European Journal of Philosophical Research. 2024. 11(1)
E-ISSN: 2413-7286
Volume-11/Issue-1/2024

Ethical Responsibility in Artificial Intelligence: A Philosophical and Legal Inquiry

Prof. Elena Rossi
Department of Philosophy and Legal Studies, University of Milan, Italy

Abstract
Artificial Intelligence (AI) has transitioned from a primarily technical field into a deeply social and ethical force shaping modern life. As AI systems are increasingly deployed in sensitive contexts—healthcare, finance, criminal justice, autonomous vehicles, and military applications—the question of ethical responsibility becomes both urgent and complex. This paper explores the philosophical foundations of responsibility, traces how traditional conceptions of moral agency apply (or fail to apply) to AI systems, and analyzes contemporary legal frameworks and proposals for governing AI accountability. We argue that existing philosophical models of responsibility must be revised to account for distributed agency, opaque decision mechanisms, and autonomous action. Furthermore, legal systems must move beyond metaphors of human-centric culpability toward institutional, hybrid, and multi-layered forms of accountability. Ultimately, this inquiry demonstrates that reconciling ethical theory with legal practice in AI governance requires a synthesis of moral, technical, and socio-legal analyses.
Keywords: AI Ethics, Moral Responsibility, Technology, Law

1. Introduction

Artificial Intelligence (AI) systems are fundamentally reshaping human environments. Autonomous decision-making systems are now embedded in daily life—curating news feeds, adjudicating creditworthiness, and optimizing logistical networks—and are poised to assume even greater roles in healthcare diagnostics, autonomous driving, public safety, and warfare. While these technologies promise efficiency, scale, and innovation, they simultaneously present unprecedented ethical challenges.

A central concern is *ethical responsibility*: identifying who (or what) is morally and legally accountable when AI systems act in ways that harm, discriminate, or otherwise produce negative consequences. Traditional frameworks of responsibility originate within moral philosophy and jurisprudence, both of which presuppose intentional agency, foresight, and personhood. AI systems, however, complicate these assumptions: they can act autonomously, produce outcomes unpredictable even to their creators, and lack clearly discernible intentional states. This paper navigates these conceptual and practical tensions.

We first explicate philosophical accounts of moral responsibility (§2), assess the applicability of those accounts to AI (§3), explore contemporary legal frameworks and proposals (§4), and conclude with recommendations for an integrated ethical and legal approach (§5).

2. Philosophical Foundations of Responsibility

Understanding ethical responsibility begins with clarifying its meaning within moral philosophy.

2.1. Traditional Theories of Moral Responsibility

Moral responsibility has long been tied to the notion of free will and agency. According to canonical views in moral philosophy:

- **Agency** implies the capacity to act intentionally, with understanding of one's decisions and their foreseeable consequences.
- **Moral responsibility** implies that an agent can be *praised*, *blamed*, *held accountable*, or *subject to sanctions* for their actions.

Prominent accounts, such as those of P.F. Strawson, assert that responsibility depends on reactive attitudes—emotional responses like resentment or gratitude—that presuppose intentional agency (Strawson 1962). Consequentialist theories, in contrast, justify responsibility in terms of its effects on future behavior rather than desert (e.g., punishment to deter harm) (Smart & Williams 1973).

2.2. Conditions for Moral Responsibility

Most philosophical accounts agree that moral responsibility requires:

1. **Control or Agency** – The agent must have causal influence over their action.
2. **Understanding or Awareness** – The agent must grasp (to some degree) the nature of their action and its morally relevant effects.
3. **Freedom from Coercion** – The agent's choices cannot be entirely determined by external forces.

Under these conditions, human actors are accountable for decisions they intentionally make. But AI, as non-human entities, complicate these criteria.

2.3. Distributed and Collective Responsibility

Modern systems often involve *distributed agency*: the result of many designers, engineers, users, and institutions. Philosophers like Hannah Arendt and more recently Collective Responsibility theorists have argued that responsibility can be distributed across groups, not merely located in a single agent (Arendt 1963; French 1984). Such models might better reflect how AI systems emerge through complex socio-technical processes.

3. AI Systems and the Challenge of Ethical Responsibility

AI systems challenge the philosophical frameworks of responsibility in several distinct ways.

3.1. Autonomy and Opacity

Many AI systems—especially those based on machine learning—operate with degrees of autonomy and internal opacity. They can identify patterns and make decisions that even their creators cannot fully explain. This raises two key issues:

1. **Unpredictability** – If neither users nor designers can foresee specific outcomes, responsibility becomes legally and ethically fraught.
2. **Lack of Intentionality** – AI systems do not possess beliefs, desires, or intentions in the human sense, undermining the applicability of intentional-based responsibility models.

