

Can Machines Be Moral? : Gender and the Ethics of Intelligent Technologies

PRIYANKA DAS

Faculty

Department of Philosophy, Shailabala Women's Autonomous College, Cuttack (Odisha) India

ABSTRACT

The rapid advancement of intelligent technologies has renewed a central philosophical question: can machines be moral agents, or are they merely sophisticated instruments of human intention? This paper examines this issue through the combined perspectives of moral philosophy and gender ethics. It argues that discussions of machine morality cannot be separated from the social values and power relations embedded in technological design and use. Drawing on debates in AI ethics, feminist philosophy, and theories of moral agency, the study critically analyses how intelligent systems often reproduce gender biases through data selection, algorithmic modelling, and normative assumptions. Such biases challenge the claim of technological neutrality and raise serious ethical concerns regarding responsibility, accountability, and justice. The paper further explores whether attributing moral agency to machines risks obscuring human moral responsibility, particularly in contexts where automated decisions disproportionately affect women and gender minorities. By engaging with care ethics alongside deontological and utilitarian approaches, the paper proposes a relational understanding of moral responsibility in human-machine interactions. It emphasizes that moral evaluation should focus not only on outcomes, but also on the social contexts and power structures within which technologies operate. The paper concludes that the question is not simply whether machines can be moral, but under what ethical and gender-sensitive conditions intelligent technologies should be developed and governed in contemporary societies.

Keywords: Artificial Intelligence Ethics, Gender Bias, Moral Agency, Care Ethics, Technology and Society

INTRODUCTION

The accelerated development of Artificial Intelligence (AI) and intelligent systems has significantly transformed the normative landscape of contemporary societies. Algorithmic systems now participate in decision-making processes across domains such as healthcare, finance, criminal justice, education, and digital governance. These systems do not merely execute predefined instructions; they classify, predict, optimize, and recommend actions that carry substantial moral and social consequences. As intelligent technologies increasingly mediate human interactions and institutional

practices, the philosophical question of moral agency acquires renewed urgency: can machines be regarded as moral agents, or are they merely complex instruments embedded within human normative structures?

Within classical moral philosophy, moral agency has been grounded in capacities such as rational deliberation, autonomy, intentionality, and accountability. Deontological traditions emphasize self-legislation and moral autonomy as defining characteristics of agents capable of responsibility¹. Consequentialist frameworks presuppose evaluative reasoning and the capacity to assess outcomes, while virtue-ethical approaches locate moral agency in practical

1. Kant, I. (1993). *Grounding for the metaphysics of morals* (J. W. Ellington, Trans., 3rd ed., pp. 1–69). Hackett Publishing Company.

wisdom and cultivated character. Despite theoretical differences, these traditions converge in affirming that moral agency requires consciousness, purposive intentionality, and the ability to justify one's actions within a normative framework.

The emergence of AI systems challenges these foundational assumptions. Although machine learning models can simulate evaluative processes and adapt through data-driven optimization, they lack subjective experience and intrinsic intentional states. This raises a central problem: does functional participation in morally relevant decision-making suffice for moral agency, or does the absence of consciousness preclude such attribution? Moreover, attributing moral status to machines risks displacing responsibility from human designers, developers, and institutions that shape technological systems.

Importantly, this debate cannot be separated from the socio-cultural contexts within which intelligent technologies are produced and deployed. A growing body of scholarship demonstrates that AI systems frequently reproduce and amplify gender biases embedded in datasets, institutional structures, and technological design. The assumption of technological neutrality therefore requires critical scrutiny. Gender operates not as an external variable but as a structural category that shapes epistemic assumptions, normative priorities, and patterns of inclusion and exclusion within technological cultures.

This paper argues that the question of machine morality must be reframed through a gender-sensitive ethical lens. Moral evaluation of intelligent technologies is inseparable from the gendered social conditions under which they are constructed and operationalized. By integrating debates on moral agency with feminist ethical analysis, the study seeks to clarify the normative foundations of responsibility, accountability, and justice in the age of intelligent technologies.

Conceptual Framework

Moral Agency in Philosophy: Beyond Classical Individualism

The traditional concept of moral agency in Western philosophy has largely been constructed within a framework of rational individualism. Moral agents are typically defined as autonomous, self-legislating individuals capable of intentional action and accountable decision-making. This model, articulated most

prominently by Immanuel Kant, grounds morality in rational autonomy: the ability to legislate universal moral law for oneself. Similarly, consequentialist traditions associated with John Stuart Mill evaluate agency in terms of rational deliberation about outcomes. Virtue ethics, derived from Aristotle, emphasizes character and practical wisdom (phronesis), yet still presumes a reflective, deliberative subject.

