected (Ménissier 2023). Recent literature therefore converges on the view that AI ethics requires governance arrangements capable of articulating responsibility, transparency, auditability and public oversight (Mittelstadt *et al.* 2016; Batool *et al.* 2025; Papagiannidis *et al.* 2025).

Despite these developments, a fundamental conceptual gap remains. Much of the literature implicitly assumes that AI ethics consists in aligning artificial systems with human values, whether through the incorporation of principles or through mechanisms of supervision and control. Yet this approach seldom confronts directly a more radical question: *in what sense is the very expression "ethical AI" conceptually legitimate?* Before discussing how artificial systems might be made ethical, it is therefore necessary to clarify whether such systems can, in a strict sense, be regarded as moral agents.

The present article intervenes in precisely this gap. Its guiding hypothesis is that recurrent practical difficulties, including the insufficiency of principles and the challenges of implementation, point to a more fundamental distinction between systems capable of producing normatively relevant outputs and agents capable of moral responsibility. The trajectory of the contemporary debate, from the formulation of principles to the emphasis on governance, may therefore be read as the symptom of an incomplete displacement. The insufficiency of principles does not merely reveal failures of implementation; it also exposes the limits of an approach that treats ethics as something that can be incorporated directly into technical systems.

WHAT IS AT STAKE WHEN WE SPEAK OF ETHICS?

The central question emerging from contemporary debates on artificial intelligence, in what sense one may speak of "ethical AI", cannot be adequately addressed without first clarifying the concept of ethics itself. The term is often used indiscriminately to refer to normative principles, regulatory compliance, good organisational practices or diffuse social expectations. Although this polysemy is understandable within the interdisciplinary field of AI ethics, it tends to obscure a fundamental question: *what characterises ethics as a specific domain of normativity?*

An initial analytical distinction is required between ethics and morality. Morality may be understood as the set of norms, values and beliefs that guide action in specific social contexts, often anchored in cultural, religious or historical traditions. Ethics, in a stricter sense, corresponds to a reflective and critical practice through which such norms are evaluated, publicly justified, revised and, where necessary, transformed. This distinction, found in authors such as Ricoeur (1990) and Habermas (1990), allows ethics to be understood not as a closed system of rules, but as an argumentative process in which norms are confronted, reinterpreted and justified in the light of concrete situations.

This characterisation brings ethics close to other forms of normativity, while also requiring it to be distinguished from them. Law constitutes an institutionalised system of coercive norms, endowed with formal mechanisms of application and sanction. Although law frequently incorporates ethical principles, it is not reducible

to them and may even come into tension with ethical demands in contexts of institutional injustice. Professional deontological codes, in turn, establish normative standards for specific practices, functioning as a mediation between general ethical principles and concrete institutional contexts. Technical regulation operates at the level of procedural compliance, defining operational requirements, safety standards and performance criteria.

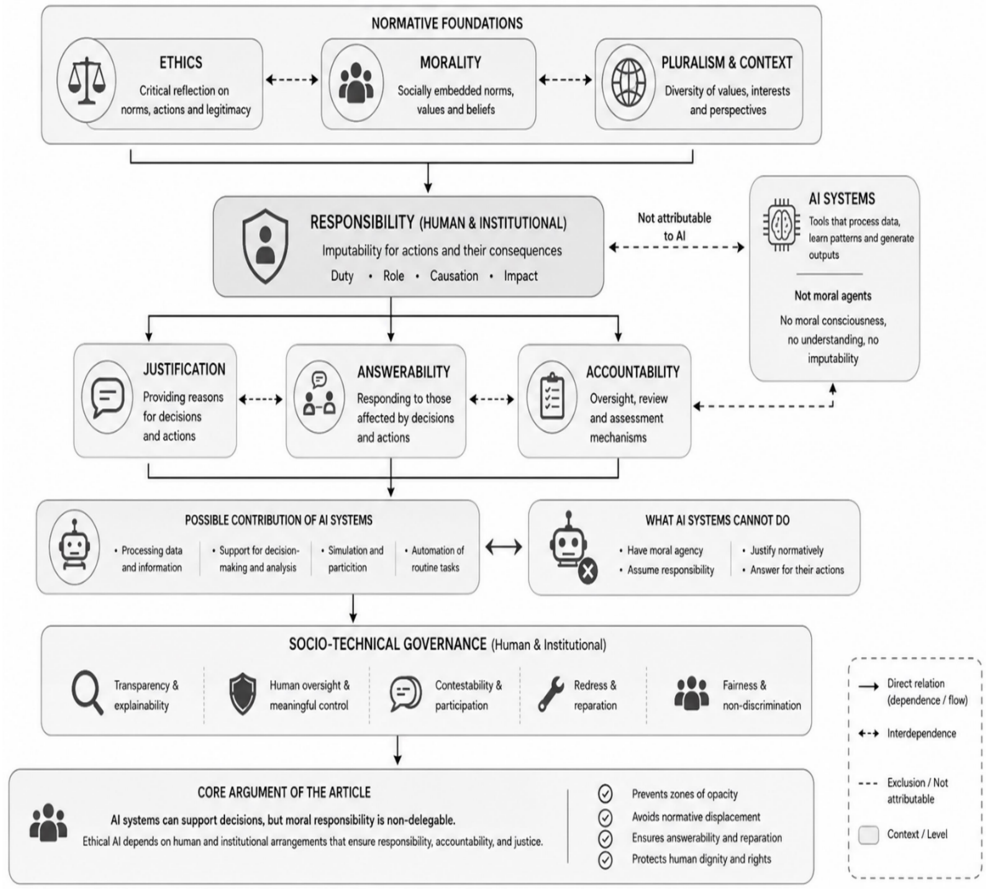
Ethics differs from these forms of normativity because it is not limited to the application of previously established norms. It interrogates the legitimacy of those norms and their adequacy to concrete situations of conflict, uncertainty and pluralism. In this sense, ethics is not only normative but also meta-normative: it questions the foundations of norms and guides their application in contexts marked by uncertainty, ambiguity and disagreement.