3.2. Responsibility Attribution

When an AI system causes harm, who bears responsibility?
Possible candidates include:

- **Developers and Programmers** – Responsible for the architecture, training data, and algorithms.
- **Deployers or Organizations** – Entities that choose to integrate AI into decision-making.
- **Manufacturers** – Suppliers of AI hardware or software.
- **Regulators or Governments** – Charged with legal oversight and standards.

Each candidate presents distinct ethical and legal complications. For instance, assigning responsibility solely to developers ignores the role of deployment contexts and end-users.

3.3. Moral Patiency vs. Moral Agency

Some scholars distinguish between *moral agency* (being capable of moral action) and *moral patiency* (being subject to moral consideration). While AI systems lack moral agency, they may nevertheless have moral *effects*—e.g., causing discriminatory outcomes—which require ethical response. The ethical focus then shifts from the AI itself to the *human actors* and structures surrounding its design, deployment, and use (Coeckelbergh 2020).

3.4. Case Studies in Ethical Ambiguity

Consider several high-profile examples:

- **Autonomous Vehicles:** When a self-driving car causes a fatality, should responsibility reside with the manufacturer who designed the control software, the owner who delegated control, or regulators who approved deployment?
- **Predictive Policing Systems:** If algorithmic bias leads to disproportionate targeting of minority communities, who bears ethical and legal responsibility—developers, law enforcement agencies, or policymakers who funded the system?
- **Healthcare Diagnostics:** When machine learning systems provide inaccurate medical recommendations, patients can suffer. Traditional malpractice concepts become difficult to apply when physicians rely on opaque AI outputs.

These cases illustrate that responsibility cannot be traced neatly to a single human or entity under conventional frameworks.

4. Legal Frameworks and Accountability for AI

Legal systems have traditionally grounded liability in human actions. AI complicates this regime.

4.1. Tort Law and Strict Liability

In tort law, liability can arise from negligence or strict liability. Negligence requires a failure to meet a duty of care, while strict liability does not require fault but arises for inherently dangerous activities.

AI systems, especially autonomous ones, provoke debate over whether harm should trigger:

- **Negligence Liability:** Holding developers, suppliers, or deployers responsible for failing to exercise due care.
- **Product Liability:** Treating AI systems as products, subject to defect theories that hold manufacturers accountable if a product is unreasonably dangerous.
- **Strict Liability Regimes:** Particularly for autonomous vehicles, where harm may be attributed to simply operating such systems.

The European Union's Product Liability Directive and national laws increasingly consider AI within existing product liability frameworks, although gaps remain concerning adaptability to self-learning systems (European Commission 2020).

4.2. Regulatory Approaches: Governance and Standards

Rather than focusing solely on ex post liability, regulatory frameworks seek to govern AI behavior in advance through norms and standards.

Examples include:

- **Algorithmic Impact Assessments** – Requiring impact evaluations for systems with significant ethical risks.
- **Mandatory Transparency Requirements** – Obliging developers to disclose training data sources, algorithmic decision logic, and performance benchmarks.
- **Certification Regimes and Technical Standards** – Establishment of oversight bodies to certify AI safety and ethical compliance.

The European Union's proposed *Artificial Intelligence Act* (AIA) uses a risk-based approach, categorizing AI systems and imposing proportionate safeguards—a departure from purely reactive liability models (European Parliament, 2021).

4.3. Human Oversight and Meaningful Human Control

One legal strategy is to require human oversight in AI deployment—especially in high-risk applications such as healthcare or air traffic control. The idea of *Meaningful Human Control* emphasizes that final decisions must remain with accountable human agents, even if assisted by AI. While this does not eliminate AI's role, it ensures a legally recognizable locus for responsibility.

4.4. Corporate and Organizational Liability

Since many AI systems are developed and deployed by corporations, legal accountability increasingly considers corporate structures:

- **Corporate Liability:** Imposing responsibility on corporations for harms caused by AI systems they own or deploy.
- **Vicarious Liability:** Holding employers responsible for actions of agents (including software agents) operating under their control.

This model resonates with approaches used in other domains, such as environmental or workplace safety regulation, where organizations, not individual employees, are held accountable for systemic harms (Calo 2015).

5. Philosophical and Legal Synthesis: Toward Responsible AI Governance

Given the limitations of both traditional philosophical responsibility and existing legal frameworks, an integrated model is necessary.

