SECURITY & DEFENCE QUARTERLY
https://securityanddefence.pl/

Artificial intelligence as a moral agent: regulatory implications and a relational–contextual extension of Moor's classification

Maciej Czyszczoń

maciej.czyszczon@doktorant.up.krakow.pl

https://orcid.org/0000-0001-9245-5894

Doctoral School, University of the National Education Commission Podchorążych 2, 31–464, Kraków, Poland

Abstract

This paper reassesses the regulatory value of James Moor's four-level typology of machine morality, considering the European Artificial Intelligence Act (AI Act) and the forthcoming European Union (EU) liability directives. It asks whether Moors categories—ethical impact, implicit, explicit, and full moral agents—still capture the morally relevant properties of today's generative, adaptive AI, and, if not, whether adding a relational–contextual dimension can better anticipate responsibility gaps. To address this gap, we introduce a novel relational–contextual dimension and a three-factor Responsibility Index (RI_3) that refines Moor's typology by cross-classifying AI systems according to complexity, autonomy, and behavioural predictability for regulatory use. Adopting a strictly conceptual design, the study combines analytic philosophy with illustrative comparisons drawn from recent EU policy debates and high-profile incidents. It refines key terms, tests their coherence against statutory risk tiers, and distils the analysis into a three-factor matrix—complexity, autonomy, and predictability—that can be operationalised by lawmakers. The evaluation confirms that Moor's typology remains a valuable baseline for distinguishing between passive and decision-making artefacts. Nevertheless, it also highlights how moral accountability is distributed within socio-technical networks. The proposed relational–contextual dimension, in conjunction with the regulatory matrix, aligns more closely with the AI Act's risk logic and highlights scenarios in which moral agency is effectively delegated to the system. Moor's framework should be retained but augmented: only by integrating relational criteria can legislators close emerging accountability gaps surrounding large-scale, autonomous AI. The matrix offers a pragmatic tool for aligning philosophical insight with concrete legal duties.

Keywords:

artificial morality, AI regulation, Moor's typology, relational agency, security and defence

Article info
Received: 30 May 2025
Revised: 30 September 2025
Accepted: 3 November 2025
Available online: 31 December 2025

Citation: Czyszczoń, M. (2025) 'Artificial intelligence as a moral agent: Regulatory implications and a relational–contextual extension of Moor's classification', *Security and Defence Quarterly*, 52(4), pp. 116–126. doi: 10.35467/sdq/213917.

Introduction

The paper begins with the classical concept of moral agency, which holds that only rational, autonomous, and intentional beings can bear ethical responsibility for their actions. In Kantian ethics, a moral agent is an individual capable of legislating for themselves, guided by the categorical imperative. Similarly, contemporary theorists such as Floridi and Sanders (2004) de-emphasise the relevance of consciousness and free will for the attribution of responsibility, arguing instead that moral agency can be ascribed to information artefacts even in the absence of these classical mental properties (Kant, 2002/1785). However, the rapid development of artificial intelligence (AI)—from autonomous vehicles to large language models—confronts us with a situation in which systems lacking human psychology shape our lives in real ways, causing harm or benefit. This raises the question of whether traditional criteria are sufficient to capture the moral significance of such artefacts. In response, the paper proposes a relational–contextual refinement of Moor's typology together with a three-factor Responsibility Index (RI_3) tailored to regulatory decision-making, with relevance to high-risk security domains (defence, border control, and critical infrastructure).

Two closely related categories are distinguished in the literature: moral agent and moral patient. An agent is "one who can do good or evil and to whom merit or blame can be attributed," while a moral patient is the object of others' ethical obligations (Moor, 2006, p. 19). Until recently, both labels were reserved for humans. Meanwhile, today's AI systems, although lacking consciousness, can make decisions with significant consequences, such as filtering credit access, directing air traffic, and diagnosing diseases. As a result, what Moor called the "policy vacuum" is becoming increasingly apparent: our legal and ethical norms are not keeping pace with the technological agency of machines (Moor, 1985, p. 266).