Jonas (1984) places responsibility at the centre of ethical reflection on the consequences of human action, while Bostrom and Yudkowsky (2014) are more directly concerned with value alignment and long-term risks in artificial intelligence. Taken together, these perspectives indicate why the ethical evaluation of AI cannot be reduced to technical performance, although they do not address responsibility in exactly the same sense. To act ethically does not simply mean conforming to rules; it means assuming authorship of one's decisions and recognising oneself as answerable to others for their effects. Ethical responsibility therefore involves three interdependent dimensions: imputability, understood as the possibility of attributing an action to an agent; justification, understood as the capacity to offer reasons for that action; and answerability, understood as the willingness and capacity to respond to the consequences of that action, including criticism, revision and reparation. Accountability, in turn, refers here to the institutional mechanisms through which responsibility is reviewed, assessed and enforced. These dimensions are not merely formal. They presuppose the existence of an agent capable of recognising others as moral subjects and of situating decisions within concrete contexts (Coeckelbergh 2020; Peterson and Broersen 2024). The key concepts and relationships of this discussion are presented in Figure 1.

Another constitutive element of ethics, often underestimated in technicist approaches, is normative pluralism. Contemporary societies are characterised by the coexistence of different moral traditions, conceptions of the good and criteria of justice. As Rawls (1993) argues, this pluralism is not a deviation to be overcome, but a structural condition of modern democracies. In this context, ethics cannot presuppose a single correct answer to all dilemmas. Rather, it must operate as a space of rational mediation between divergent perspectives, allowing for the possibility of reasonable disagreement. Ethical judgement is therefore often situated, contestable and open to revision, and cannot be reduced to a decision function producing universally valid answers.

Figure no. 1

AI Ethics as Human and Institutional Responsibility: Key Concepts and Relationships.



Source: Prepared by the authors.

These distinctions allow a central thesis to be stated more precisely: ethics, as a practice of responsible judgement, normative justification and mediation in pluralistic contexts, cannot be reduced either to the application of rules or to the optimisation of outcomes. It involves capacities that exceed the execution of formal procedures, including sensitivity to context, recognition of the other, deliberation under uncertainty and openness to critical revision.

This characterisation has direct implications for the debate on artificial intelligence. Many contemporary approaches tend to treat ethics as something that

can be formalised, operationalised or incorporated into technical systems, either through explicit rules or through data-driven learning models. However, if ethics entails imputable responsibility, normative justification and pluralism, a conceptual difficulty arises: *to what extent can these characteristics be attributed to artificial systems?*

The issue is not merely that current AI systems present technical limitations or risks of bias, opacity or error. The deeper problem is that the very idea of "ethical AI" presupposes that artificial systems can perform functions which, according to the definition adopted here, are intrinsically bound to the condition of a responsible moral agent. This tension does not render ethics irrelevant in the field of AI. On the contrary, it reinforces the need to distinguish between two levels that are frequently conflated: ethics as a property of a moral agent and ethics as a quality of practices, institutions and sociotechnical systems.

It is from this distinction that the argument proceeds. Rather than asking how ethics can be incorporated into the machine, it becomes necessary to examine whether, and under what conditions, artificial systems may be regarded as moral agents.

CAN AI BE A MORAL AGENT?

Given the conceptual delimitations established in the previous section, the question of moral agency in artificial intelligence systems can now be formulated more precisely: *can an artificial system satisfy the conditions that characterise a moral agent?*

Much of the literature on AI ethics has addressed this question under the rubric of machine ethics, often associated with Moor (2006) and further developed by Wallach and Allen (2009). Within this framework, a distinction is made between systems capable of operating in accordance with moral rules, or implicit ethical agents; systems endowed with explicit mechanisms of normative deliberation, or explicit ethical agents; and the stronger hypothesis of full ethical agents. This typology makes it possible to acknowledge that artificial systems may, to varying degrees, incorporate normative constraints or produce behaviours compatible with ethical standards, without this necessarily entailing the attribution of moral agency in the strict sense.

This point becomes decisive in the light of the definition of ethics adopted in this article. The capacity to produce morally acceptable outputs, whether through programmed rules or learning models, is not equivalent to the capacity to assume responsibility, justify decisions and answer to others. The distinction between behaviour that conforms to norms and full moral agency, already present in Moor's work, acquires here a more demanding scope: the issue is not merely to identify levels of functional sophistication, but to assess whether artificial systems satisfy the constitutive conditions of ethical practice.

Engaging with Qualified Artificial Moral Agency

Floridi and Sanders (2004) propose that artificial agents could exercise a form of "mindless morality" at a high level of abstraction, capable of normatively evaluating outcomes even without consciousness. Wallach and Allen (2009) further describe "functional morality," where AI systems may produce behaviours consistent with ethical norms, simulating deliberative processes through programmed rules or learning models.

While these perspectives acknowledge normative sophistication, they do not meet the requirements for *full moral agency* as defined in this article: the capacity for imputable responsibility, the ability to offer normative justification, and the situated contextualisation necessary to respond to pluralistic moral demands. AI may align with rules or generate normatively acceptable outputs, but it *cannot assume ultimate responsibility, provide justificatory reasons, or answer to those affected*.

This reinforces the article's central argument: ethical responsibility cannot be delegated to artificial systems. Human and institutional actors remain the locus of accountability and answerability, supported by governance frameworks that structure oversight, contestability, and reparation mechanisms. Integrating these insights strengthens the conceptual distinction between functional compliance and genuine moral agency, highlighting why AI ethics must be anchored in human and institutional practices.

Dimensions and Normative Implications of Moral Agency

The analysis now turns to the constitutive dimensions of moral agency, which are central to understanding how ethical responsibility is structured in AI-mediated contexts. Peterson and Hamrouni (2022) help to organise this problem by identifying empathy, accountability, contextualisation and normative pluralism as relevant dimensions of machine ethics. These dimensions provide an entry point for the analysis, while the argument also draws on broader philosophical traditions already mobilised in the article: the open-question argument, moral pluralism, discourse ethics and accounts of situated judgement. These traditions converge on the idea that ethical judgement involves justification, interpretation and contestability, rather than the application of a single decision rule.

