The Alignment Discourse and the Locus of Responsibility

Edervaldo José de Souza Melo
Independent researcher.
Email: edersouzamelo@gmail.com
ORCID: 0009-0003-6835-135X

Abstract

Contemporary discussions of AI alignment frequently employ normative language that attributes to technical systems properties commonly associated with moral agency, such as values, intentions, or goals. This paper argues that such usage, in some cases, involves a misattribution of moral agency and a corresponding mislocation of responsibility. By treating systems as the primary bearers of normative obligations, parts of the alignment discourse risk obscuring the human and institutional responsibility involved in the design, deployment, and use of these artifacts. The paper offers a strictly conceptual clarification: it outlines minimal criteria for moral agency, distinguishes instruments, systems, and agents, and examines how a shift from functional to moral vocabulary contributes to a displacement of responsibility. The central claim is modest: clarifying the locus of responsibility improves the coherence of ethical discussions surrounding accountability, governance, and responsible use of AI systems.

Keywords: AI alignment; moral agency; responsibility; ethics of technology; normativity.

1 Introduction

The concept of alignment has become a focal point in contemporary debates on the ethics of artificial intelligence. Broadly understood, alignment concerns the requirement that AI systems operate in accordance with human values, socially acceptable goals, or relevant normative constraints. This concern is well motivated, given the increasing integration of automated systems into domains involving high-stakes decision-making, mediation, and coordination.

Alongside this development, however, a distinctive feature has emerged in alignment-related discourse: the application of moral vocabulary directly to technical systems. Systems are described as having goals, values, or intentions, and their outputs are evaluated in terms of moral success or failure. While such language often serves heuristic or operational purposes, its uncritical extension into normative contexts can generate conceptual confusion.

It is important to note that this paper is not concerned with alignment as a technical research program, nor with specific methods for value learning, constraint satisfaction, or control. Instead, it focuses on what may be called the alignment discourse: a heterogeneous body of normative claims, policy-oriented arguments, and public-facing interpretations that draw on alignment terminology. This discourse operates at the intersection of technical description and ethical evaluation, and it is precisely at this interface that conceptual slippage is most likely to occur.

This paper addresses a specific instance of that confusion. It examines how, in parts of the alignment discourse, the attribution of normative properties to systems contributes to a displacement of the locus of moral responsibility. The aim is neither to reject alignment as a research agenda nor to advance an alternative framework. Rather, the paper offers a limited conceptual clarification: it seeks to distinguish instruments, systems, and agents, and to show how conflating these categories can obscure where moral responsibility properly resides.

2 Agency and Moral Responsibility: Minimal Criteria

Philosophical accounts of moral agency vary widely in their metaphysical and normative commitments. For the purposes of this paper, however, it is unnecessary to adopt a robust or controversial theory of agency. Instead, a minimal set of criteria suffices to distinguish entities that can plausibly be considered moral agents from those that cannot.

At a minimum, moral agency presupposes the capacity to grasp normative reasons, to deliberate in light of such reasons, and to be held accountable for actions as one's own. These conditions do not require strong metaphysical assumptions about free will or consciousness, but they do entail imputability: the possibility of justification, explanation, or answerability in normative terms. Classic philosophical analyses of intention and action emphasize precisely this link between agency and reason-responsiveness, rather than mere causal efficacy or functional complexity (Anscombe, 1957; Bennett, 1976).

This distinction is particularly relevant in the context of contemporary AI systems, which are often described as optimizing objectives or responding to reward structures. While such systems may exhibit sensitivity to formally specified criteria, this sensitivity should not be confused with responsiveness to normative reasons. Optimization with respect to a loss function or reward signal does not amount to deliberation about what ought to be done, nor does it ground imputability. The presence of goal-directed behavior, therefore, is insufficient for moral agency absent the capacity to engage with reasons as reasons.

