Artificial Intelligence as a Moral Phenomenon: Rethinking the Ethical Neutrality of Algorithmic Systems through Rousseau

Abstract

Much of the contemporary discussion on Artificial Intelligence (AI) assumes that AI systems are ethically neutral tools whose moral significance depends largely on how human beings choose to design, deploy, regulate, or use them. This article challenges that assumption by arguing that AI should be understood not merely as a technological instrument but as a moral phenomenon. The central claim is that AI is already morally constituted before it is put to use because human values, assumptions, priorities, and judgments are embedded in its design, training data, algorithms, and modes of deployment. At the same time, AI increasingly shapes the environments within which human moral decisions are made, thereby influencing social practices, institutional processes, and patterns of human interaction.The study adopts a conceptual and philosophical approach, drawing particularly on Jean-Jacques Rousseau's account of moral formation. Rousseau's insight that moral order emerges within humanly created social structures provides a useful framework for understanding AI as a system that is itself a product of collective human values and choices. Through a critical engagement with arguments for technological neutrality, alongside perspectives from philosophy of technology, information ethics, and algorithmic governance, the article demonstrates that AI cannot be separated from questions of morality. AI systems classify, predict, recommend, and make decisions in ways that embody normative assumptions and shape human understanding of the world.The article concludes that AI should be regarded as a morally saturated socio-technical infrastructure rather than a neutral computational tool. Recognising this shifts ethical inquiry beyond questions of external regulation toward a deeper examination of the moral constitution of AI itself. In doing so, the study contributes to ongoing debates in AI ethics and offers a framework for understanding the increasingly pervasive role of AI in contemporary society.

Keywords: Algorithmic Ethics, Artificial Intelligence, Ethical Neutrality, Moral Phenomenon, Rousseau.

Introduction

Artificial Intelligence (AI) has emerged as one of the most consequential developments of the contemporary age. From healthcare and finance to education, communication, and governance, algorithmic systems increasingly mediate the decisions through which individuals and institutions organise social life. Yet, despite the growing influence of AI, a persistent assumption continues to shape much of the discourse surrounding it: the assumption that AI is fundamentally an ethically neutral technology whose moral significance derives only from the purposes to which human beings put it. Within this view, AI is understood as an advanced computational instrument capable of processing information, identifying patterns, and generating predictions, while remaining intrinsically detached from moral content. Ethical concerns arise only subsequently through human use, policy regulation, or external evaluation.

This assumption of neutrality is not merely a practical position; it is a philosophical one. It presupposes a distinction between technology and morality in which technological systems exist prior to ethical judgment and acquire moral significance only when they enter the sphere of human action. Consequently, AI is often portrayed as a passive tool whose ethical implications are derivative rather than constitutive. The present article challenges this understanding. It argues that AI should not be conceived merely as a technological instrument but as a moral phenomenon. This claim does not imply that AI possesses consciousness, autonomy, or moral agency. Rather, it suggests that AI is already morally constituted through the values, assumptions, priorities, and judgments embedded within its design, training, optimisation, and deployment. At the same time, AI increasingly shapes the social and informational environments within which human beings make moral decisions. It is therefore implicated not only in the outcomes of moral action but also in the conditions under which moral action becomes possible.

To develop this argument, the article draws upon the moral and political philosophy of Jean-Jacques Rousseau. While Rousseau is rarely discussed within contemporary AI ethics, his reflections on moral formation offer important conceptual resources for understanding the relationship between human agency, collective organisation, and technological systems. In *The Social Contract*, Rousseau advances the view that moral life does not emerge in isolation but within structures of association governed by shared norms, obligations, and forms of collective organisation (Rousseau, 1762/2002). Human beings become moral not simply through individual rationality but through participation in institutions and practices that shape their understanding of freedom, responsibility, and social obligation.

This insight is especially significant for contemporary discussions of AI. Increasingly, algorithmic systems function as organising structures through which social choices are classified, prioritised, distributed, and constrained. Recommendation systems shape access to information; predictive models influence employment, credit, and security decisions; and automated systems mediate interactions between citizens and institutions. If moral order emerges within collectively constructed systems, as Rousseau suggests, then these technological structures cannot be regarded as morally empty instruments standing outside

the sphere of ethical life. Rather, they become part of the very architecture through which moral and social realities are produced and sustained.

Rousseau's relevance to the present inquiry lies therefore not primarily in his political theory but in his account of the social constitution of moral life. His philosophy directs attention away from morality understood solely as a matter of individual intention and toward the structures through which human action is organised. Moral order, in this sense, is not merely psychological; it is institutional and systemic. Human beings construct social worlds, and those worlds in turn shape human conduct. This perspective provides a valuable framework for understanding AI as a humanly constructed socio-technical system already embedded with normative commitments. Decisions concerning data collection, classification categories, optimisation objectives, risk thresholds, and performance metrics are never value-free. They reflect judgments about what counts as relevant information, desirable outcomes, legitimate forms of prediction, and acceptable distributions of risk and benefit.

Seen from this perspective, AI systems cannot be understood as neutral computational artefacts. They are products of collective human choices and institutional priorities, carrying into operation the assumptions and values that informed their creation. Rousseau's account of moral formation therefore provides a conceptual bridge between the social origins of morality and the moral character of technological systems. If human institutions inevitably embody normative commitments, then AI systems, as products of human design and collective organisation, cannot escape moral constitution.