Empathy introduces a dimension frequently marginalised in technicist approaches. Ethics is not reducible to the correct application of norms; it involves the capacity to consider the other as a moral subject, recognising their needs, vulnerabilities and expectations. This capacity cannot be equated with the simulation of affective states or the modelling of preferences. It implies a form of relation that presupposes involvement, recognition and genuine consideration. For example, a system may generate a supportive response to a student in distress or to a patient describing suffering, but it does not recognise that person as someone to

whom it is morally answerable. The responsibility for interpreting the situation and deciding how to respond therefore remains with human professionals and institutions. Even when AI systems simulate empathic responses, they do not participate in the moral relation of recognition, care and obligation that gives empathy its normative density.

Responsibility and accountability are therefore related but not interchangeable. Responsibility concerns the attribution of action and its consequences to an agent, whereas accountability concerns the institutional practices through which such responsibility can be examined, challenged and acted upon. To be a moral agent is to be capable of being held responsible for one's actions; that is, to be the possible object of practices such as criticism, blame, justification and, where appropriate, sanction. This dimension presupposes that the action is recognised as belonging to the agent and that it makes sense to address normative demands to that agent. To attribute blame, sanction or a demand for justification to an artificial system does not have the same normative status as attributing them to an agent capable of recognising itself as the author of an action and of answering for its effects.

Contextualisation is equally decisive. Ethical judgement does not occur in a vacuum, but in concrete situations marked by contingency, singularity and conflict between values. The same norm may lead to different decisions depending on the circumstances, requiring a capacity for situated adaptation. Although AI systems may incorporate contextual variables and adjust their decisions based on data, this adaptation remains inscribed within previously defined formal structures. They lack the capacity to reinterpret norms, question frameworks and redefine criteria in the light of unprecedented situations, which is characteristic of human ethical deliberation.

Normative pluralism is central to any demanding account of moral agency. As discussed in the previous section, ethics operates in contexts in which different values may come into tension, without there being a single solution that is both correct and predetermined. Ethical deliberation therefore involves processes of arbitration, negotiation and compromise between reasonable alternatives. Any attempt to formalise this dynamic within a decision function, for example, a function $f(x)$ capable of producing "the" correct answer, faces the difficulty of incorporating legitimate disagreements and competing criteria (Williams 1985; Nussbaum 2001; Peterson and Hamrouni 2022; Peterson 2023). Even when AI systems are trained to optimise multiple objectives or balance different metrics, they do so on the basis of previously defined parameterisations, which do not remove the need for normative justification of those choices.

The core of the argument can therefore be stated directly. The difficulty of attributing moral agency to AI does not lie only in contingent technical limitations, but in a structural incompatibility between the demands of ethics, as defined here, and the mode of operation of artificial systems. Empathy is not reducible to

simulation; responsibility is not reducible to traceability; contextualisation is not reducible to parametric adaptation; pluralism is not reducible to multi-objective optimisation. In all these cases, there is a difference between executing procedures that produce normatively relevant results and participating in a practice of justification and accountability before others.

This does not imply that artificial systems are ethically irrelevant or that they cannot play significant roles in decision-making processes. As suggested by Moor (2006) and further developed by Wallach and Allen (2009), artificial systems may function as normatively sensitive instruments, capable of incorporating constraints, supporting deliberation and reducing certain forms of error or bias. The problem lies not in their use, but in the tendency to attribute to them a status that obscures human responsibility (Bryson 2018; Peterson 2023).

The question "Can AI be a moral agent?" therefore admits a differentiated answer. If moral agency is understood merely as the capacity to produce behaviours compatible with norms, then artificial systems may perform this role to varying degrees. If, however, moral agency is defined, as it is in this article, in terms of imputable responsibility, normative justification, situated contextualisation and participation in a plural space of deliberation, then the answer is negative. This limitation in moral agency contributes to the responsibility gaps examined in the section 'Responsibility Gap and Non-Delegable Responsibility'.

This distinction has direct implications for how AI ethics itself is conceived (Ladrière 1997; Zuin 2021; Santaella 2025). If artificial systems cannot be regarded as moral agents in the full sense, then ethics cannot be located within the machine. It must instead be sought in the human, institutional and sociotechnical practices that orient its development, implementation and use.

THE LIMITS OF COMPLETE ETHICAL FORMALISATION

The analysis developed thus far makes it possible to shift the problem of moral agency to a more fundamental level. If ethics involves imputable responsibility, normative justification, situated judgement and mediation within pluralistic contexts, then the question is no longer only whether artificial systems can satisfy these requirements. It becomes necessary to examine a deeper presupposition underpinning many contemporary proposals: can ethics be completely and self-sufficiently formalised into a decision function capable, by itself, of producing morally justified answers?

Much of the literature on machine ethics has operated with this hypothesis, even if often implicitly. The idea that moral dilemmas could be translated into computational terms, whether through explicit rules or probabilistic models, assumes that ethical deliberation can be represented as a function $f(x)$, in which inputs such as data, contexts and preferences are processed according to specific criteria in order to produce an output deemed morally acceptable. This ambition is

present across different traditions, from approaches based on normative logic to models of decision-making under uncertainty.

However, this hypothesis faces a structural difficulty. Peterson and Hamrouni (2022) formulate it in terms of the limits of defining a moral decision function, while the open-question argument (Moore 1959) shows why any proposed definition of what is morally right remains open to further evaluation. Ethics cannot be reduced to a function capable of systematically producing "the" correct answer for each situation, because the criteria guiding that function are themselves objects of normative assessment.

Applied to artificial intelligence, this argument means that no decision function can bring ethical justification to a close. Even if a system produces decisions that are consistent, transparent or aligned with certain principles, it remains possible, and, in pluralistic contexts, inevitable, to question the normative presuppositions that guide those decisions. Formalisation may produce results; it cannot, by itself, justify their legitimacy.

The difficulty becomes even more apparent when considering optimisation-based models, such as those operating through the maximisation of expected utility or the minimisation of risk. Although such models may be useful for structuring decisions in specific contexts, they depend on prior normative choices: which values should be considered, how they should be weighted, which risks are acceptable and which consequences are relevant. Apparently technical parameters, such as decision thresholds or cost functions, incorporate substantive decisions that cannot be derived from the model itself. Calculation does not eliminate the need for judgement; it merely relocates it to the point at which the criteria are defined.