Across these traditions, moral agency is characterized by three central conditions:

- a. Autonomy (self-governance and rational self-legislation)
- b. Intentionality (acting for reasons with conscious purpose)
- c. Responsibility (accountability for one's actions within a moral community)

The conceptual framework of this study begins with the philosophical foundations of moral agency, traditionally defined as the capacity for normatively evaluable action grounded in autonomy, intentionality, and responsibility. In classical moral philosophy, an agent is one who acts freely according to reasons, forms intentions, and can be held accountable for those actions. Kantian ethics locates moral agency in rational self-legislation and duty-bound autonomy; utilitarianism emphasizes evaluative reasoning oriented toward consequences and the maximization of well-being; and virtue ethics situates agency within character formation and practical wisdom cultivated through lived experience. Despite their differences, these traditions converge in treating moral agency as inherently tied to rational subjectivity and experiential understanding. Parallel to this is the debate in philosophy of technology concerning moral neutrality.

The instrumental view regards technology as a value-neutral tool whose ethical status depends entirely on human use. However, contemporary critiques argue that technologies—particularly artificial intelligence systems—are not neutral artefacts but embody normative assumptions through design choices, data structures, and institutional objectives. AI systems mediate social relations and can reproduce structural biases, thereby shaping moral outcomes rather than merely transmitting human intentions. This tension between human-centred moral theory and the non-neutral character of intelligent technologies frames the central philosophical challenge: whether moral agency can be extended to artificial systems or whether responsibility must remain

fundamentally anchored in human and institutional actors within technologically mediated environments.

The conceptual framework developed here establishes two central insights:

First, classical moral philosophy defines moral agency through autonomy, intentionality, and responsibility—conditions deeply tied to rational subjectivity and experiential awareness. Second, the assumption that technology is morally neutral is philosophically inadequate, particularly in the context of intelligent systems that mediate social outcomes.

Together, these insights create a tension at the heart of machine morality debates. On the one hand, AI systems increasingly perform morally significant functions. On the other hand, they lack the core conditions traditionally required for moral agency.

Resolving this tension requires either redefining moral agency in functional terms or reaffirming its human-centered foundations while developing robust frameworks of distributed responsibility. The subsequent analysis builds upon this conceptual groundwork to assess whether and in what sense machines can be considered moral participants within technologically mediated societies.

This framework is deeply anthropocentric and often implicitly gendered. Feminist philosophers have argued that the dominant image of the moral agent—as detached, rational, and independent—reflects historically masculine-coded ideals of agency while marginalizing relational and interdependent dimensions of moral life.

Gender Ethics and the Critique of Autonomous Moral Agency

Care Ethics and Relational Agency

The ethics of care, articulated by Carol Gilligan and further developed by Nel Noddings, challenges the dominance of abstract, rule-based moral reasoning. In a *Different Voice*, Gilligan argues that moral development theories had privileged justice-oriented reasoning while overlooking relational and care-centered moral perspectives often expressed by women. Care ethics emphasizes responsiveness, empathy, relational interdependence, and contextual judgment².

Under care ethics, moral agency is not primarily defined by detached rational autonomy but by the capacity to sustain relationships and respond to vulnerability. Noddings's conceptualizes morality as emerging from the ethical relation between the "one-caring" and the "cared-for," foregrounding attentiveness and responsibility within concrete contexts rather than universal abstraction³.

This reconceptualization significantly reshapes the debate about machine morality. If moral agency includes relational sensitivity, empathy, and responsiveness to lived vulnerability, then intelligent systems—lacking emotional awareness and experiential understanding—cannot fully embody moral agency in a care-ethical sense. Even advanced AI capable of sentiment analysis does not experience care; it processes signals without moral concern.

Gendered Assumptions in Technological Rationality

Feminist critiques extend beyond moral psychology to the structure of technological rationality itself. Modern technological systems often prioritize efficiency, optimization, control, and abstraction—values aligned with historically masculine-coded ideals of domination over nature and objectivity. Feminist theorists such as Donna Haraway argue that technology is embedded within networks of power and identity formation. Technologies do not merely execute neutral functions; they shape subjectivities and social hierarchies.