This position stands in contrast to influential tendencies within contemporary AI ethics that treat ethics as an external corrective applied to otherwise neutral technologies (Jobin et al., 2019). It aligns more closely with traditions in the philosophy of technology that question the possibility of technological neutrality itself. Langdon Winner (1986), for example, argues that technologies often embody specific forms of power and political order within their very design. Similarly, Martin Heidegger (1977) contends that modern technology is not merely a collection of tools but a mode of revealing that shapes how reality appears to human understanding. Technology, on this view, does not simply facilitate action; it structures perception, interpretation, and possibility. Such insights suggest that the moral significance of AI cannot be reduced to questions of usage alone, since technological systems participate in shaping the horizons within which human judgment is exercised.

The historical development of computation further reinforces this point. From early calculating devices and formal systems of logic to modern computational theory, efforts to mechanise reasoning have always reflected particular conceptions of knowledge, order, and rationality. Contemporary AI extends this trajectory in unprecedented ways through machine learning systems trained on vast datasets generated within specific historical and social contexts. These systems do not merely process information; they inherit patterns, classifications, exclusions, and biases embedded within the data from which they learn (Barocas & Selbst, 2016; Bender et al., 2021; Crawford, 2021). What appears as technical output often carries the sediment of prior human judgments and institutional practices.

The implications are profound. If AI systems are shaped by human value structures and simultaneously influence the environments in which human decisions occur, then the distinction between a morally neutral technology and its ethical consequences becomes increasingly difficult to sustain. The moral question is no longer simply how AI should be regulated after it has been developed, but how moral assumptions become embedded within AI systems from the outset and how these systems subsequently participate in the formation of social and moral life.

Accordingly, this article adopts a methodological approach grounded in conceptual analysis, hermeneutic interpretation, normative reconstruction, and critical philosophy of technology. Rather than treating AI as a purely technical object, it examines the moral conditions that underlie its design, operation, and deployment. Central to this analysis is the distinction between a moral agent and a moral phenomenon. A moral agent is capable of intention, responsibility, and accountability for action. A moral phenomenon, by contrast, need not possess consciousness or agency. Its significance lies in its capacity to shape morally relevant forms of perception, judgment, action, and social organisation. AI systems are understood here as moral phenomena in precisely this sense. They do not act morally; rather, they organise the informational and institutional environments within which moral action occurs.

The central argument of this article, therefore, is that AI should be understood as a morally constituted socio-technical infrastructure rather than a neutral computational tool. By examining AI through Rousseau's account of moral formation and situating that analysis within broader debates in philosophy of technology and AI ethics, the article seeks to demonstrate that morality is not something added to AI from the outside. It is already present within the structures through which AI is imagined, designed, trained, and integrated into human life.

Overview of the History of AI

The history of Artificial Intelligence (AI) is often narrated as a story of technological innovation, progressing from primitive calculating devices to increasingly sophisticated computational systems. While this account is not incorrect, it is incomplete. To understand AI adequately, one must situate it within a much longer intellectual history concerned with the human desire to externalise reasoning, organise knowledge, and extend the capacities of judgment beyond the immediate limits of individual cognition. AI is therefore not simply a recent technological achievement. It is the latest manifestation of a longstanding philosophical and cultural aspiration to render aspects of human thought amenable to formal representation and systematic execution (Russell & Norvig, 2021, pp. 1–15).

The earliest expressions of this aspiration can be discerned in devices such as the abacus, which provided a structured means of performing calculation outside the human mind (Campbell-Kelly et al., 2014, pp. 3–11). Although such instruments cannot meaningfully be described as intelligent, they nonetheless mark an important conceptual moment in the history of cognition. They represent an early separation between human reasoning and the

symbolic operations through which reasoning may be expressed. What begins as an aid to calculation gradually develops into a broader project of representing thought itself through formal systems capable of operating independently of immediate human judgment.

This intellectual trajectory acquires greater philosophical significance in the early modern period. René Descartes' rationalist project sought to establish a method by which knowledge could be generated through systematic and orderly reasoning (Descartes, 1641/1996). Descartes did not anticipate contemporary AI, nor did he propose the mechanisation of intelligence in its modern form. Nevertheless, his conviction that rational inquiry could be structured according to formal procedures contributed to a broader intellectual climate in which thought increasingly came to be understood as analysable, decomposable, and reproducible. The significance of Cartesian rationalism for the present discussion lies not in any direct technological influence but in its affirmation of the possibility that aspects of human reasoning might be represented through ordered systems independent of particular experiences or subjective contexts.

This movement toward the formalisation of thought reaches a decisive stage in the work of Alan Turing. Through his theory of computation, Turing demonstrated that symbolic operations could be executed mechanically through the application of formal rules (Turing, 1950). His conception of the universal machine established the theoretical foundations of modern computing and provided a framework within which intelligence itself could be imagined as a process susceptible to computational simulation. Yet even at this formative stage, the apparent neutrality of computation warrants closer examination. Computational systems do not emerge in a vacuum. Their design presupposes decisions regarding what counts as relevant information, what forms of reasoning are desirable, and which outcomes are worth pursuing. Such decisions are never purely technical. They are shaped by assumptions about value, purpose, and significance that precede computation itself (Afisi, 2021, 2026).

The transition from classical computation to contemporary AI introduces an even more profound transformation. Earlier computational systems largely operated through predefined rules explicitly specified by programmers. Modern AI, by contrast, increasingly relies on machine learning models that infer patterns from vast quantities of data rather than following predetermined logical instructions (Goodfellow et al., 2016, pp. 1–18). This development has significant philosophical implications. AI systems no longer simply execute human commands; they learn from traces of human activity accumulated across social, economic, and institutional contexts. Historical data become the medium through which algorithmic systems acquire their predictive capacities.