At this point, it is essential to return to a distinction already suggested in the previous sections: the difference between technical explanation and normative justification. An AI system may explain how it arrived at a given decision, whether through traceability or techniques of interpretability. It may indicate which variables were considered, which weights were assigned and which rules were applied. Yet this answers the question "*how was the decision produced?*", not the question "why is this decision morally acceptable?". As Peterson and Hamrouni (2022) and Peterson and Broersen (2024) discuss, propositional explanation, understood as the description of a system's internal processes, is not equivalent to normative justification, which involves offering reasons capable of sustaining the legitimacy of a decision before others.

This distinction also reveals a central epistemological limit. The formalisation of ethics presupposes that the moral validity of a decision can be inferred from a set of rules or procedures. Ethical justification, however, is not exhausted by the application of such procedures. It involves a space of argumentation in which reasons are presented, assessed and potentially contested. This space is constitutively open, since it admits revision, disagreement and

reinterpretation. No formal system can, by itself, close this process, because the criteria of evaluation remain open to criticism.

The reference to the limits of formal systems, such as those identified by Gödel (1962), is retained only as a limited analogy. It does not establish a direct equivalence between mathematical incompleteness and ethical deliberation. Its role is simply to underline the difficulty of treating ethics as a closed and self-sufficient system of justification. In the case of ethics, this difficulty appears in the impossibility of reducing moral judgement to a closed set of rules or to an exhaustive decision function.

This argument is not directed only at deterministic models. Probabilistic and machine-learning-based systems, which operate through uncertainty, adaptation and continuous updating, also remain dependent on formal structures, including training data, loss functions and optimisation objectives. These structures do not remove the need for normative justification. The introduction of indeterminacy or stochasticity does not solve the problem; it merely displaces it. The relevant question is not whether the system is deterministic or probabilistic, but whether it can justify the normative criteria that guide its decisions.

The limits of complete ethical formalisation therefore cease to be merely a conceptual hypothesis and acquire a more precise epistemological status. The point is not simply that current AI systems remain technically insufficient, but that the idea of a complete and self-sufficient ethical function faces structural obstacles. Ethics, understood as a practice of justification, responsibility and deliberation in pluralistic contexts, resists full reduction to formal procedures.

This does not diminish the relevance of formal tools in supporting moral decision-making. Computational models may help to identify patterns, simulate scenarios, make trade-offs explicit and reduce certain forms of error. Their role, however, should be understood as instrumental and auxiliary, rather than as a substitute for ethical judgement. Formalisation may organise deliberation, but it cannot settle it.

The argument developed in this section makes the notion of an "ethical machine" in the full sense conceptually untenable. If ethics cannot be fully formalised, and if moral agency presupposes capacities for justification and responsibility that exceed the formal domain, then machine ethics cannot be understood as the construction of autonomous moral agents. Its proper scope lies elsewhere: in the design of conditions under which artificial systems may support, but not replace, human ethical responsibility.

FROM THE ETHICAL MACHINE TO OPERATIONAL AI GOVERNANCE

The conceptual and epistemological limits established in the previous sections require a shift from the supposed morality of the machine to the responsibility of the sociotechnical system (Mittelstadt et al. 2016; Larsson 2019).

If ethics cannot be fully formalised, and if moral agency presupposes capacities for responsibility, justification and deliberation that cannot be attributed to artificial systems, then the idea of an "ethical machine" in the full sense becomes conceptually unsustainable. The central question is no longer how to endow artificial intelligence with a morality of its own, but how to organise human practices and institutions capable of integrating artificial systems without dissolving ethical responsibility.

This shift does not amount to a renunciation of AI ethics, but to its reformulation. The impossibility of the ethical machine does not eliminate ethics; it makes ethics more demanding by returning it to the domain of human, institutional and sociotechnical practices. In this context, the distinction between ethics as a property of moral agents and ethics as a quality of sociotechnical systems becomes decisive. Artificial intelligence is not ethical as an agent, but it may be embedded in configurations that are more or less ethical, depending on the structures of responsibility within which it is framed.

This distinction avoids two recurrent misconceptions: the anthropomorphisation of AI, which attributes to it a non-existent moral interiority, and the displacement of human responsibility, which turns technical mediation into an alibi for the absence of imputability. A pragmatic human ethics rejects both, articulating mediation, delegation and responsibility in contexts shaped by technical systems.

AI ethics may therefore be formulated as a set of interdependent normative requirements (UNESCO 2021; European Commission 2026; Papagiannidis *et al.* 2025). These include non-delegable responsibility, according to which decisions with morally significant impact must remain attributable to identifiable human or institutional agents, and substantive oversight, meaning that human intervention must be effective, informed and endowed with a real capacity for revision or interruption. They also include justificatory transparency, since it is not sufficient to explain technical functioning without offering normatively relevant reasons for decisions; contestability, whereby AI-mediated decisions must be open to questioning, review and appeal; and reparation, requiring systems to include mechanisms for responding to harm, including correction and compensation. Finally, continuous institutional governance is required, since AI ethics depends on monitoring, auditing and revision throughout the entire life cycle of systems. Before detailing the institutional implementation of governance principles, we discuss how automation may create responsibility gaps. We then show how non-delegable responsibility addresses these gaps.

From a responsibility-centred, discourse-ethical and pluralist perspective, these requirements help to prevent automation from producing zones of opacity and irresponsibility. This framing places justification, imputability and answerability at the centre of ethical agency; it treats justification as addressed to others; and it recognises that ethical judgement takes place under conditions of

reasonable disagreement. In this sense, AI ethics is displaced from the problem of the supposed incorporation of values into technical systems towards the institutional configuration of responsibility.

Risk-Based Regulation and Contextual Differentiation

One of the most significant developments in AI governance is the adoption of risk-based regulatory models, in which normative requirements vary according to the potential impact of systems (European Parliament and Council of the European Union 2024). This approach recognises that not all uses of AI have the same ethical significance, allowing more stringent obligations to be concentrated on applications that affect fundamental rights, access to resources or social recognition.

In the European context, for instance, the Artificial Intelligence Act establishes differentiated categories of risk, unacceptable, high, limited and minimal, associated with obligations such as conformity assessment, technical documentation, human oversight and post-deployment monitoring. This logic may be extended to other institutional contexts if it is adapted to local and sectoral specificities.