A systematic review of the field reveals that no classification simultaneously tracks an artefact's complexity, degree of delegated autonomy, and behavioural predictability—the very variables on which the European Union (EU) risk-based regulation now relies. Existing taxonomies—from Moor's four-level ladder of machine morality to Floridi and Sanders' functionalist model—offer valuable heuristics for what makes an artefact ethically salient. However, they remain essentially one-dimensional, treating complexity, autonomy, and behavioural predictability as loosely related attributes, rather than as interacting variables that modulate accountability. This omission has practical consequences. Under the EU Artificial Intelligence Act, the same system may transition between limited-, high-, and unacceptable-risk tiers as its architecture evolves; however, no current framework anticipates the "responsibility gaps" that emerge when such transitions occur.

This deficiency is the specific research gap that the present paper addresses. The present study, therefore, asks the following: How can a multi-factor classification close the responsibility gaps that emerge when adaptive AI surpasses the descriptive power of agent-centric scales? To answer, it introduces the RI_3 matrix, a three-factor extension of Moor's ladder that cross-references complexity, autonomy, and predictability, thereby translating philosophical distinctions into auditable compliance triggers. The following section elaborates on Moor's four-level typology that underpins the rest of this analysis.

James Moor's Typology

James Moor stresses that the condition for including an artefact in ethical discourse is not its possession of intentionality but rather its capacity to produce effects that human norms

describe as good or bad. The first level of Moor's (2006) typology is ethical impact agents: objects and systems that, regardless of their internal lack of sentience, evidently influence human and environmental well-being.

Technological development inevitably carries measurable axiological consequences. By providing precise time cues, a digital watch primarily enhances its wearer's punctuality in day-to-day interactions but does not necessarily ensure wider contractual reliability. Similarly, the Y2K bug, lacking malicious intent, posed a significant risk of economic disruption, which in value terms represented a threat to the security and equitable distribution of goods. In both cases, moral evaluation applies not to the mental states of a watch or binary code but to the foreseeable consequences of their operation within a web of human relationships.

Thus, Moor's (1985, p. 271) thesis that "every technology is in some sense moral" assumes the importance of consequences over the tool's ontological status and explains the emergence of the previously noted policy vacuum. This approach reconfigures the classical, anthropocentric map of ethics: wherever there is potential for harm or benefit, there too minimally begins the jurisdiction of normative assessment, obligating designers and users to reflect on the values their creations embody in practice.

The second category in Moor's typology is implicit ethical agents—systems in which value-sensitive limitations have been embedded at the design stage to prevent evident violations of human-protected values. As Moor notes, such artefacts do not understand morality but implement ethical assumptions through technical limitations, effectively shifting the focus of moral assessment from intention to architectural design.

A classic example is the onboard Terrain Awareness and Warning System (TAWS) or Airborne Collision Avoidance System (ACAS), which enforces safe flight paths regardless of the pilot's decision, protecting passenger lives. Similarly, an automated teller machine (ATM) that refuses to dispense more money than the account balance prevents unauthorised appropriation of funds. In both cases, an engineered principle of non-maleficence becomes integral to the machine's operational logic. The user—sometimes against their intentions—is guided towards ethically correct action.

Modern ethics-by-design practices further develop this intuition by equipping systems with multi-layered safeguards, including sensor redundancy, anomaly detection algorithms, and emergency "kill switches." The Big Red Button concept, studied by the Google DeepMind, illustrates attempts to ensure human override of harmful AI decisions. However, as Russell (2019, pp. 160–161, 196–197) warns, the tool must be resilient against circumvention by learning systems. IEEE Standards Association (2021) and EU Trustworthy AI guidelines also call for auditable safety protocols. For this reason, most current AI applications, from autonomous vehicles to content-filtering systems, are better viewed as implicit, rather than explicit, ethical agents: their moral behaviour is hard-coded as safety constraints, rather than derived from deliberative value reasoning.