It is at this point that questions of moral significance become unavoidable. Scholars such as Barocas and Selbst (2016) have shown that algorithmic systems frequently reproduce patterns of inequality embedded within the datasets from which they learn, even in the absence of explicit discriminatory intent. Likewise, Crawford (2021) demonstrates that contemporary AI is inseparable from the social infrastructures, labour practices, and institutional arrangements

that make its operation possible. Far from existing as detached computational mechanisms, AI systems are embedded within histories of power, exclusion, and social organisation.

Advocates of technological neutrality often respond that AI merely reflects the world as it is. According to this position, algorithmic systems do not generate moral content; they simply identify and reproduce patterns already present in data (Floridi, 2014, pp. 94–99). On this account, AI functions as a mirror rather than a participant in moral life. While this argument possesses intuitive appeal, it remains philosophically insufficient. It overlooks the fact that data are never given in a raw or value-free form. Decisions regarding what data are collected, how categories are defined, which variables are included, and what objectives algorithms are designed to optimise involve normative judgments at every stage of the process (Mittelstadt et al., 2016, pp. 4–8). Furthermore, AI systems do not merely reflect social realities; they increasingly contribute to their reorganisation through processes of classification, recommendation, prediction, and ranking (Beer, 2017, pp. 3–8; Zuboff, 2019, pp. 93–102).

These considerations invite a deeper philosophical interrogation of the relationship between technology and morality. It is here that Rousseau's account of moral formation offers a particularly illuminating perspective. Rousseau argues that moral life emerges through systems of collective human organisation rather than existing independently of them (Rousseau, 2002; Cohen, 2010; Binns, 2022). Human institutions embody the values, priorities, and forms of association through which individuals come to understand themselves and others. If this insight is extended to technological systems, AI can no longer be regarded as a passive instrument that merely processes information. Rather, it appears as a structured expression of collective human judgments, carrying within itself the normative assumptions that shaped its development.

From this standpoint, the history of AI is not simply a history of technical refinement. It is also a history of evolving conceptions of knowledge, rationality, authority, and social organisation. Each stage in the development of computational systems reflects particular assumptions about what ought to be measured, predicted, classified, and controlled. The movement from mechanical calculation to machine learning therefore represents more than an increase in computational power. It reflects a deepening effort to organise human life through increasingly complex systems of representation and decision-making.

This interpretation finds support within broader traditions in the philosophy of technology. Langdon Winner (1986) argues that technological artefacts often embody specific forms of power and political ordering. Technologies are not neutral instruments subsequently employed for social purposes; rather, they frequently carry social arrangements within their design and operation. Similarly, Heidegger (1977) contends that technology shapes the way reality is disclosed to human understanding. Technology is not merely a means to an end but a mode of revealing that influences how the world becomes intelligible. Taken together, these perspectives challenge the assumption that AI represents a purely technical evolution detached from moral significance.

Recent analyses of digital capitalism further reinforce this critique. Zuboff (2019) argues that contemporary computational systems increasingly function within economic structures dedicated to the extraction, prediction, and commodification of human behaviour. Under such conditions, AI becomes more than a computational artefact. It becomes part of an extensive socio-economic infrastructure through which behaviour is monitored, interpreted, and influenced. The moral implications of AI therefore cannot be separated from the institutional environments within which it operates.

Seen in this light, the history of AI reveals not a neutral progression of technological sophistication but a layered process through which human capacities for reasoning, prediction, and decision-making have been externalised within increasingly complex systems of social organisation. Rousseau's theory of moral formation helps illuminate the significance of this development. Humanly constructed systems do more than coordinate action; they shape the normative conditions under which individuals understand themselves, relate to others, and exercise judgment (Rousseau, 2002). Contemporary AI systems increasingly perform a comparable function. Through mechanisms of classification, prediction, recommendation, and visibility, they participate in structuring social expectations and influencing practical decisions.

A frequently cited example is the use of predictive algorithms within criminal justice systems, particularly tools such as COMPAS in the United States. Although presented as objective instruments of risk assessment, these systems have been criticised for reproducing patterns of racial inequality embedded within historical data (Barocas & Selbst, 2016; Ugwulebo, 2026). Such cases illustrate that AI does not merely process information in a neutral manner. Rather, it can reproduce and amplify moral and social assumptions inherited from the environments in which it is developed and trained.

The history of AI therefore points beyond the evolution of computational capacity alone. It reveals the expansion of humanly constructed systems through which social life is interpreted, organised, and governed. Read through Rousseau's account of moral formation, this history suggests that AI is neither morally empty nor external to ethical life. Like other products of collective human activity, it bears the imprint of the values, judgments, and forms of association from which it emerges. The development of AI is consequently not only a technological story but also a moral one.

Moral Philosophy and the Morality of AI

Any serious attempt to argue that Artificial Intelligence (AI) constitutes a moral phenomenon must begin by clarifying the meaning of morality itself. This preliminary task is necessary because much of the resistance to attributing moral significance to AI arises from a particular understanding of morality—one that locates the moral exclusively within the domain of conscious intention, rational deliberation, and autonomous agency. If morality is defined solely in these terms, then the conclusion appears straightforward: machines cannot be moral because they neither possess consciousness nor exercise intentional choice. Yet such a

conclusion rests upon a narrower conception of morality than the history of moral philosophy permits.

Classical ethical traditions reveal a more complex picture. In Aristotle's moral philosophy, ethical life is not reducible to isolated acts of conscious intention but is cultivated through habits, practices, and forms of communal existence within which character is shaped (Aristotle, 2009). Moral evaluation concerns not only what individuals intend but also the social and institutional contexts that sustain particular ways of living. Likewise, although Kant places considerable emphasis on rational autonomy and moral agency, the exercise of moral judgment remains inseparable from the structured conditions within which practical reason operates (Kant, 1996). In both traditions, morality extends beyond isolated moments of choice to encompass the broader frameworks that make moral action intelligible.