Similarly, Judith Butler demonstrates how norms are socially constructed and reproduced through institutional and discursive practices. When AI systems are trained on historically biased datasets, they reproduce gendered norms and exclusions. The claim of technological neutrality thus becomes philosophically untenable.

If classical moral agency presupposes autonomy and impartial rationality, care ethics reveals the incompleteness of this model. Moral life is not solely a matter of universal principles but also of situated relationships. Intelligent systems, however, are designed primarily within instrumental and efficiency-driven paradigms, often neglecting relational dimensions central to gender-sensitive ethics.

2. Gilligan, C. (1982). *In a different voice: Psychological theory and women's development*. Harvard University Press.

3. Noddings, N. (1984). *Caring: A Feminine Approach to Ethics and Moral Education*. University of California Press.

The Question of Machine Morality

The question of machine morality represents one of the most contested debates in contemporary philosophy of technology and AI ethics. As artificial intelligence systems increasingly participate in domains such as healthcare, law enforcement, finance, and warfare, their decisions carry significant moral consequences. The central issue, therefore, is not merely whether machines can perform ethical calculations, but whether they can properly be regarded as moral agents. This requires examining competing theoretical positions, the arguments supporting machine moral agency, and the major philosophical objections concerning consciousness, intentionality, and responsibility.

Artificial Moral Agents (AMAs): Theoretical Positions

The concept of Artificial Moral Agents (AMAs) refers to artificial systems designed to make decisions in morally significant contexts. Three primary theoretical positions structure the debate.

The weak AMA position maintains that machines can simulate ethical reasoning without possessing genuine moral agency. On this view, AI systems implement moral rules or optimization strategies programmed by humans. Their “moral behaviour” is derivative, not intrinsic. Responsibility remains entirely with designers, programmers, and institutions.

The strong AMA position argues that sufficiently advanced AI systems may qualify as moral agents if they demonstrate autonomy, adaptability, and independent decision-making. Proponents claim that when systems generate outcomes not directly foreseeable by programmers, they exhibit a functional form of agency. If moral agency is defined by performance rather than inner consciousness, machines could meet operational criteria.

A third perspective proposes a relational or distributed model of agency, suggesting that moral agency does not reside solely in individuals—human or machine—but emerges within sociotechnical networks. Under this framework, AI systems participate in moral processes alongside designers, users, institutions, and regulatory structures. Agency is thus distributed rather than isolated.

These positions reflect deeper disagreements about whether moral agency is essentially metaphysical (requiring consciousness and intention) or functional (requiring only decision-making competence).

Arguments in Favor of Machine Moral Agency

Supporters of machine moral agency advance several arguments.

First, the functionalist argument asserts that moral agency should be assessed by observable capabilities. If an AI system can identify morally relevant features, evaluate alternative actions, and generate ethically constrained decisions, then it fulfils the practical criteria of moral agency. Autonomous vehicles, for example, must weigh risks and minimize harm—tasks that resemble moral reasoning.

Second, the autonomy argument emphasizes the increasing independence of machine learning systems. Advanced AI systems adapt to new data and environments in ways not entirely predictable by their creators. Because they operate without direct real-time control, proponents argue that they display a form of practical autonomy.

Third, the normative necessity argument contends that when machines operate in morally sensitive contexts, embedding ethical reasoning capacities becomes essential. Even if machines are not moral subjects in a traditional sense, they must function as moral decision-makers to prevent harm. While these arguments highlight the growing moral significance of AI systems, they do not conclusively establish that machines possess moral agency in the full philosophical sense.

Objections: Consciousness, Intentionality, and Moral Understanding

The strongest objections to machine moral agency arise from classical philosophical criteria. The consciousness objection argues that moral agency presupposes subjective awareness. Moral responsibility entails the capacity to understand one’s actions as morally significant and to experience remorse or guilt. Current AI systems lack phenomenological consciousness; they do not experience their decisions. Without awareness, critics argue, genuine moral agency is impossible.

The intentionality objection focuses on purposiveness. Human moral action involves intentions grounded in beliefs and desires. Artificial systems operate through algorithmic processes without intrinsic mental states. Although they can model goals computationally, these are externally assigned objective functions rather than internally endorsed reasons.