A system that can reliably prevent norm-violating behaviours is no longer a passive artefact but an operative entity of delegated agency: designers transfer part of their normative authority to the machine, allowing it to act as their replacement in real time (Floridi and Sanders, 2004; Coeckelbergh, 2015). Because the machine selects and enforces behavioural restrictions without further human intervention, observers begin to treat its outputs as intentional—although limited—acts. It is precisely this shift from causal effectiveness to normative authorship that justifies speaking of moral agency, even if the system lacks consciousness or moral understanding.

Moor (2006) defines an explicit ethical agent, the third category of his typology, as a system capable of apparent moral reasoning: a machine that not only follows programmed safety boundaries but can identify value conflicts and select solutions based on ethical norms. At this level, epistemic difficulties arise. Ethical theories must be formalised into operational rules for AI to make decisions like humans. Attempts to modularise the canonical paradigms—deontology and utilitarianism—show that their principles are both heterogeneous and mutually competitive. Systems such as MoralDM (Moral Decision-Making) and two well-known prototype agents often referred to as "Jeremy" and "W.D."—developed in machine ethics research to test hybrid rule- and learning-based approaches—address this conflict by contextually switching between utility calculation and applying absolute obligations. However, in practice, they require a fully specified "moral ontology" of the world, which has yet to be successfully developed (Cervantes et al., 2019, pp. 501–532). Critics further point out that any programmed hierarchy of rules inevitably reflects the beliefs of its designers, raising the risk of hidden bias and ethnocentrism (Allen et al., 2006).

Computational issues are equally complex. Resolving ethical dilemmas requires exploring a decision space that grows exponentially with the number of stakeholders and potential outcomes. In time-sensitive scenarios, such as an autonomous vehicle collision, the algorithm must reach a result within milliseconds, even though computing a "minimal harm" scenario may be nondeterministic polynomial-time (NP)-complete (Goodall, 2014, pp. 93–102). NP-complete problems scale exponentially, meaning that no polynomial-time solution is known; therefore, real-time vehicles must rely on heuristics that may sacrifice optimality. Hybrid architectures (e.g. GenEthand LIDA) attempt to circumvent this barrier by combining top-down rules with bottom-up learning. However, they require constant monitoring to ensure that learning does not undermine deontological constraints (Cervantes et al., 2020, pp. 117–125). As a result, Moor's vision of explicit ethical agents largely remains a research programme, as it still lacks a strong ethical formalism and efficient algorithms capable of ensuring consistency, speed, and cultural relevance in machine decision-making.

The highest level in Moor's taxonomy is the full ethical agent—a being empowered with all the attributes traditionally attributed to a mentally healthy adult human: self-awareness, the ability to distinguish right from wrong, intentionality, and free will (Moor, 2006). Moor acknowledges that no existing AI system meets these strict criteria, and the concept remains, at least for now, a speculative construct serving to define the upper limit of the debate.

Critical Perspective on Moor's Typology

The debate on full ethical agents questions the necessity of consciousness and free will for full moral agency. Advocates of functionalism, like Dennett (1992), argue that replicating the right cognitive processes is sufficient: if a system can consistently generate "drafts" of events, recognise values, and update its states based on feedback, it meets the functional criteria for moral agency, despite lacking a classical "self."[1] From this perspective, Moor's insistence on free will is a remnant of anthropocentric intuition, rather than a logical necessity.

Chalmers (1996), on the other hand, challenges the opposing view: without phenomenal consciousness—the ability to experience qualitative mental states—true moral responsibility is impossible, as valuation and feeling form the irreducible core of ethical

[1] *Functionalism holds that mental states are constituted by their causal roles.*

decision-making. If *qualia* are inaccessible to machines, full moral agents will remain in the realm of fiction.[2]

An additional critical view is raised by Bryson (2010, pp. 63–74), who warns that granting machines the status of autonomous individuals poses a risk of offsetting human responsibility. In her view, granting machines the status of agents may undermine human responsibility by shifting blame from designers to the supposed autonomy of the algorithm. Bryson (2009, pp. 5–12) compares AI to a car or a computer: tools may be highly complex, but that does not make them moral persons. AI systems should remain tools, not moral partners. As a result, Moor's concept of full ethical agent primarily functions as an experimental framework today, revealing how far contemporary technologies fall short of the ideal and forcing reflection on which human traits truly constitute the minimal threshold for moral agency, and which are merely cultural baggage embedded in our expectations.