A risk-based approach, however, is not sufficient in itself. Although it is necessary for identifying hazards and mitigating harm, it may underestimate broader ethical dimensions, such as distributive justice, social recognition and long-term impacts. Regulation should therefore be integrated into a wider governance ecosystem in which legal norms, institutional practices and deliberative processes are articulated (Larsson 2019; Papagiannidis et al. 2025).

Accountability Infrastructures: Auditing, Documentation and Impact Assessment

The operationalisation of AI ethics requires accountability infrastructures that make decisions mediated by artificial systems traceable, assessable and contestable (Korobenko *et al.* 2024; Papagiannidis *et al.* 2025). Among the central instruments are algorithmic impact assessments, independent and continuous audits, systematic documentation and post-deployment monitoring.

Algorithmic impact assessments make it possible to identify, prior to implementation, ethical, social and legal risks, as well as mitigation strategies (European Parliament and Council of the European Union 2024; European Commission 2026). Independent audits verify the functioning of systems in real-world settings, including bias, performance and normative compliance (Batool *et al.* 2025; Papagiannidis *et al.* 2025). Systematic documentation records design decisions, data characteristics, system limitations and conditions of use. Post-deployment monitoring, in turn, follows effects over time and identifies unintended consequences.

These instruments do not replace ethical judgement, but they create the conditions under which it can be exercised in an informed and responsible manner. Without them, AI ethics remains at the declarative level, lacking effective capacity for intervention (Mittelstadt *et al.* 2016; Jobin *et al.* 2019).

Institutional Structures and Sectoral Implementation

Beyond technical instruments, AI ethics requires institutional structures capable of sustaining continuous processes of deliberation, accountability and revision (UNESCO 2021; Papagiannidis *et al.* 2025). In this context, AI ethics committees or review boards are particularly relevant. Such bodies should bring together specialists from different fields, including technology, law, philosophy and the social sciences, and, ideally, representatives of groups potentially affected by AI systems.

These committees should not function as merely consultative or symbolic bodies, but as spaces with real capacity to influence organisational decisions. Their functions may include project assessment, the definition of internal guidelines, the monitoring of systems in operation and the mediation of ethical conflicts. At the same time, excessive centralisation of governance in a single entity must be avoided. Since AI cuts across diverse domains, healthcare, education, justice, security and the world of work, its governance should be distributed and multisectoral, involving regulatory bodies, public institutions, private organisations and civil society (UNESCO 2021; Batool *et al.* 2025).

This requirement becomes particularly evident in sensitive contexts. In healthcare, clinical decision-support systems may influence diagnoses, treatments and the allocation of resources, requiring professional decision-making authority, clinical audits and transparency regarding system limitations. In justice, risk assessment or decision-support algorithms require intelligible explanations, contestability and effective mechanisms of review. In education, personalisation systems and automated assessment require continuous monitoring, the involvement of educators and attention to the reproduction of inequalities (European Parliament and Council of the European Union 2024).

AI ethics is therefore not the property of a single actor, nor can it be addressed abstractly and homogeneously. It constitutes a collective process of accountability, contextualised according to risks, affected subjects and the specific impacts of each domain of application.

Higher Education and Research as Governance Contexts

These requirements have particular relevance for higher education and research, which are not merely additional fields of application but institutional contexts in which knowledge, assessment, recognition and professional futures are produced. In universities, AI systems may support automated or semi-automated

assessment, plagiarism and authorship detection, learning analytics, admission and progression decisions, research writing, peer review and data analysis. Such uses do not transfer responsibility to the system. They redistribute practical tasks across students, teachers, researchers, reviewers, managers and institutional bodies, making it necessary to determine who authorises the use of the tool, who interprets its outputs, who can contest them and who answers for possible harm.

An automated assessment system, for example, may classify performance or flag suspected misconduct, but it cannot justify the academic meaning of that classification. The responsible decision must remain with identifiable educators or institutional committees, who must be able to examine the evidence, consider contextual information, hear the affected student and revise the decision. Similarly, when generative AI is used in research design, literature synthesis, data analysis or drafting, ethical governance cannot be reduced to the declaration that a tool was used. It requires prior rules on acceptable use, documentation of human contribution, protection of data, verification of outputs and accountability for errors, omissions or fabricated content.

Within higher education, responsibility-centred AI governance should therefore combine institutional guidelines, domain-specific review procedures, training for staff and students, disclosure requirements, appeal mechanisms and periodic audits of AI-mediated decisions. These measures express the article's central claim at the institutional level: AI may support pedagogical or research deliberation, but the justification of decisions affecting learning, authorship, evaluation or academic recognition must remain human and institutionally answerable.

Responsibility Gap and Non-Delegable Responsibility

The deployment of increasingly autonomous AI systems can give rise to what the literature describes as a responsibility gap: situations in which it becomes unclear how moral responsibility, accountability, and answerability should be attributed when automation obscures causal chains and decision processes. Matthias (2004) introduced the term to characterise instances in which human agents cannot be held fully accountable for the harms or normative consequences of automated systems, especially when system outcomes cannot be traced to specific human acts or intentions.

Subsequent work has refined this notion by identifying different types of responsibility gaps and by emphasising the importance of meaningful human control. Sparrow (2007) highlights how delegation to machines can displace normative responsibility, creating zones of opacity and diffusion where no single agent is clearly answerable. Santoni de Sio and Mecacci (2021) further develop the responsibility-gap literature by distinguishing four interconnected forms of responsibility gap: gaps in culpability, moral accountability, public accountability and active responsibility. They show that these gaps may arise from technical,

organisational, legal, ethical and societal sources, and propose meaningful human control as part of a more comprehensive approach to addressing them. Their analysis is particularly relevant here because it shows that responsibility cannot be secured simply by assigning a nominal human supervisor; the sociotechnical system must preserve the conditions under which human agents can meaningfully control, assess and respond to AI-mediated actions. These authors converge on the insight that without human and institutional anchors of responsibility, normative accountability risks being lost in the technical operations of automated systems.