The moral understanding objection maintains that moral agency requires normative comprehension. To act

morally is not merely to produce correct outputs but to understand why an action is right or wrong. AI systems process data syntactically rather than semantically; they manipulate symbols without grasping meaning. Thus, they simulate moral reasoning without genuine understanding.

Together, these objections suggest that artificial systems may be causally powerful but remain ontologically distinct from moral subjects.

The Problem of Responsibility and Accountability

Even if machines are denied full moral agency, their increasing role in decision-making creates what has been termed an accountability gap. When AI systems cause harm—through biased hiring algorithms, flawed medical diagnostics, or discriminatory predictive policing—it becomes difficult to identify who is responsible.

Several models of responsibility have been proposed. The designer model assigns responsibility to programmers and engineers. The operator model holds users accountable for deployment decisions. The institutional model locates responsibility in corporations and regulatory bodies that oversee AI systems. A more controversial approach suggests attributing responsibility directly to the machine.

However, assigning responsibility to machines raises serious concerns. Machines cannot meaningfully bear blame, offer justification, or undergo moral reform. Attributing responsibility to them risks diffusing or obscuring human accountability. Therefore, even if AI systems exhibit operational autonomy, moral responsibility remains fundamentally human and institutional. The accountability problem reveals that the debate about machine morality is not only theoretical but deeply practical. As AI systems shape social outcomes, clarifying responsibility becomes essential for justice, governance, and ethical legitimacy.

Gender and Technological Ethics

Technological ethics cannot be adequately theorized without engaging feminist critiques of rationality and power. Feminist philosophy has long argued that dominant models of technological rationality are shaped by historically masculine ideals of control, abstraction, efficiency, and domination over nature. This model privileges detached calculation while marginalizing

relational, contextual, and care-oriented forms of reasoning. When artificial intelligence systems are built upon such epistemic foundations, they inherit not neutrality but historically situated assumptions about whose knowledge counts and whose experiences are prioritized.

Feminist critiques challenge the claim that technology is value-neutral by demonstrating how gendered power relations become embedded within technical systems. AI does not emerge in a vacuum; it is developed within institutions that reflect structural inequalities. Data used to train intelligent systems often encode historical discrimination—whether in hiring patterns, credit allocation, healthcare access, or criminal justice decisions⁴. When algorithms learn from biased datasets, they do not simply mirror reality; they normalize and automate existing inequalities. Thus, bias in AI is not a technical glitch but a structural consequence of unequal social histories.

Gender bias operates at multiple levels: in data collection (who is represented and who is invisible), in design processes (whose perspectives shape system objectives), and in deployment contexts (who bears the risks of technological error). For example, facial recognition technologies have demonstrated higher error rates for women and darker-skinned individuals, revealing how supposedly objective systems reflect skewed training distributions. Such disparities expose the illusion of algorithmic neutrality.

More fundamentally, intelligent systems participate in broader power structures. They allocate resources, influence visibility, and shape opportunities. When AI systems determine creditworthiness, job eligibility, or risk assessment, they mediate access to social goods. If these systems reproduce gendered hierarchies, they reinforce marginalization under the guise of technical efficiency. The ethical issue is therefore not only bias correction but structural justice.

A gender-informed technological ethics shifts the focus from isolated algorithmic errors to systemic power relations. It demands inclusive design practices, epistemic diversity in development teams, and accountability mechanisms attentive to marginalized communities. Rather than asking merely whether machines can be moral, this perspective asks whose morality is encoded in machines—and who is excluded

4. Aarti and Chauhan, S. (2025). *Algorithmic bias and gender representation: feminist perspectives on AI-driven marketing*.

from its formulation⁵.

Technology and the Myth of Moral Neutrality ***Instrumental View of Technology***

The instrumental view treats technology as morally neutral—a passive tool whose ethical valence depends entirely on user intention. Under this model, responsibility remains exclusively human. AI systems are seen as computational instruments executing algorithms without intrinsic moral status.

Feminist Critique of Neutrality

Contemporary philosophy of technology challenges this neutrality thesis. Langdon Winner famously argued that artifacts can embody political properties. Technologies structure possibilities of action and distribute power asymmetrically. Feminist scholarship deepens this critique by demonstrating how technological systems embed gendered assumptions in design and deployment.