Echoing Bryson's warning against anthropomorphising machines, subsequent critiques of Moor's typology focus mainly on the charge of anthropocentrism. Floridi and Sanders (2004, p. 349) argue that the four-tiered scale overly ties moral agency to the categories of consciousness and free will. In contrast, in real socio-technical networks, responsibility is distributed and may also apply to 'mind-less' artefacts (hence, mind-less morality).[3] Their concept of "levels of abstraction" enables a contextual analysis of AI agency as a function of the system-user-institution relationship, rather than being solely based on the machine's internal properties. The dispute between Bryson and Floridi reveals a deeper tension between the ambition to expand the category of agency and the fear of prematurely "dehumanising" ethics. Coeckelbergh (2009, pp. 181–189) takes this debate even further by proposing a theory of virtual moral agency, where the key factor is the social perception of the robot, not its ontology; an artefact may thus become morally relevant even without possessing any mental states.

Critics, regardless of their ontological stance, point to three practical challenges. First, the responsibility gap: autonomous learning systems may act unpredictably, making it challenging to identify a responsible party, as shown in examples discussed by Matthias (2004, pp. 175–183) and Asaro (2007, pp. 18–24). Second, interpretability: research into the minimal level of transparency required for moral agents shows that without comprehensible explanations of decision rules, systems cannot be effectively controlled or certified (Vijayaraghavan and Badea, 2024, pp. 1–22). Third, value alignment: the more an algorithm adapts to data, the higher the risk that its criteria for good and evil will diverge from social norms, necessitating dynamic mechanisms of oversight and rule updates.

As a result, even if Moor's typology provides a practical unifying framework, the contemporary debate has shifted emphasis from ontology to operational issues and questions: Who bears responsibility for system failures? How can the processes of a system be made transparent? How can we ensure that machine learning does not erode fundamental ethical values? Without addressing these concerns, the concept of AI moral agency remains both theoretically debatable and practically unsafe.

Illustrative Case Studies

The academic debate surrounding the notion of artificial moral agency has intensified in recent years, offering valuable refinement to Moor's original typology. Among the most

[2] *Qualia denote the subjective, phenomenal aspects of experience.*
[3] *Mindless morality refers to assigning moral relevance to systems that lack phenomenal consciousness.*

influential voices, Gunkel (2023) revisits the ontological assumptions behind machine morality. He contends that the growing entanglement of human and artificial actors necessitates a shift in how agency and responsibility are conceptualised. This perspective lends weight to the claim that a strictly agent-centric taxonomy overlooks morally relevant contexts in which accountability is co-constituted by both human stakeholders and adaptive artefacts.

Coeckelbergh (2014, pp. 61–77), in his contribution, extends this relational turn by suggesting that moral status should not depend on internal attributes, such as consciousness and intentionality, but instead on the nature and structure of interactions with human agents. His relational conception of artificial agency aligns with the complexity–autonomy–predictability matrix (hereinafter referred to as the RI_3 matrix) proposed here, which frames machine morality as an emergent property of socio-technical systems, rather than a function of internal capacities alone.

This theoretical foundation can be extended to large language models deployed in high-stakes environments (e.g., medicine or finance), where probabilistic and context-sensitive outputs complicate attribution and can generate responsibility gaps (Matthias, 2004). These gaps underscore the need for regulatory mechanisms that can track accountability throughout the system's entire lifecycle, an objective reflected in the layered obligations outlined in the Artificial Intelligence Act adopted by the European Parliament and Council of the European Union (2024) (Regulation (EU) 2024/1689; hereinafter referred to as the AI Act).

Two illustrative cases emphasise the urgency of achieving this objective. In October 2023, a Cruise autonomous vehicle fatally injured a pedestrian in San Francisco after its self-learning perception module failed to recognise the human figure under dim lighting. According to preliminary media reports, over-the-air software updates, deployed after initial regulatory clearance, had altered the vehicle's behaviour, highlighting the structural inadequacy of traditional agency-based taxonomies when applied to self-adaptive systems (Coeckelbergh, 2014; Gunkel, 2023).