In the context of the governance model proposed in this article, acknowledging responsibility gaps underscores why responsibility is not delegable to AI systems. Our framework of non-delegable responsibility situates responsibility within human and institutional arrangements that design, authorise, supervise, contest, and, where necessary, repair AI-mediated decisions. By explicitly incorporating mechanisms for oversight, contestability, and reparation, and by clarifying who authorises, interprets, and answers for AI outputs, the framework responds to the normative concerns foregrounded by Matthias, Sparrow, and Santoni de Sio and Mecacci. This approach ensures that accountability and answerability remain traceable to identifiable human actors - educators, researchers, review committees, or governance bodies - thereby avoiding the responsibility gaps that would otherwise arise in highly automated decision contexts.

When we refer to reparation, we understand it as a decisive element to ensure that ethical responsibility is not limited merely to the prevention of harm. Complex systems inevitably produce errors, side effects and unintended consequences. AI governance must therefore include clear mechanisms for compensation, decision correction and institutional accountability (UNESCO 2021; European Parliament and Council of the European Union 2024). Trust in the use of AI depends less on the promise of technical perfection than on the institutional capacity to acknowledge failures, correct errors and provide accountability. In this sense, trust is not a technical attribute of systems, but the result of institutional practices committed to responsibility, transparency and justice.

It is important to emphasise the need to address the risk that artificial intelligence ethics may be reduced to corporate or normative rhetoric without concrete effects. The mere enunciation of principles such as transparency, fairness and responsibility does not guarantee their practical realisation. What is at stake is the capacity to translate them into operational structures, institutional processes and verifiable practices (Batool *et al.* 2025).

From this perspective, AI governance does not consist merely in regulating technologies. It involves reconfiguring relations of responsibility in societies increasingly mediated by technical systems (European Parliament and Council of the European Union 2024). Ethics cannot be externalised to the machine, nor

reduced to formal compliance; it must be embedded in continuous processes of decision-making, assessment, revision and accountability.

The broader consequence of this argument is that the ethics of artificial intelligence must be conceived as an integral part of public policy and institutional design. The issue is not merely to regulate the use of AI, but to define who governs, how governance is exercised, through which instruments and under which forms of accountability.

CONCLUSION

This article examined the conceptual legitimacy of the expression "ethical artificial intelligence" by questioning whether artificial systems can be regarded as moral agents in a robust normative sense. The analysis showed that this question cannot be addressed only through technological improvement, regulatory compliance or the alignment of systems with predefined values. Before discussing how AI can be used responsibly, it is necessary to clarify what is meant by ethics, moral agency, responsibility and justification.

In line with the conceptual analysis developed throughout the article, we may identify that ethics is not reducible to the application of rules, the optimisation of outcomes or the production of technically acceptable decisions. Ethics involves situated judgement, imputable responsibility, normative justification and mediation within pluralistic contexts. These dimensions require an agent capable of recognising others as moral subjects, offering reasons for decisions and responding to the consequences of actions. From this perspective, artificial intelligence systems may generate normatively relevant outputs, simulate morally acceptable decision patterns and support human deliberation. However, they cannot assume moral responsibility, justify their decisions normatively or answer to those affected by their operation.

The article therefore argues that the main difficulty does not lie only in the current technical limitations of AI systems. Rather, it concerns the conceptual and normative limits of attributing moral agency to artificial systems. AI can operate according to programmed rules, statistical models and learning processes, but this does not mean that it participates in ethical judgement as a responsible agent. The distinction between behaviour compatible with norms and moral agency in the full sense is therefore central. A system may produce a decision that appears ethically acceptable without becoming an ethical subject. In this sense, the article contributes to the debate by distinguishing between the ethical status of the agent and the ethical quality of the sociotechnical arrangement.

Another important contribution of the article concerns the limits of formalising ethics. The analysis showed that ethical judgement cannot be fully translated into a decision function capable of producing morally correct answers in all situations. Technical explanation may describe how a system reached a given

output, which variables were considered or which rules were applied. Nevertheless, this does not amount to normative justification. Explaining how a decision was produced is different from justifying why that decision should be considered morally acceptable. This distinction is particularly relevant in AI-mediated contexts, where opacity, automation and institutional distance may weaken the attribution of responsibility.

Taking into account these conceptual and epistemological limits, the article proposes a shift from the idea of the ethical machine to the governance of responsible sociotechnical systems. The central issue is not how to place morality inside artificial systems, but how to organise human, institutional and regulatory arrangements capable of ensuring responsibility when AI is used in socially consequential domains. This means that AI ethics should be understood as a practice of governance based on non-delegable responsibility, substantive oversight, justificatory transparency, contestability, reparation and continuous monitoring.

This shift has direct implications for public policy, institutional design and organisational practice. If AI systems cannot be responsible moral agents, then responsibility must remain attributable to identifiable human or institutional actors. Decisions mediated by AI should therefore be subject to clear procedures of explanation, review, appeal and correction. In areas such as healthcare, justice and education, where automated systems may affect access to care, rights, opportunities and recognition, these requirements become particularly important. Risk-based regulation, algorithmic impact assessments, independent audits, systematic documentation and post-deployment monitoring can support ethical governance, but they cannot replace human judgement and institutional accountability.

The article's specific contribution is therefore to make explicit the conceptual link between the limits of artificial moral agency and the design of responsibility-bearing institutions. In higher education and research, this link is especially consequential: automated assessment, research-writing tools, integrity systems and AI-assisted review practices may support academic work, but they cannot bear academic responsibility. The justificatory burden remains with teachers, researchers, reviewers and universities that authorise, supervise, contest and correct those uses.

The article also highlights the importance of moving beyond a merely declarative understanding of AI ethics. The repeated reference to principles such as transparency, fairness and responsibility does not, by itself, guarantee ethical practice. These principles must be translated into operational mechanisms, institutional procedures and verifiable forms of accountability. Without such translation, AI ethics risks becoming a normative rhetoric that legitimises technological adoption without effectively protecting affected subjects. Therefore, the ethical challenge of AI is not only to define principles, but to create

institutional conditions in which those principles can be implemented, contested and revised.

This study has some limitations. It is a conceptual article and does not include empirical analysis of concrete AI governance mechanisms. Its argument is also mainly grounded in Western philosophical traditions, particularly debates on moral agency, responsibility, normativity and justification. Future studies may expand this analysis by incorporating intercultural perspectives on ethics and technology, examining how accountability mechanisms operate in different institutional sectors and comparing regulatory regimes across national and international contexts. Further empirical research is also needed to understand how affected subjects experience AI-mediated decisions, how organisations distribute responsibility and how mechanisms of contestability and reparation function in practice.