Nevertheless, these traditions remain primarily centred upon the human subject as the locus of moral responsibility. The challenge presented by AI is not that machines have become moral agents in the traditional sense, but that moral consequences increasingly emerge through systems that operate between human intentions and social outcomes. Contemporary algorithmic systems participate in the organisation of social life in ways that complicate conventional distinctions between action, agency, and responsibility. As a result, understanding AI requires a conception of morality capable of addressing not only agents but also the structures through which moral experience is mediated.

Such a perspective can be found in moral traditions that emphasise the social and institutional dimensions of ethical life. MacIntyre (1984), for example, argues that moral reasoning is embedded within historically constituted practices and traditions rather than existing as an abstract exercise of individual rationality. Similarly, Hume's moral philosophy locates ethical judgment within sentiments, habits, and forms of social interaction that precede purely rational calculation (Hume & Hume, 1978). These approaches shift attention away from morality as a property possessed exclusively by individual minds and toward morality as something sustained through shared practices, institutions, and systems of social organisation.

This broader understanding of morality becomes particularly important when examining AI. Algorithmic systems increasingly shape what individuals encounter, how information is prioritised, what options appear available, and which forms of behaviour are rewarded or discouraged (Birhane, 2021). Recommendation engines influence cultural consumption; predictive systems affect access to employment, credit, and security; and classification algorithms shape institutional judgments about risk, relevance, and opportunity. Under such conditions, the moral question can no longer be confined to what individuals intend. It must also address how technological systems participate in shaping the conditions under which intentions themselves are formed.

At this point, an important objection must be considered. A significant body of scholarship maintains that AI cannot properly be described as moral because it lacks consciousness, intentionality, and genuine understanding (Searle, 1980). Machines may simulate intelligent behaviour, but they do not possess awareness of the meanings associated with their actions.

Consequently, attributing morality to AI appears to involve a category mistake, confusing computational processes with ethical agency.

This objection is persuasive only if moral significance is equated entirely with moral agency. Yet the two concepts are not identical. To acknowledge that something is morally significant is not necessarily to claim that it is morally responsible. Many of the institutions that profoundly shape human life—legal systems, bureaucratic structures, financial markets, and administrative organisations—possess neither consciousness nor intention. Nevertheless, they are routinely subjected to moral evaluation because they influence the distribution of opportunities, burdens, rights, and responsibilities within society (Winner, 1986, pp. 22–29). Their significance lies not in their capacity to choose but in their capacity to structure human action and social outcomes.

A useful illustration can be found in algorithmic risk-assessment systems employed within criminal justice institutions. Systems such as COMPAS do not possess moral awareness, nor can they be held responsible in the manner of human agents. Yet their classifications influence decisions regarding bail, sentencing, and assessments of future risk. The consequences of these classifications affect liberty, fairness, and access to social opportunity. The moral significance of such systems therefore derives not from their possession of agency but from their role in shaping outcomes that carry profound ethical consequences (Barocas & Selbst, 2016; Ugwulebo, 2026).

It is at this juncture that Rousseau's contribution becomes particularly illuminating. Rousseau does not conceive morality primarily as a matter of isolated individual intention. Rather, he understands moral life as emerging within the structures through which individuals are constituted as members of a political and social community (Rousseau, 2002). Human beings do not become moral in isolation; they become moral through participation in shared forms of association governed by norms, obligations, and collective understandings of freedom and responsibility.

This insight is vividly expressed in Rousseau's discussion of the general will:

"Whoever refuses to obey the general will shall be constrained to do so by the whole body; which means nothing other than that he shall be forced to be free" (Rousseau, 2002, pp. 18–19).

The significance of this claim lies not in its political implications alone but in its broader account of moral formation. For Rousseau, moral order is not something imposed upon social structures from the outside. It is generated within the very institutions and practices through which collective life is organised. Individuals come to understand obligation, responsibility, and freedom through participation in structured forms of social existence.

Applied to contemporary AI, this insight suggests an important shift in perspective. The central question is not whether AI behaves morally in the way human beings do. Rather, it is whether AI systems participate in shaping the moral environments within which human agency is exercised. Once framed in this way, the moral significance of AI becomes difficult

to dismiss. Algorithmic systems increasingly mediate access to information, opportunities, social recognition, and institutional resources. In doing so, they contribute to the formation of the contexts within which individuals deliberate, judge, and act.

Rousseau's importance here lies precisely in his refusal to separate moral formation from the structures that sustain social life. Institutions do not merely regulate behaviour after moral subjects have already been formed; they participate in the formation of those subjects. They help define what counts as obligation, fairness, legitimacy, and freedom. Contemporary AI systems increasingly occupy a similar position. By organising visibility, classification, recommendation, and prediction, they shape the informational environments within which moral understanding develops.

This argument is further supported by contemporary scholarship on algorithmic governance. Mittelstadt et al. (2016, pp. 3–8) demonstrate that AI systems incorporate normative assumptions through choices regarding data selection, optimisation objectives, classification criteria, and thresholds of acceptable risk. Such decisions are not ethically neutral technical operations. They represent judgments concerning what ought to count as fair, relevant, efficient, or desirable. In this respect, moral commitments become embedded within the operational logic of algorithmic systems themselves.

A critic might respond that moral responsibility remains entirely with human designers, developers, and institutions rather than with AI systems as such (Floridi& Sanders, 2004, pp. 349–351). In an important sense, this is undoubtedly correct. The present argument does not seek to transfer moral responsibility from human beings to machines. Rather, it seeks to recognise that moral effects are increasingly mediated through technological systems whose operations may be partially autonomous, opaque, and difficult to contest (Pasquale, 2015). Human responsibility remains, but it is exercised through structures that themselves shape moral outcomes.