A second illustrative case concerns the clinical deployment of a large language model as a decision-support interface, in which hallucinated recommendations and undue deference to fluent outputs can undermine professional judgement unless robust validation and meaningful human oversight are maintained. Even when oversight is formally assigned, the perceived authority of language-based outputs may shape decision-making and contribute to responsibility gaps within the broader socio-technical network (Matthias, 2004).

These scholarly and empirical developments support the paper's central thesis: Moor's typology, while structurally sound, must be extended with relational parameters to remain normatively and analytically robust. By integrating moral responsibility as a distributed and dynamic phenomenon, one shaped by ongoing interactions, rather than static properties, the proposed RI_3 matrix provides a more accurate and operationalisable model for AI governance.

Normative and Regulatory Implications

The normative and regulatory implications for the presented typology confirm that even implicit AI systems lacking self-awareness are now subject to strict legal and technical requirements. At the level of *lex lata*, the cornerstone is the AI Act, which is gradually introducing prohibitions, transparency obligations, and a risk-assessment regime, becoming

fully applicable in 2026 (Digital Strategy). This act classifies most contemporary systems as "high-risk," mandating algorithmic audits, documentation of the design process, and mechanisms for meaningful human control, thereby embedding Moor's implicit ethical agents within a legal framework of accountability.

Based on these obligations, algorithm designers now bear an extensive *ex ante* duty of care. The AI Act couples risk management and human oversight with lifecycle monitoring, while the draft AI Liability Directive (AILD) and the revised Product Liability Directive (PLD) promise a strict and fault-based liability for defective code (European Parliament and Council of the European Union, 2024). Proposals for professional licencing and an "AI Hippocratic Oath" would expose individual engineers to malpractice sanctions (Sharma, 2024). Meanwhile, legal theorists advocate for the imposition of fiduciary duties of loyalty and care whenever information asymmetry arises (Custers *et al.*, 2025). Since autonomy increases harm-forecasting uncertainty, scholars anticipate the use of complementary tools, including mandatory third-party audits (Remolina, 2025, pp. 51–70), compulsory insurance pools (Saul Ewing LLP, 2025), and burden-shifting rules when designers withhold evidence (Kennedys Law, 2024). Despite discussions of distributed responsibility, designers remain the primary focus; they must deliver transparent, explainable, and value-aligned systems or face escalating regulatory and financial exposure (Novelli *et al.*, 2025). The following section demonstrates how these heightened designers' obligations transform both ethics-by-design standards and liability architecture.

Discussion

As part of this preventive infrastructure, the revised PLD 2024/2853 came into force on 8 December 2024. Member states must transfer the directive into national law by 9 December 2026. Unlike the AI Act, the PLD is rooted in ex-post liability, maintaining the EU's strict, no-fault model while adapting it to the realities of digital products. Its innovations include an expanded concept of defect, presumption of causation for complex technologies, and enhanced disclosure requirements. In this respect, the PLD offers a harmonised remedy for victims of AI-related harm, particularly in cases where the AI Act's *ex ante* controls prove insufficient.

Another regulatory element, the proposed AILD, was designed to harmonise national tort law by introducing rebuttable presumption of fault and mandatory disclosure obligations when petitioners encounter evidentiary asymmetry caused by the opacity of AI systems. While the European Parliament had targeted adoption for February 2026, the Commission's updated work programme, released in February 2025, unexpectedly flagged the proposal for potential withdrawal due to a lack of political consensus. At the time of writing, the AILD remains in legislative limbo—formally active but procedurally suspended. If adopted, it would fill the intermediate space between prevention and compensation by refining fault-based liability for AI-specific harm. If withdrawn, member states will revert to their tort doctrines, likely reintroducing the very kind of fragmentation that Moor characterised as a normative "policy vacuum."