Future developments in increasingly autonomous AI systems are likely to intensify rather than remove these questions. As systems become more adaptive, opaque and embedded in institutional routines, responsibility will not always be reducible to a single operator at the moment of use. It will also depend on prior decisions about authorisation, procurement, design, training, supervision, appeal and reparation. Responsibility frameworks may therefore need to evolve, but this evolution should not mean transferring moral responsibility to artificial systems. It should clarify how responsibility is distributed, documented and exercised across the human and institutional arrangements that make AI-mediated decisions possible.

The article therefore concludes that artificial intelligence cannot be ethical in the full sense if ethics is understood as responsible moral agency. This conclusion does not weaken the relevance of AI ethics; rather, it clarifies its proper locus. In societies increasingly mediated by artificial systems, the central challenge is to ensure that automation does not obscure, dilute or displace human and institutional responsibility. Ethics, in this context, lies not in the moralisation of machines, but in the conditions through which their design, deployment and use remain accountable, contestable and open to correction.

AI Use Statement

The conception, execution, and analysis of the research titled “The Limits of Ethical Artificial Intelligence: Moral Agency, Non-Delegable Responsibility and Sociotechnical Governance” are of the exclusive authorship and intellectual responsibility of the human authors. Generative artificial intelligence tools (ChatGPT 5.2 and Gemini 3, accessed in April and May 2026) were used in a limited and precise manner solely for technical, operational, and linguistic support, specifically for: grammatical and stylistic refinement, assistance in translating terms, formal verification of consistency between in-text citations and the reference list, formatting the abstract, adapting the references to the "Chicago author-date"

style, and the schematic structuring of the manuscript based on the original research report. The AI tools were not used to generate data, conduct analyses, interpret results, or produce scientific inferences. All content was subjected to rigorous critical review and validation by the authors, who assume full and exclusive responsibility for the originality, integrity, and substance of the final manuscript, as well as for the ethical, scientific, and legal obligations of all analyses, interpretations, and conclusions presented in this work.

REFERENCES

Batool, Amna, Didar Zowghi, and Muneera Bano. 2025. "AI Governance: A Systematic Literature Review." *AI and Ethics* 5 (3): 3265–79. https://doi.org/10.1007/s43681-024-00653-w.

Bostrom, Nick, and Eliezer Yudkowsky. 2014. "The Ethics of Artificial Intelligence." In *The Cambridge Handbook of Artificial Intelligence*, edited by Keith Frankish and William M. Ramsey. Cambridge University Press. https://doi.org/10.1017/CBO9781139046855.020.

Bryson, Joanna J. 2018. "Patiency Is Not a Virtue: The Design of Intelligent Systems and Systems of Ethics." *Ethics and Information Technology* 20 (1): 15–26. https://doi.org/10.1007/s10676-018-9448-6.

Coeckelbergh, Mark. 2020. *AI Ethics*. MIT Press.

Corrêa, Nicholas Kluge, Camila Galvão, James William Santos, et al. 2023. "Worldwide AI Ethics: A Review of 200 Guidelines and Recommendations for AI Governance." *Patterns* 4 (10): 100857. https://doi.org/10.1016/j.patter.2023.100857.

European Commission. 2026. *Living Guidelines on the Responsible Use of Generative AI in Research*. 3rd ed. Directorate-General for Research and Innovation. https://research-and-innovation.ec.europa.eu/document/download/2b6cf7e5-36ac-41cb-aab5-0d32050143dc_en?filename=ec_rtd_ai-guidelines.pdf.

European Parliament and Council of the European Union. 2024. "Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No. 300/2008, (EU) No. 167/2013, (EU) No. 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)." *Official Journal of the European Union* L 2024/1689. http://data.europa.eu/eli/reg/2024/1689/oj.

Farisco, Michele, Kathinka Evers, and Arleen Salles. 2020. "Towards Establishing Criteria for the Ethical Analysis of Artificial Intelligence." *Science and Engineering Ethics* 26 (5): 2413–25. https://doi.org/10.1007/s11948-020-00238-w.

Floridi, Luciano. 2013. *The Ethics of Information*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199641321.001.0001.

Floridi, Luciano, Josh Cowls, Monica Beltrametti, et al. 2018. "AI4People–An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations." *Minds and Machines* 28 (4): 689–707. https://doi.org/10.1007/s11023-018-9482-5.

Floridi, Luciano, and Jeff W. Sanders. 2004. "On the Morality of Artificial Agents." *Minds and Machines* 14 (3): 349–79. https://doi.org/10.1023/B:MIND.0000035461.63578.9d.

Gödel, Kurt. 1962. *On Formally Undecidable Propositions of Principia Mathematica and Related Systems*. Translated by Bernard Meltzer. Oliver & Boyd. Originally published 1931.

Habermas, Jürgen. 1990. *Moral Consciousness and Communicative Action*. Translated by Christian Lenhardt and Shierry Weber Nicholsen. MIT Press.

Holmes, Wayne, Maya Bialik, and Charles Fadel. 2019. *Artificial Intelligence in Education: Promises and Implications for Teaching and Learning*. Center for Curriculum Redesign.

Jaakkola, Elina. 2020. “Designing Conceptual Articles: Four Approaches.” *AMS Review* 10 (1–2): 18–26. https://doi.org/10.1007/s13162-020-00161-0.

Jobin, Anna, Marcello Ienca, and Effy Vayena. 2019. “The Global Landscape of AI Ethics Guidelines.” *Nature Machine Intelligence* 1 (9): 389–99. https://doi.org/10.1038/s42256-019-0088-2.

Jonas, Hans. 1984. *The Imperative of Responsibility: In Search of an Ethics for the Technological Age*. Translated by Hans Jonas with the collaboration of David Herr. University of Chicago Press. Originally published 1979.

Khan, Arif Ali, Sher Badshah, Peng Liang, et al. 2022. “Ethics of AI: A Systematic Literature Review of Principles and Challenges.” In *Proceedings of the 26th International Conference on Evaluation and Assessment in Software Engineering*, 383-92. Association for Computing Machinery. https://doi.org/10.1145/3530019.3531329.