Heidegger's philosophy of technology strengthens this position. For Heidegger (1977), technology is not simply a collection of instruments available for human use. It is a mode of revealing that shapes how reality becomes intelligible. Technological systems disclose certain possibilities while obscuring others. They influence not only what human beings do but also how they perceive, interpret, and understand the world. AI systems increasingly perform this revelatory function. They frame attention, structure interpretation, and influence decision-making in ways that extend beyond the intentions of individual users.

Morality in the age of AI must therefore be understood in a more distributed and structural sense than traditional accounts of moral agency alone permit. Moral significance does not reside exclusively within conscious subjects. It is also present within the systems, institutions, and technological arrangements that organise human action and shape social life. Rousseau's account of moral formation provides an important philosophical foundation for this claim because it situates morality within collective structures rather than isolated acts of subjectivity (Rousseau, 2002, pp. 6–9; MacIntyre, 1984, pp. 216–225).

From this perspective, AI does not become morally significant because it replaces human agency. It becomes morally significant because it increasingly shapes the conditions within which human agency is exercised. Its influence is exercised through the structuring of perception, judgment, opportunity, and action. It is this capacity to participate in the formation of moral conditions, rather than any possession of consciousness or intention, that justifies understanding AI as a moral phenomenon.

Beyond AI Ethics and the Ethics of AI: Ethical Neutrality as a Practical Illusion

Contemporary discussions of Artificial Intelligence (AI) are frequently organised around a distinction between *AI ethics* and the *ethics of AI* (Floridi, 2014; Jobin et al., 2019; Ugwulebo, 2026b). At first glance, the distinction appears both useful and intuitive. *AI ethics* is generally concerned with the normative principles that ought to guide the design, development, and deployment of AI systems, while the *ethics of AI* seeks a broader philosophical understanding of the implications of AI for human life, social relations, and moral existence. The distinction has undoubtedly contributed to conceptual clarity within contemporary scholarship. Yet it often carries with it an assumption that deserves closer scrutiny: the assumption that AI is initially a morally neutral technology whose ethical significance emerges only through its application, regulation, or social consequences.

The present argument does not reject the analytical distinction between *AI ethics* and the *ethics of AI* as such. Rather, it questions the neutrality thesis that frequently underlies that distinction. The concern is not with the usefulness of distinguishing between ethical governance and philosophical reflection, but with the presupposition that morality enters AI only after technological systems have already been designed and put into operation. Such a presupposition subtly reinforces the belief that technology and morality exist as separate domains, intersecting only at the point of use. It is precisely this assumption that requires reconsideration.

To clarify the argument, the neutrality thesis under examination is not the simplistic claim that AI systems are free from bias or unaffected by social conditions. Few serious scholars maintain such a position today. The stronger and more sophisticated version of neutrality asserts that moral significance originates primarily from human decisions regarding how AI is used, regulated, or evaluated, while the technological system itself remains analytically distinct from moral content. On this view, ethical concerns arise around AI but not within AI. Technology serves as the instrument, while morality enters only through external human intervention.

The difficulty with this position is that it misconstrues the nature of technological systems themselves. AI does not emerge from an ethical vacuum. It is conceived, designed, trained, and deployed within socio-technical environments already structured by particular values, priorities, and assumptions about the world. What appears as a technical decision often rests upon deeper judgments concerning what should be measured, optimised, classified, predicted, or controlled. Consequently, the boundary between technical design and moral evaluation is far less distinct than the neutrality thesis suggests.

This point is evident in Helen Nissenbaum's account of contextual integrity. Nissenbaum (2010) argues that information systems function within normative expectations governing the appropriate flow of information. Questions about privacy, disclosure, access, and use are therefore inseparable from judgments about what ought to occur within particular social contexts. Information technologies do not simply process data; they operate within normative frameworks that give meaning to the movement of information itself. Ethical considerations are therefore present from the outset rather than appearing only after technological deployment.

A similar conclusion emerges from David Beer's analysis of algorithmic power. Beer (2017) observes that algorithms do not merely organise information. They participate in shaping visibility, recognition, and social ordering. Through processes of ranking, recommendation, filtering, and classification, algorithmic systems influence what becomes salient and what recedes into obscurity. Their significance lies not only in what they compute but also in how they structure social attention. Such influence cannot adequately be described in morally neutral terms because it bears directly upon questions of access, inclusion, authority, and opportunity.

The critique of neutrality becomes even more compelling when considered alongside Zuboff's analysis of surveillance capitalism. Zuboff (2019) demonstrates that contemporary digital infrastructures are embedded within economic systems organised around the extraction, prediction, and modification of human behaviour. Human experience is increasingly translated into data that can be analysed, commodified, and used to shape future conduct. Under these conditions, AI functions not as an isolated technological artefact but as part of a broader architecture of behavioural influence and value extraction. The moral significance of such systems is not accidental; it is inseparable from the purposes and institutional arrangements that sustain them.

Defenders of neutrality may nevertheless insist that AI remains fundamentally non-moral because it possesses neither intention nor purpose of its own. Moral responsibility, they argue, belongs entirely to designers, corporations, policymakers, and users rather than to the technological systems themselves (Floridi, 2014). In one sense, this claim is correct. The present argument does not seek to attribute moral agency to machines. Yet the absence of moral agency does not entail the absence of moral significance. The question is not whether AI intends, but whether AI participates in shaping the conditions under which human action, judgment, and responsibility are exercised.

This distinction becomes particularly important in light of contemporary research on algorithmic decision-making. Mittelstadt et al. (2016) show that algorithmic systems embody normative assumptions through decisions concerning data selection, optimisation objectives, classification categories, and thresholds of acceptable risk. Such choices are not merely technical procedures. They represent judgments about fairness, relevance, efficiency, and legitimacy. Moral commitments become operationalised within computational structures long before questions of regulation or ethical oversight arise.