Together, these instruments outline a layered liability architecture. The AI Act imposes forward-looking obligations aimed at preventing harm; the PLD provides a harmonised ex-post remedy for when harm occurs; and the AILD, if introduced, would tailor non-contractual fault-based liability to the epistemic challenges posed by AI systems. For regulators and system designers, understanding this interplay is essential for operationalising the paper's proposed RI_3 matrix—complexity, autonomy, and predictability—as each factor corresponds to distinct legal triggers across the evolving EU framework. This point is

especially critical for defence and security deployments, including autonomous weapon platforms, border control analytics, and cyber defence early warning, where high autonomy and low predictability trigger the strictest oversight and assurance regimes.

The findings of this study reaffirm the lasting but limited diagnostic value of James Moor's four-level typology when applied to contemporary generative and self-learning AI systems. The comparative analysis confirms that the distinction between ethical impact, implicit, explicit, and full moral agents remains illuminated in legacy contexts, such as expert systems and rule-based robotics (Moor, 2006, pp. 18–21). However, it weakens in the face of anomalies characteristic of probabilistic, generative systems deployed in high-stakes decision-support, where outputs may exceed predefined boundaries and give rise to "responsibility gaps" that Moor's schema was not designed to interpret (Gunkel, 2023; Matthias, 2004).

Further study led to a relational–contextual extension of Moor's typology. This perspective, inspired by relational ethics (Coeckelbergh, 2014, pp. 61–77) and interactionist epistemology, shifts moral analysis away from the system's internal architecture and towards the socially embedded practices in which that architecture is deployed. Case-based comparisons illustrate the hypothesis that AI systems embedded in dense human feedback loops, such as GPT-4 in therapeutic contexts, exhibit the most intense accountability ambiguities. These instances demonstrate that moral salience emerges not merely from system design but also from its situated use, trust dynamics, and institutional context.

To operationalise this insight, the paper introduced an RI_3 matrix encompassing complexity, autonomy, and predictability, each corresponding to distinct regulatory obligations outlined in the AI Act. Complexity aligns with systemic risk management for high-risk systems; autonomy is balanced with obligations for meaningful human oversight; and predictability is supported by interpretability and traceability requirements. This matrix provides a pragmatic terminology for translating philosophical distinctions into regulatory frameworks and appears compatible with upcoming liability instruments, such as the AILD and PLD.

Moreover, the discussion resonates with long-standing debates over moral agency and moral patiency. While Floridi and Sanders (2004, pp. 349–379) argue for the possibility of attributing moral agency to information artefacts, and Bryson (2010, pp. 63–74) advises against anthropomorphising technological systems, the relational–contextual approach reframes the debate. It posits agency not as an intrinsic, metaphysical property but as a role that emerges from socio-technical entanglements. This perspective clarifies why accountability may shift—from developers to end-users or institutional deployers—as AI systems migrate from isolated environments to embedded, real-world applications.

These findings contribute to the field in three key ways. First, they demonstrate that Moor's typology, although still foundational, necessitates relational augmentation to accommodate the dynamics of modern AI. Second, they present a scalable regulatory framework that can be embedded within risk-based governance. Third, they reorient ontological debates towards actionable questions of relational responsibility, providing ethicists and lawmakers with a nuanced yet applicable utility. In this way, the study positions Moor's legacy as a living conceptual resource—one that evolves in response to the changing topography of artificial moral agency.

Conclusion and Future Directions

This paper critically reviews the regulatory utility of James Moor's four-level typology of machine morality in the context of increasingly autonomous AI systems. While the typology remains conceptually robust, our analysis suggests that it insufficiently captures the relational and contextual dynamics at play in contemporary socio-technical systems. The typology's explanatory power decreases when it is set against the anticipatory logic of the EU's AI Act, as detailed earlier in the Normative and Regulatory Implications section.