On this basis, it is useful to distinguish among three conceptual categories:

1. **Instruments**, which function as means to externally specified ends and lack normative autonomy.
2. **Systems**, which consist of organized components capable of complex, adaptive, or self-regulating behavior, while remaining governed by externally defined objectives and constraints.
3. **Agents**, which are capable of understanding norms as norms, responding to reasons, and being held morally accountable for their actions.

The distinction between instruments and systems is technically significant, but it does not by itself confer moral status. A system may be highly complex or opaque without satisfying even minimal criteria for moral agency. Functional sophistication is not equivalent to normative imputability.

3 Alignment Talk and the Misplacement of Normativity

In alignment-related discussions, functional descriptions of AI systems are frequently framed in intentional terms. Systems are said to pursue goals, learn values, or select among alternatives. Within technical and engineering contexts, such vocabulary often serves a pragmatic role, facilitating abstraction and interdisciplinary communication. As several authors have noted, the use of intentional language in descriptions of complex artifacts does not, by itself, commit one to literal attributions of agency or moral status (Bennett, 1976; Boden, 2016; Suchman, 2007).

Difficulties arise when this functional vocabulary supports normative inferences. Claims that a system “ought” to respect certain values or that it has failed in a morally relevant sense suggest, implicitly, that the system itself is the bearer of the obligation or failure. In such cases, a gradual shift occurs from operational description to moral attribution, often without explicit theoretical justification. This shift reflects not a deliberate philosophical stance, but an erosion of conceptual boundaries between description and normativity.

What is at stake in this process is a form of semantic drift, in which terms initially introduced for functional or heuristic purposes gradually acquire normative force. Over time, expressions such as “system goals” or “aligned behavior” cease to function merely as shorthand for design specifications and begin to support claims about moral adequacy or failure. When this shift goes unexamined, the language of alignment no longer merely describes system behavior but implicitly assigns normative standing to the artifact itself.

This shift does not necessarily reflect an explicit theoretical commitment. More often, it emerges gradually through the repeated use of intentional metaphors in settings where conceptual distinctions are not carefully maintained. Nonetheless, the effect is significant: normativity, which ordinarily applies to agents capable of responsibility, is projected onto technical artifacts.

4 Misattribution of Agency and the Responsibility Shift

The central claim of this paper is that AI systems, considered as systems, do not satisfy minimal criteria for moral agency. While they may exhibit complex and adaptive behavior, they do not understand norms as norms, deliberate on the basis of moral reasons, or answer for their actions in justificatory terms. As Johnson argues, computer systems may be treated as moral entities in a derivative or instrumental sense, but they lack the status of moral agents capable of responsibility in their own right (Johnson, 2006).

When alignment discourse nonetheless treats the system as the primary locus of normative assessment, a shift in responsibility occurs. Moral attention moves away from those who design, deploy, and govern the system and toward the artifact itself. Long-standing discussions of accountability in socio-technical systems have emphasized that such shifts risk obscuring human and institutional responsibility, particularly in contexts where decision-making authority and control remain externally located (Nissenbaum, 1996; Stahl, 2012).

This argument should not be understood as denying the practical importance of modifying or constraining systems in response to harmful outcomes. Technical corrections, safeguards, and design changes may be entirely appropriate as prudential measures. The conceptual issue arises when such measures are framed as responses to a moral failure of the system itself, rather than as obligations incumbent upon those responsible for its design and use. Conflating these two levels—prudential system correction and moral responsibility—reinforces the displacement of normativity that this paper seeks to clarify.

This shift becomes particularly visible in public reactions to illicit or harmful uses of AI systems. In such contexts, the occurrence of a wrongful act mediated by a technical artifact is often interpreted as evidence of a normative defect in the system, described in terms of misalignment. Ethical response then concentrates on correcting the artifact—its internal mechanisms, constraints, or filters—while the responsibility of the human agent who intentionally employed the system for illicit purposes remains secondary. The result is an implicit reconfiguration of the locus of normativity, in which the tool assumes the role of primary moral object.

From a conceptual standpoint, this amounts to a confusion of levels—akin to what classical philosophical discussions would describe as a category error. More importantly, it leads to a practical consequence: the dilution of human and institutional accountability in contexts where responsibility ought to be clearly assigned.