At this point, Rousseau's philosophy offers a particularly illuminating perspective. Rousseau rejects the notion that moral order is externally imposed upon an otherwise neutral social world. Instead, he understands moral life as emerging from the very processes through which human beings organise themselves collectively (Rousseau, 2002; Cohen, 2010; Binns, 2022). Social institutions do not first exist as neutral mechanisms to which moral values are later attached. Rather, they embody judgments concerning authority, obligation, freedom, inclusion, and common purpose from the moment of their formation.

This insight has significant implications for understanding AI. If collective forms of organisation inevitably embody normative commitments, then technologies produced within such forms of organisation cannot be morally empty. Human beings do not first construct value-free systems and subsequently decide how to use them. The act of construction itself already reflects assumptions about what ought to be promoted, protected, measured, coordinated, or excluded. From a Rousseauian perspective, neutrality therefore appears less as a genuine condition and more as a practical illusion generated by our tendency to separate technological form from social purpose.

The force of this argument lies precisely in its refusal to anthropomorphise technology. AI systems need not possess consciousness, intentionality, or moral awareness in order to be morally significant. Their significance arises from their role within larger structures of human action and social organisation. They are not isolated computational entities operating independently of human values. Rather, they are institutionalised infrastructures through which values are translated into operational procedures and social outcomes (Bommasani et al., 2021).

Heidegger's philosophy of technology further deepens this critique. Heidegger (1977) argues that technology is not simply an instrument available for human use; it is a mode of revealing that shapes how reality appears to us. Technology discloses some possibilities while concealing others. It frames the horizon within which understanding and action become possible. AI systems increasingly perform such a function. Through prediction, recommendation, classification, and optimisation, they influence what can be seen, known, anticipated, and acted upon. This capacity to structure perception and possibility introduces an unavoidable normative dimension into technological operation.

From this perspective, ethics cannot be understood as a corrective mechanism applied to otherwise neutral systems. The task of ethical inquiry is not to add morality to AI after the fact but to uncover and critically examine the moral architecture already embedded within technological forms. What is commonly described as *AI ethics* becomes, in part, an exercise in revealing the normative assumptions concealed within systems that often present themselves as objective or value-free. Likewise, the *ethics of AI* becomes inseparable from questions concerning the moral constitution of technological systems themselves.

The distinction between *AI ethics* and the *ethics of AI* therefore becomes less stable than it initially appears. Both approaches frequently presuppose a separation between technology and morality that cannot be sustained in practice. AI systems are not neutral entities awaiting

ethical interpretation. They are already shaped by historical, social, economic, and institutional judgments before they are ever deployed in the world.

A more adequate understanding of AI begins from the recognition that technological systems are always already ethically structured. Ethics does not enter AI from the outside; it is present within the assumptions, classifications, priorities, and forms of organisation through which AI is constituted. Rousseau's account of moral formation reinforces this conclusion. If moral order is inseparable from the systems human beings collectively construct, then AI, as a product of such collective construction, cannot stand outside the moral sphere. It is part of the continuing formation of social life and moral experience, whether this fact is explicitly acknowledged or not.

Rethinking AI as a Moral Phenomenon

To rethink Artificial Intelligence (AI) as a moral phenomenon is not merely to propose a new vocabulary for discussing technology. It is to challenge a deeply entrenched assumption about the relationship between technology and morality itself. Much of the contemporary discourse on AI, whether within technical research, policy development, or ethical governance, continues to proceed from the premise that AI is fundamentally a neutral instrument whose moral significance arises only through human use, misuse, regulation, or oversight (Jobin et al., 2019, pp. 390–392). Ethical questions are therefore treated as secondary considerations, emerging only after the technological system has already been constituted. The argument developed in this article suggests otherwise. It contends that such a conception fails to capture the manner in which AI is produced, embedded, and experienced within social life.

At the heart of this reconsideration lies Rousseau's account of moral formation. Rousseau's enduring significance extends beyond questions of political legitimacy and the social contract narrowly understood. His deeper philosophical contribution lies in the recognition that moral life is not external to the structures through which human beings organise themselves collectively. Moral order does not arrive after institutions have been established; rather, it emerges through the very processes by which those institutions are constituted and sustained (Rousseau, 2002; Cohen, 2010; Binns, 2022). Human beings create social worlds, and in creating them, they simultaneously create the normative conditions that shape responsibility, obligation, freedom, and social belonging.

This insight provides an important lens through which AI may be reconsidered. Artificial intelligence systems are not natural occurrences. They are products of human choices, institutional priorities, economic interests, and technical judgments. Decisions concerning data selection, model architecture, optimisation objectives, classification categories, and deployment contexts are never merely computational matters. They involve assumptions about what counts as relevant knowledge, desirable outcomes, acceptable risks, and legitimate forms of social ordering. Such decisions reveal that moral considerations are not

added to AI after its construction; they are already implicated in the processes through which AI comes into being.

From this perspective, the prevailing neutrality thesis becomes increasingly difficult to sustain. The assumption that AI exists as a morally empty instrument awaiting ethical evaluation depends upon a separation between technology and value that rarely exists in practice. Information systems do not simply mediate human activity from a position of detachment. As Floridi (2014) observes, they increasingly shape the environments within which human action unfolds. Yet mediation itself is not a neutral activity. To mediate is to organise access, prioritise possibilities, structure visibility, and influence judgment. Every act of mediation contains assumptions about what should be emphasised, ignored, encouraged, or constrained.