5.1. Multi-Layered Accountability Structures

Rather than searching for a *single responsible agent*, accountability should be *distributed*:

1. **Design Level:** Developers and architects should be accountable for building systems with ethical constraints, transparency, and bias mitigation.
2. **Deployment Level:** Organizations must evaluate context-specific risks, incorporate human oversight, and monitor system performance.

3. **Regulatory Level:** Governments and supranational bodies should establish standards, compliance mechanisms, and enforcement capabilities.
4. **Public Participation:** Civil society and users must have access to oversight processes, complaint mechanisms, and redress options.

This layered accountability acknowledges the socio-technical complexity of AI systems.

5.2. Ethical Frameworks for AI Decision Making

Philosophers and ethicists have proposed frameworks to guide AI ethics:

- **Principle of Beneficence and Non-Maleficence:** Maximizing positive outcomes and minimizing harm.
- **Principle of Justice:** Ensuring fair distribution of benefits and burdens, preventing discrimination.
- **Principle of Autonomy:** Respecting human agency and consent.
- **Principle of Explainability and Transparency:** Ensuring that decisions can be traced and justified.

These principles align with existing bioethical models (Beauchamp & Childress 2001) and can extend to AI governance.

5.3. Revising Responsibility for AI's Unique Ontology

Some philosophers argue for *new models of moral responsibility* that accommodate non-human actors. For example:

- **Collective Responsibility:** Treating groups of designers, organizations, and users as co-responsible agents.
- **Extended Responsibility:** Assigning responsibility to systems insofar as they embody human values and choices.
- **Functional Responsibility:** Focused less on intentionality and more on *capability to prevent harm* (Floridi & Sanders 2004).

These models preserve ethical accountability without requiring AI to be moral agents in the traditional sense.

5.4. International Legal Harmonization

AI systems operate globally, with development and use crossing national boundaries. Harmonizing legal standards—similar to environmental treaties or human rights conventions—can prevent regulatory arbitrage and ensure a baseline of ethical accountability.

The *UN Guiding Principles on Business and Human Rights* and UNESCO's *Recommendation on the Ethics of AI* signal emerging international consensus on ethical AI governance, though enforcement mechanisms are still nascent.

6. Conclusion

AI's transformative power demands a reevaluation of ethical and legal responsibility. Traditional models of moral responsibility—grounded in individual intentional agency—are insufficient for technologies with autonomous behavior, opaque reasoning, and distributed development contexts. Assigning responsibility for AI is not merely a legal technicality; it reflects deeper societal values about justice, autonomy, human dignity, and control.

This paper has argued that:

- AI's opacity and autonomy challenge old paradigms of moral agency and liability.

- Legal systems must adapt toward multi-layered accountability structures that span design, deployment, and regulation.
- Philosophical models of responsibility should incorporate collective, extended, and functional frameworks that respect AI's socio-technical embeddedness.
- Effective governance requires harmonized regulatory standards, transparency requirements, and mechanisms for meaningful human oversight.

Ultimately, ethical responsibility in AI cannot be resolved solely within philosophy or law; it demands an interdisciplinary dialogue among ethicists, legal scholars, engineers, policymakers, and affected communities. Only through integrated inquiry can we ensure that AI technologies advance human welfare without compromising accountability, fairness, or justice.

References

Books and Articles

Arendt, Hannah. *Eichmann in Jerusalem: A Report on the Banality of Evil*. Viking Press, 1963.

Beauchamp, T.L., & Childress, J.F. *Principles of Biomedical Ethics*. Oxford University Press, 2001.

Calo, Ryan. "Robotics and the Lessons of Cyberlaw." *California Law Review* 103.3 (2015): 513–563.

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

European Commission. *White Paper on Artificial Intelligence: A European Approach to Excellence and Trust*, 2020.

European Parliament and Council. *Proposal for a Regulation Laying Down Harmonized Rules on Artificial Intelligence (Artificial Intelligence Act)*, 2021.

Floridi, Luciano & Sanders, Jeff. *On the Morality of Artificial Agents*. *Minds and Machines* 14.3 (2004): 349–379.

French, Peter A. "The Corporation as a Moral Person." *American Philosophical Quarterly* 21.3 (1984): 207–215.

Smart, J.J.C., & Williams, B. *Utilitarianism: For and Against*. Cambridge University Press, 1973.

Strawson, P.F. "Freedom and Resentment." *Proceedings of the British Academy* 48 (1962): 1–25.

Legal and Policy Reports

UNESCO. *Recommendation on the Ethics of Artificial Intelligence*, 2021.

UN Human Rights Council. *Report of the Special Rapporteur on the Rights to Freedom of Peaceful Assembly and of Association*, 2020.