To address this gap, an RI_3 matrix—complexity, autonomy, and predictability—is proposed as an operational bridge between philosophical classification and normative governance. This matrix not only complements Moor's levels of machine morality but also enables regulators to more precisely identify responsibility gaps that emerge in environments of epistemic opacity and socio-technical entanglement. By moving beyond an exclusively agent-centric approach, the framework accommodates both structural and relational modes of ethical risk. The same three-dimensional lens can also guide decision-makers in security-critical arenas, from autonomous weapon platforms and cyber-defence early-warning systems to biometric border-control gates, where high complexity, delegated autonomy, and limited behavioural predictability jointly demand the most stringent oversight and contingency planning (De Spiegeleire et al., 2017; Yampolskiy, 2018).

Nonetheless, several limitations restrict the scope of this inquiry. The selected case studies are geographically and regionally narrow; further cross-jurisdictional and longitudinal research is required to assess the generalisation of the proposed matrix across legal and cultural contexts. Moreover, while the matrix aligns qualitatively with emerging liability doctrines, its thresholds for actionable unpredictability or autonomy require empirical adjustment in dialogue with computer scientists and risk evaluators. A final ontological requirement remains: the analysis assumes a relatively stable boundary between moral agency and moral patiency (Coeckelbergh, 2014, pp. 61–77; Floridi and Sanders, 2004, pp. 349–379). Should future AI systems evolve to possess self-modification or goal-generation capacities, the foundational assumptions of Moore's law may require radical revision.

Future research should proceed along three converging paths. Conceptually, scholars should elaborate the relational metrics of moral salience by incorporating insights from social robotics and interactionist epistemologies. Also, large-N studies of diverse AI deployments, including in mobility, healthcare, and generative media, should test whether the matrix reliably predicts accountability gaps. Normatively, interdisciplinary teams should translate the heuristic into auditable compliance indicators that correspond with the AI Act's risk tiers and the forthcoming AILD. Advancing along these trajectories will both validate and refine the framework presented here, ensuring that ethical theory remains both philosophically rigorous and practically responsive to the accelerating pace of artificial agency.

Funding

The research received no external funding.

Disccloser Statement

No potential conflict of interest was reported by the author. The author read and agreed to the published version of the manuscript

Data Availability Statement

Not applicable.

References

Allen, C., Wallach, W. and Smit, I. (2006) 'Why machine ethics?', *IEEE Intelligent Systems*, 21(4), pp. 12–17. doi: 10.1109/MIS.2006.62.

Asaro, P.M. (2007) 'Robots and responsibility from a legal perspective', in *Proceedings of the IEEE international conference on intelligent robots and systems*. San Diego, CA: IEEE, pp. 18–24.

Bryson, J.J. (2009) 'Building persons is a choice', *Ethics and Information Technology*, 11(1), pp. 5–12. doi: 10.1007/s10676-009-9189-3.

Bryson, J.J. (2010) 'Robots should be slaves', in Wilks, Y. (ed.) *Close engagements with artificial companions: Key social, psychological, ethical and design issues*. Amsterdam: John Benjamins, pp. 63–74. doi: 10.1075/nlp.8.11bry.

Cervantes, S., López, S., Castro-Sánchez, L. and Ramos, J. (2019) 'Artificial moral agents: A survey of the current status', *Science and Engineering Ethics*, 26(2), pp. 501–532, doi: 10.1007/s11948-019-00151-x.

Cervantes, S., López, S. and Cervantes, J-A. (2020) 'Toward ethical cognitive architectures for the development of artificial moral agents', *Cognitive Systems Research*, 64, pp. 117–125. doi: 10.1016/j.cogsys.2020.08.010.

Chalmers, D.J. (1996) *The conscious mind: In search of a fundamental theory*. Oxford: Oxford University Press.

Coeckelbergh, M. (2009) 'Virtual moral agency, virtual moral responsibility: On the moral significance of the appearance, perception, and performance of artificial agents', *AI & Society*, 24(2), pp. 181–189. doi: 10.1007/s00146-009-0208-3.

Coeckelbergh, M. (2014) 'The moral standing of machines…', Philosophy & Technology, 27(1), pp. 61–77. doi: 10.1007/s13347-013-0133-8.

Coeckelbergh, M. (2015) 'Artificial agents, good care, and modernity', Theoretical Medicine and Bioethics, 36(4), pp. 265–277, doi: 10.1007/s11017-015-9331-y.