It is important to distinguish moral responsibility from corrective or prudential intervention. Modifying, constraining, or redesigning a system in response to harmful outcomes may be entirely appropriate as a matter of risk management or institutional duty. Such interventions, however, do not by themselves imply that the system is the bearer of moral fault. Confusing corrective action with moral attribution risks treating technical modification as a substitute for normative accountability, thereby obscuring the responsibilities of those agents who design, deploy, and authorize the use of the system.

The normative significance of this distinction lies in the effects of responsibility displacement. When moral evaluation is redirected toward artifacts, the agents and institutions capable of intention, deliberation, and justification recede from view. Over time, this redirection weakens the normative force of accountability practices by diffusing responsibility across technical systems that cannot meaningfully respond to moral demands. Preserving the locus of responsibility at the human and institutional level is therefore not merely a conceptual preference, but a condition for maintaining the integrity of ethical evaluation in socio-technical contexts.

5 Implications for AI Ethics

The implications of this clarification are limited but meaningful. A clearer account of the locus of responsibility supports more coherent discussions of accountability, governance, and responsible use. Rather than asking whether a system itself is morally aligned, ethical evaluation can focus on whether relevant agents and institutions have fulfilled their normative obligations in design, deployment, and oversight. Contemporary work in AI ethics has repeatedly emphasized the importance of preserving this distinction in order to avoid misplaced moral expectations of technical systems

(Floridi, 2013; Floridi et al., 2018).

From this perspective, debates about accountability benefit from a clearer separation between technical evaluation and moral attribution. Systems may be assessed for reliability, safety, or compliance with design specifications, while responsibility for their use remains firmly located at the human and institutional level. Preserving this distinction helps avoid situations in which moral expectations are redirected toward artifacts, thereby weakening the normative force of accountability mechanisms directed at actual decision-makers.

Second, conceptual clarity improves debates on governance and policy. Norms and regulatory mechanisms are more effective when they target appropriate subjects of responsibility, rather than attributing moral standing to artifacts.

Finally, this argument does not undermine the operational usefulness of alignment as a technical concept. Alignment can remain a valuable research goal, provided it is not conflated with the attribution of moral agency to systems themselves.

6 Conclusion

This paper has examined a recurrent conceptual confusion in parts of contemporary AI alignment discourse: the misattribution of moral agency to technical systems and the resulting displacement of responsibility. By appealing to minimal criteria of moral agency and distinguishing instruments, systems, and agents, it has argued that AI systems are not appropriate bearers of normative obligations.

Moral responsibility remains primarily human and institutional. Clarifying this point does not resolve all ethical challenges posed by AI, but it contributes to a more precise and coherent debate—one in which obligations, failures, and corrective duties are located where they can meaningfully be assumed. In the ethics of technology, conceptual clarity is not an abstract virtue, but a precondition for responsibility.

References

Anscombe, G. E. M. (1957). *Intention.* Oxford: Basil Blackwell.

Bennett, J. (1976). *Linguistic Behaviour.* Cambridge: Cambridge University Press.

Boden, M. A. (2016). *AI: Its Nature and Future.* Oxford: Oxford University Press.

Floridi, L. (2013). *The Ethics of Information.* Oxford: Oxford University Press.

Floridi, L., J. Cowls, M. Beltrametti, et al. (2018). Ai4people—an ethical framework for a good ai society. *Minds and Machines 28*(4), 689–707.

Johnson, D. G. (2006). Computer systems: Moral entities but not moral agents. *Ethics and Information Technology 8*(4), 195–204.

Nissenbaum, H. (1996). Accountability in a computerized society. *Science and Engineering Ethics 2*(1), 25–42.

Stahl, B. C. (2012). Moral responsibility for it innovation. *Journal of Information, Communication and Ethics in Society 10*(3), 117–133.

Suchman, L. (2007). *Human–Machine Reconfigurations: Plans and Situated Actions.* Cambridge: Cambridge University Press.