Contemporary scholarship further reinforces this point. Crawford (2021) demonstrates that AI systems emerge from material infrastructures shaped by histories of labour, extraction, inequality, and institutional power. Likewise, critical studies of algorithmic governance reveal how systems inherit and reproduce patterns embedded within the social realities from which their data are drawn. What appears as technical output often reflects deeper assumptions concerning classification, authority, legitimacy, and value. AI therefore does not arise within a moral vacuum. It is born within already existing social worlds and carries traces of those worlds into its operation.

A familiar objection nevertheless persists. AI, it is argued, cannot properly be regarded as a moral phenomenon because it lacks consciousness, intentionality, and the capacity for moral reasoning (Searle, 1980, pp. 417–424). Only beings capable of deliberation and responsibility can be considered moral in any meaningful sense. This objection remains persuasive if morality is understood exclusively through the lens of moral agency. However, the argument advanced here does not attribute moral agency to AI. Rather, it distinguishes between moral agency and moral significance.

This distinction is crucial. Many structures that profoundly influence human life possess neither consciousness nor intention. Institutions, bureaucracies, markets, and legal systems are not moral agents in the conventional sense, yet they remain subject to moral evaluation because they shape opportunities, distribute burdens, and influence social outcomes. Their significance lies not in their capacity to choose but in their capacity to organise human existence. AI systems increasingly occupy a similar position. As Barocas and Selbst (2016) have shown, algorithmic systems can generate unequal outcomes even in the absence of explicit discriminatory intent. Likewise, Mittelstadt et al. (2016) demonstrate that normative assumptions are embedded in decisions concerning data classification, optimisation metrics, and thresholds of acceptable risk. Such findings suggest that moral relevance does not depend solely upon intention; it may also arise through structure, mediation, and consequence.

Rousseau's framework provides a particularly useful way of understanding this phenomenon. For Rousseau, morality is not confined to isolated acts of individual choice. It is embedded within the forms of association through which collective life is organised. Human beings

become moral subjects through participation in institutions, practices, and structures that shape their understanding of obligation and freedom (Rousseau, 2002). Moral formation is therefore not merely personal; it is social, institutional, and systemic.

When viewed through this lens, AI appears less as a neutral tool and more as a participant in the organisation of moral life. It does not make moral decisions in the way human beings do, nor does it replace human responsibility. Rather, it increasingly shapes the conditions within which moral judgment is exercised. Through processes of classification, prediction, recommendation, and ranking, AI influences what individuals perceive, what options become available, and what courses of action appear reasonable or desirable. Its significance lies not in possessing morality but in participating in the formation of the contexts within which morality is lived.

This interpretation resonates strongly with broader traditions in the philosophy of technology. Heidegger's analysis of technology as a mode of revealing offers an important insight in this regard. Technology, for Heidegger (1977), is not simply instrumental. It discloses reality in particular ways, making certain possibilities visible while concealing others. AI systems increasingly perform such a function. They frame attention, organise information, and influence how individuals and institutions interpret the world around them. Similarly, Beer (2017) demonstrates how algorithmic systems exercise diffuse forms of social power that shape behaviour and social organisation without necessarily operating through direct coercion.

The implications become even more profound when considered alongside Zuboff's (2019) analysis of surveillance capitalism. Within contemporary digital economies, AI systems are frequently embedded within infrastructures designed to predict, influence, and modify behaviour. Human experience becomes a source of behavioural data, while prediction becomes a mechanism of economic value creation. Under such conditions, AI does more than process information; it participates in shaping the horizons of human choice itself. The moral significance of such systems therefore extends beyond questions of technical performance or regulatory compliance. It reaches into the very conditions under which autonomy, agency, and social participation are exercised.

Taken together, these considerations point toward a necessary philosophical conclusion. AI should not be understood as a morally neutral technology to which ethical concerns are subsequently attached. Rather, it should be understood as a socio-technical formation already embedded within networks of value, power, meaning, and collective organisation. Its moral significance arises not because it possesses consciousness or moral agency, but because it participates in structuring the conditions through which moral life unfolds.

Rousseau's framework does more than support this conclusion; it provides its conceptual foundation. If moral order emerges through systems of collective human construction, then technologies produced within those systems cannot stand outside moral reality. AI is one of the contemporary forms through which collective human judgments become institutionalised, operationalised, and reproduced. To regard it as ethically neutral is therefore to overlook both

its origins and its effects. AI is not external to moral life. It has become one of the increasingly significant ways through which moral life is organised, mediated, and experienced in the contemporary world.

Conclusion

This article has argued that Artificial Intelligence (AI) should be understood not merely as a technological artefact or computational instrument, but as a moral phenomenon embedded within the fabric of contemporary social life. Against the widely held assumption that AI is ethically neutral and acquires moral significance only through its application, regulation, or misuse, the analysis has demonstrated that moral considerations are already present within the conditions of its design, development, and operation. The question, therefore, is not simply what human beings do with AI, but how AI itself participates in shaping the environments within which human judgment, action, and social relations unfold.

Drawing upon Jean-Jacques Rousseau's account of moral formation, the article has sought to show that morality is not something external to systems of collective human construction. Rather, moral order emerges through the very structures, institutions, and practices by which human beings organise their common life. Rousseau's significance for contemporary debates on AI lies precisely in this insight. If moral life is constituted within the forms through which collective existence is organised, then technological systems produced within those forms cannot stand outside moral reality. AI systems are not detached mechanisms operating beyond the sphere of value; they are products of human choices, social priorities, institutional arrangements, and historical inheritances. As such, they inevitably carry the imprint of the moral worlds from which they emerge.