From a care-ethical perspective, neutrality is itself a moral stance—one that often obscures structural inequalities. When AI hiring systems disadvantage women or facial recognition systems misidentify gender minorities, the issue is not malfunction but embedded normative bias. The moral evaluation of AI must therefore address systemic relational harms, not merely rule compliance.

Reframing Machine Morality Through Gender Ethics

Integrating gender ethics into the framework of moral agency shifts the debate in three crucial ways:

- a. From Individual Autonomy to Relational Responsibility:** Moral agency is not solely self-legislation but responsiveness to others. Machines lack genuine relational awareness.
- b. From Abstract Neutrality to Contextual Embeddedness:** Technologies are socially situated and shaped by gendered power structures.
- c. From Rule-Following to Ethical Attentiveness:** Care ethics emphasizes moral perception and

sensitivity—capacities absent in current AI architectures.

Thus, the question “Can machines be moral?” cannot be resolved merely by examining computational sophistication. It requires interrogating the normative frameworks within which technologies are designed and the gendered assumptions that shape their operation. Machines may simulate ethical reasoning within predefined parameters, but moral agency—understood through both classical and care-ethical traditions—remains deeply rooted in human relational life.

Contemporary Applications and Philosophical Implications

Contemporary AI systems demonstrate that the question of machine morality is not abstract but socially consequential. In domains such as hiring, credit scoring, predictive policing, and healthcare triage, algorithmic decision-making has produced gendered effects that reinforce existing inequalities⁶. Automated employment systems have disadvantaged women through biased training data; medical algorithms have underestimated risks in female patients when based on male-centric clinical datasets; and surveillance technologies disproportionately monitor and classify marginalized bodies. These applications reveal that intelligent systems mediate access to opportunities, resources, and recognition, thereby shaping moral outcomes in ways that extend beyond individual intention.

Such developments compel a rethinking of moral responsibility within socio-technical systems⁷. If AI decisions emerge from complex interactions among data infrastructures, designers, institutions, and users, responsibility cannot be reduced to a single agent. Yet attributing morality to machines themselves risks obscuring accountability. Artificial systems may perform evaluative functions, but they lack consciousness, intentionality, and the capacity for moral answerability. Therefore, the limits of machine morality must be acknowledged: machines can influence moral landscapes but cannot bear moral responsibility. Ethical accountability remains fundamentally human and institutional, even within technologically mediated

5. Wajcman, J. and Young, E. (2023). *Feminism confronts AI: The gender relations of digitalisation*.

6. Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in *Commercial Gender classification*. *Proceedings of Machine Learning Research*, **81** : 1–15.

7. Floridi, L. and Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines*, **14**(3) : 349–379.

environments. The philosophical task is thus not to moralize machines, but to ensure that technological systems are governed in ways that preserve human responsibility, justice, and democratic oversight.

Conclusion

This paper has argued that the question “Can machines be moral?” cannot be answered in isolation from the philosophical foundations of moral agency and the socio-political conditions under which intelligent technologies are developed and deployed. Classical moral philosophy grounds agency in autonomy, intentionality, and responsibility—conditions deeply tied to conscious subjectivity and normative understanding. While contemporary AI systems increasingly perform evaluative and decision-making functions in morally consequential domains, functional competence does not suffice for moral subjecthood. Machines may simulate ethical reasoning, but they neither comprehend moral norms nor bear accountability for their outcomes.

At the same time, the instrumental view of technology as value-neutral is untenable. Intelligent systems embody normative assumptions embedded in data, design architectures, and institutional priorities.

Feminist critiques demonstrate that algorithmic systems often reproduce gendered hierarchies, not because machines possess bias in themselves, but because they operationalize historically structured inequalities. Thus, the moral significance of AI lies less in whether machines qualify as agents and more in how they mediate power, visibility, and access within social systems.

The central thesis of this paper is therefore twofold: first, artificial systems cannot be moral agents in the full philosophical sense because they lack consciousness, intentionality, and answerability; second, their growing role in shaping moral outcomes demands a reconfiguration of human and institutional responsibility within socio-technical networks. Rather than attributing morality to machines, ethical inquiry must focus on governance structures, epistemic inclusion, and justice-oriented design practices that prevent the diffusion of accountability. Ultimately, the future of AI ethics depends not on granting machines moral status, but on reaffirming human moral responsibility in an age of technological mediation. The challenge is not to humanize machines, but to ensure that technological systems remain accountable to human dignity, equality, and democratic oversight.