Custers, B., Lahmann, H. and Scott, B.I. (2025) 'From liability gaps to liability overlaps: shared responsibilities and fiduciary duties in AI and other complex technologies', AI & Society, 40, pp. 4035–4050. doi: 10.1007/s00146-024-02137-1.

Dennett, D.C. (1992) *Consciousness explained*. London: Penguin.

De Spiegeleire, S., Maas, M. and Sweijs, T. (2017). *Artificial intelligence and the future of defence: Strategic implications for small- and medium-sized force providers*. The Hague: The Hague Centre for Strategic Studies. Available at: https://hcss.nl/report/artificial-intelligence-and-the-future-of-defense/ (Accessed: 1 May 2025).

European Parliament and Council of the European Union (2024) 'Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)', *Official Journal of the European Union*, L 2024, p. 1689.

Floridi, L. and Sanders, J.W. (2004) 'On the morality of artificial agents', *Minds and Machines*, 14(3), pp. 349–379. doi: 10.1023/B:MIND.0000035461.63578.9d.

Goodall, N.J. (2014) 'Machine ethics and automated vehicles', in Meyer, G. and Beiker, S. (eds.) *Road vehicle automation*. Cham: Springer, pp. 93–102. doi: 10.1007/978-3-319-05990-7_9.

Gunkel, D.J. (2023) *The machine question (revisited)*. Cambridge, MA: MIT Press.

IEEE Standards Association (2021) *IEEE standard model process for addressing ethical concerns during system design (IEEE Std 7000-2021)*. Piscataway, NJ: IEEE.

Kant, I. (2002, 1785) *Groundwork of the metaphysics of morals*. Translated by M. Gregor. Cambridge: Cambridge University Press.

Kennedys Law (2024) *AI liability in the EU: Burden-shifting rules explained*. White paper. Available at: https://kennedyslaw.com/thought-leadership/article/ai-liability-in-the-eu-burden-shifting-rules-explained/ (Accessed: 26 May 2025).

Kurzweil, R. (2005) *The singularity is near: When humans transcend biology*. New York: Viking.

Matthias, A. (2004) 'The responsibility gap: Ascribing responsibility for the actions of learning automata', *Ethics and Information Technology*, 6(3), pp. 175–183. doi: 10.1007/s10676-004-3422-1.

Moor, J.H. (1985) 'What is computer ethics?', *Metaphilosophy*, 16(4), pp. 266–275. doi: 10.1111/j.1467-9973.1985.tb00173.x.

Moor, J.H. (2006) 'The nature, importance, and difficulty of machine ethics', *IEEE Intelligent Systems*, 21(4), pp. 18–21. doi: 10.1109/MIS.2006.80.

Novelli, C., Taddeo, M. and Floridi, L. (2024) 'Accountability in artificial intelligence: what it is and how it works', *AI & Society*, 39, pp. 1871–1882, doi: 10.1007/s00146-023-01635-y.

Remolina, N. (2025) 'AI Governance and Algorithmic Auditing in Financial Institutions: Lessons From Singapore', SSRN (Research Paper). Available at: https://ssrn.com/abstract=5199968 [Accessed: 31 March 2025].

Russell, S.J. (2019) *Human compatible: Artificial intelligence and the problem of control*. New York, NY: Viking.

Saul Ewing LLP (2025) *The Use of AI in the Insurance Policy Lifecycle and Legal Implications.* White paper, 20 February 2025. Available at: https://www.saul.com/ (Accessed: 26 May 2025).Sharma, P. (2024) *Toward an AI Hippocratic oath*. Tilburg: TechReg Press.

Vijayaraghavan, A. and Badea, C. (2024) 'Minimum levels of interpretability for artificial moral agents', *AI & Ethics*, 4, pp. 1–22. doi: 10.1007/s43681-024-00536-0.

Yampolskiy, R.V. (Ed.). (2018) *Artificial intelligence safety and security*. Chapman and Hall/CRC. doi: 10.1201/9781351251389.