The discussion has further shown that the dominant language of neutrality obscures more than it reveals. AI systems do not merely process information; they classify, prioritise, predict, recommend, and structure access to opportunities and resources. In doing so, they increasingly participate in shaping the conditions under which individuals perceive reality, exercise judgment, and make decisions. Their significance therefore extends beyond questions of technical efficiency or functional performance. They have become part of the architecture through which contemporary moral life is mediated and organised.

To recognise AI as a moral phenomenon is not to attribute consciousness, intention, or moral agency to machines. Such a claim would be philosophically unwarranted. Rather, it is to acknowledge that moral significance is not exhausted by agency alone. Moral life is also constituted through the structures that shape human action, the institutions that organise social relations, and the systems that influence what can be known, valued, or chosen. AI increasingly occupies such a position. Its moral importance lies not in replacing human responsibility but in shaping the contexts within which responsibility is exercised.

The broader implication of this argument is that ethical reflection on AI must move beyond the language of external governance and corrective regulation, important though these remain. The deeper philosophical task is to examine the moral assumptions, value commitments, and forms of power already embedded within technological systems. Ethics, in

this sense, is not an activity performed after technology has been created. It is a mode of inquiry directed toward uncovering the normative structures that are present from the beginning.

Ultimately, the significance of AI for moral philosophy may not lie in whether machines can become moral agents, but in how their emergence compels us to rethink the location of morality itself. AI confronts us with the possibility that moral life is not confined to individual intentions or isolated acts of choice. It is also woven into the systems through which human beings organise knowledge, coordinate action, and imagine their collective future. If this is so, then AI must be understood not as something external to moral life, but as one of the increasingly influential ways through which moral life is constituted, mediated, and experienced in the twenty-first century.

References

Afisi, O. T. (2026), The interplay of philosophical logic and computer science: Foundations, logical connectives, and contemporary computational applications" *Advanced Research Journal*, Vol.13, Issue 2, pp. 109-123, DOI:https://doi.org/10.71350/30621925113

Afisi, O. T. (2021) Karl Popper's Social Engineering: Piecemeal or 'Many-Pieces-at-Once'?. In: Afisi O.T. (eds) Karl Popper and Africa: Knowledge, Politics and Development. Springer, Cham. https://doi.org/10.1007/978-3-030-74214-0_3

Aristotle. (2009). *Nicomachean Ethics* (W. D. Ross, Trans.). Oxford University Press.

Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *California Law Review*, 104(3), 671-732.

Beer, D. (2017). The Social Power of Algorithms. *Information, Communication & Society*, 20(1), 1-13.

Bender, E. M., Gebru, T., McMillan-Major, A., &Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610-623.

Binns, R. (2022). Algorithmic Sovereignty and the Social Contract. *Philosophy & Technology*, 35(2), Article 38.

Birhane, A. (2021). Algorithmic Injustice: A Relational Ethics Approach. *Patterns*, 2(2), 100205.

Bommasani, R., et al. (2021). *On the Opportunities and Risks of Foundation Models*. Stanford Centre for Research on Foundation Models.

Campbell-Kelly, M., Aspray, W., Ensmenger, N., & Yost, J. (2014). *Computer: A History of the Information Machine* (3rd ed.). Westview Press.

Crawford, K. (2021). *Atlas of AI*. Yale University Press.

Cohen, J. (2010). *Rousseau: A Free Community of Equals.* Oxford University Press.

Descartes, R. (1996). *Meditations on First Philosophy*(J. Cottingham, Trans.). Cambridge University Press. (Original work published 1641)

Floridi, L. (2014). *The Ethics of Information*. Oxford University Press.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Heidegger, M. (1977). *The Question Concerning Technology and Other Essays*(W. Lovitt, Trans.). Harper & Row.

Hume, D. (1978). *A Treatise of Human Nature* (2nd ed., L. A. Selby-Bigge & P. H. Nidditch, Eds.). Oxford University Press. (Original work published 1739-1740)

Jobin, A., Ienca, M., &Vayena, E. (2019). The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence*, 1, 389-399.

Kant, I. (1996). *The Metaphysics of Morals*(M. Gregor, Trans.). Cambridge University Press.

MacIntyre, A. (1984). *After Virtue*. University of Notre Dame Press.

Mittelstadt, B., Allo, P., Taddeo, M., Wachter, S., &Floridi, L. (2016). The Ethics of Algorithms: Mapping the Debate. *Big Data & Society*, 3(2), 1-21.

Nissenbaum, H. (2010). *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press.

Ojomo, P. A. (2011). Environmental Ethics: An African Understanding. *African Journal of Environmental Science and Technology*, 5(8), 572-578. https://doi.org/10.5897/AJEST10.300

Pasquale, F. (2015). *The Black Box Society*. Harvard University Press.

Rousseau, J.-J. (2002). *The Social Contract* (G. D. H. Cole, Trans.). Dover Publications.

Russell, S., & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach*(4th ed.). Pearson.

Searle, J. R. (1980). Minds, Brains, and Programs. *Behavioural and Brain Sciences*, 3(3), 417-

424.

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind*, 59(236), 433-460.

Ugwulebo, O.C. (2026) *AI Ethics Versus Ethics of AI: Reclaiming Philosophy in Artificial Intelligence Advancement* [Manuscript submitted for Publication]. Department of Philosophy, Lagos State University, Nigeria.

737 Ugwulebo, O.C. (2026) *Cartesian Dualism and AI Adjudication: Avoiding Ghost In The*
738 *Machine Justice Through Human Expertise.*[Manuscript submitted for Publication].
739 Department of Philosophy, Lagos State University, Nigeria.
740 Winner, L. (1986). *The Whale and the Reactor*. University of Chicago Press.
741 Zuboff, S. (2019). *The Age of Surveillance Capitalism*. PublicAffairs.