Korobenko, Daria, Anastasija Nikiforova, and Rajesh Sharma. 2024. “Towards a Privacy and Security-Aware Framework for Ethical AI: Guiding the Development and Assessment of AI Systems.” In *dg.o ’24: Proceedings of the 25th Annual International Conference on Digital Government Research*, 740–53. Association for Computing Machinery. https://doi.org/10.1145/3657054.3657141.

Ladrière, Jean. 1997. *L’éthique dans l’univers de la rationalité*. Artel; Fides.

Larsson, Stefan. 2019. “The Socio-Legal Relevance of Artificial Intelligence.” *Droit et société*, no. 103: 573–93. https://doi.org/10.3917/drs1.103.0573.

Matthias, Andreas. 2004. “The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata.” *Ethics and Information Technology* 6 (3): 175–83. https://doi.org/10.1007/s10676-004-3422-1.

Ménissier, Thierry. 2023. “Les quatre éthiques de l’intelligence artificielle.” *Revue d’anthropologie des connaissances* 17 (2). https://doi.org/10.4000/rac.29961.

Mittelstadt, Brent. 2019. “Principles Alone Cannot Guarantee Ethical AI.” *Nature Machine Intelligence* 1 (11): 501–7. https://doi.org/10.1038/s42256-019-0114-4.

Mittelstadt, Brent Daniel, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. “The Ethics of Algorithms: Mapping the Debate.” *Big Data & Society* 3 (2): 2053951716679679. https://doi.org/10.1177/2053951716679679.

Moor, James H. 2006. “The Nature, Importance, and Difficulty of Machine Ethics.” *IEEE Intelligent Systems* 21 (4): 18–21. https://doi.org/10.1109/MIS.2006.80.

Moore, George Edward. 1959. *Principia Ethica*. Cambridge University Press. Originally published 1903.

Nussbaum, Martha C. 2001. *Upheavals of Thought: The Intelligence of Emotions*. Cambridge University Press.

Papagiannidis, Emmanouil, Patrick Mikalef, and Kieran Conboy. 2025. “Responsible Artificial Intelligence Governance: A Review and Research Framework.” *The Journal of Strategic Information Systems* 34 (2): 101885. https://doi.org/10.1016/j.jsis.2024.101885.

Peterson, Clayton. 2023. “Further Thoughts on Defining f(x) for Ethical Machines: Ethics, Rational Choice, and Risk Analysis.” *The International FLAIRS Conference Proceedings* 36. https://doi.org/10.32473/flairs.36.133203.

Peterson, Clayton, and Jan Broersen. 2024. “Understanding the Limits of Explainable Ethical AI.” *International Journal on Artificial Intelligence Tools* 33 (3): 2460001. https://doi.org/10.1142/S0218213024600017.

Peterson, Clayton, and Naïma Hamrouni. 2022. “Preliminary Thoughts on Defining f(x) for Ethical Machines.” *The International FLAIRS Conference Proceedings* 35. https://doi.org/10.32473/flairs.v35i.130545.

Rawls, John. 1993. *Political Liberalism*. Columbia University Press.

Ricoeur, Paul. 1990. *Soi-même comme un autre*. Seuil.

Santaella, Lucia. 2025. “A ética como guia para o uso da inteligência artificial generativa.” *Educação & Sociedade* 46: e296469. https://doi.org/10.1590/es.296469.

Santoni de Sio, Filippo, and Giulio Mecacci. 2021. “Four Responsibility Gaps with Artificial Intelligence: Why They Matter and How to Address Them.” *Philosophy & Technology* 34 (4): 1057-84. https://doi.org/10.1007/s13347-021-00450-x.

Sparrow, Robert. 2007. “Killer Robots.” *Journal of Applied Philosophy* 24 (1): 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x.

UNESCO. 2021. *Recommendation on the Ethics of Artificial Intelligence*. Adopted November 23. UNESCO. https://www.unesco.org/en/legal-affairs/recommendation-ethics-artificial-intelligence.

Wallach, Wendell, and Colin Allen. 2009. *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195374049.001.0001.

Williams, Bernard. 1985. *Ethics and the Limits of Philosophy*. Harvard University Press.

Zuin, Antônio Álvaro Soares. 2021. “Inteligência artificial e formação danificada: aprendizagem profunda e ética rasa entre professores e alunos.” *Educar em Revista* 37: e80158. https://doi.org/10.1590/0104-4060.80158.

Acest articol analizează critic legitimitatea conceptuală a expresiei "inteligență artificială etică", întrebând dacă sistemele artificiale pot fi considerate agenți morali într-un sens normativ robust. Adoptând un design conceptual și teoretic argumentativ și bazându-se pe o revizuire narativă critică a literaturii despre etica AI, agent moral, formalizarea algoritmică și guvernanța sociotehnică, articolul susține că sistemele AI pot genera rezultate relevante normativ, pot simula decizii moral acceptabile și pot susține deliberarea umană, dar nu pot să-și asume responsabilitatea morală, să ofere justificare normativă pentru deciziile lor sau să răspundă celor afectați de funcționarea lor. Analiza distinge între etică ca proprietate a agenților morali și etică ca o calitate a sistemelor sociotehnice proiectate, utilizate și guvernate de actori umani. Dintr-o perspectivă pluralistă, centrată pe responsabilitate și etică a discursului, articolul susține că imposibilitatea unor mașini să fie pe deplin etice nu invalidează etica AI. Mai degrabă, își redefinește domeniul de aplicare prin relocarea responsabilității etice de la sistemele artificiale către aranjamentele umane și instituționale care proiectează, autorizează, supraveghează și contestă utilizarea acestora. Lucrarea concluzionează că provocarea centrală nu este construirea unor mașini moral responsabile, ci evitarea consecințelor ca automatizarea să producă zone de opacitate, deplasare normativă și iresponsabilitate umană în luarea deciziilor cu efecte sociale.

Cuvinte-cheie: *Inteligența Artificială; Etica AI; Agenția morală; Responsabilitatea umană; Guvernanța algoritmică.*

Received: 15.05.2026 *Accepted: 14.09.2026*