Trustworthy AI and Mechanical Expressivism: The Systemic Risk of Ontological Hermeneutical Injustice

Marios Thomas Dogkas[1]

[1]National and Kapodistrian University of Athens, Greece

Abstract

This paper fundamentally challenges current approaches to "Trustworthy AI," arguing that superficial technical de-biasing cannot solve the ethical alignment problem. As humans and generative algorithms increasingly share cognitive labor to form dynamic "Hybrid Epistemic Agents," the text introduces two radical philosophical concepts. First, "Mechanical Expressivism" posits that Large Language Models (LLMs) do not possess moral beliefs or track objective ethical truths. Instead, they function as automated statistical mirrors, mathematically expressing the compressed, probabilistic average of human normative attitudes found in their massive training datasets. Second, this dynamic leads to an "Ontological Hermeneutical Injustice". Because generative AI mathematically optimizes for majoritarian linguistic regularities, it systematically treats minority perspectives, low-frequency patterns, and complex non-normative interpretations as statistical "noise". This mechanically filters out diverse viewpoints, creating a structural erosion of our collective interpretive resources and locking users into conceptual conformism. Ultimately, relying on statistical models for moral reasoning actively homogenizes human thought. To achieve truly human-centered AI, we must shift our focus toward "Epistemic Sustainability"—a commitment to actively safeguarding human interpretive diversity outside the majoritarian loops of statistical calculation.

1. Introduction

The rapid integration of Large Language Models (LLMs) into daily cognitive and decision-making routines marks a profound shift in human-computer interaction. We have transitioned from utilizing passive cognitive artifacts—such as static databases or standard calculators—to forming dynamic "Hybrid Epistemic Agents," wherein cognitive labor is continuously and fluidly shared between human users and generative algorithms. As these models increasingly process and output normative judgments, the challenge of AI alignment has taken center stage in the pursuit of safe and reliable technology. However, current frameworks for "Trustworthy AI" frequently attempt to solve ethical alignment purely through technical risk management and algorithmic de-biasing. In doing so, they overlook a foundational philosophical puzzle: What is the exact ontological and metaethical nature of the moral language generated by a machine?

When an LLM outputs a normative claim, it does not do so by tracking an objective moral reality, nor does it possess the semantic grounding required for genuine ethical deliberation. This paper argues that to accurately assess the societal implications and ethical requirements of generative models, we must conceptualize their linguistic behavior through a novel metaethical framework: Mechanical Expressivism. Drawing upon classical non-cognitivist theories, Mechanical Expressivism posits that LLMs function as automated statistical mirrors; they do not hold moral beliefs, but rather mathematically express the compressed, probabilistic average of the normative attitudes embedded within their vast training corpora.

Recognizing LLMs as mechanical expressivists exposes a critical systemic vulnerability at the heart of Trustworthy AI. When human agents increasingly rely on these statistical outputs for interpretive and moral reasoning, a structural failure of trust emerges, which this paper identifies as an Ontological Hermeneutical Injustice. Because the underlying mathematics of generative AI systematically favor linguistic regularities and majoritarian viewpoints, they mechanically filter out statistical anomalies, minority perspectives, and complex non-normative interpretations. We demonstrate that this phenomenon

TRUST-AI 2026 workshop at IJCAI-ECAI 2026, August 15-16, 2026

dogkasm29l@gmail.com (M. T. Dogkas)

is not merely a resolvable technical bias, but an ontological symptom of statistical meaning-making that inherently erodes our shared interpretive resources. Ultimately, achieving truly human-centered and trustworthy AI requires moving beyond superficial technical fixes to actively safeguard interpretive diversity against the systemic risks of statistical normativity.

2. The Metaethical Ontology of Generative AI

2.1. The Critique of Artificial Moral Realism

To establish a philosophically rigorous framework for Trustworthy AI, it is first necessary to dismantle the intuitive but flawed assumption that Large Language Models (LLMs) can track, process, or output objective moral truths. When generative AI produces normative statements—such as declaring an action "right" or "wrong"—it visually mimics human ethical reasoning. However, evaluating AI outputs through the lens of traditional moral realism reveals a profound ontological deficit in the nature of machine-generated moral language.

According to Richard Boyd's formulation of moral realism, moral statements are not merely expressions of attitude, but propositions that are either true or false. Crucially, the truth or falsity of these moral statements is "largely independent of our moral opinions, theories, etc." (Boyd, 1988, p. 307). For a moral realist, genuine ethical judgments succeed in referring to objective, mind-independent moral properties. This referential success requires a causal, truth-tracking relationship between the epistemic agent and the real-world moral phenomena they are describing. If an agent—whether human or artificial—is to make a realist moral judgment, their cognitive mechanisms must be connected to, and regulated by, the objective moral facts of the world (Boyd, 1988, pp. 321, 336).

Applying this realist criterion to LLMs commits a fundamental category mistake regarding the nature of artificial cognition. To understand why, we must analyze the AI system at the correct Epistemic Level of Abstraction (LoA), a methodological framework developed by Luciano Floridi. A Level of Abstraction consists of a specific set of observables through which a system is analyzed, preventing the conflation of different ontological properties (Floridi, 2011, p. 52). If we analyze an LLM at a superficial behavioral LoA—focusing merely on its linguistic output—the system appears to possess moral agency and semantic understanding. However, when analyzed at its fundamental architectural LoA, an LLM is revealed to be a system of pure syntactic manipulation and statistical compression, operating entirely without semantic grounding.

This architectural reality aligns directly with John Searle's famous critique of strong Artificial Intelligence. Through his "Chinese Room" thought experiment, Searle demonstrates that formal symbol manipulation (syntax) is fundamentally distinct from, and insufficient for, genuine understanding (semantics) (Searle, 1980, pp. 417, 422). An artificial program operates strictly by applying computational rules to uninterpreted symbols. As Searle argues, "the program is purely formal, but the intentional states are not in that way formal. They are defined in terms of their content, not their form" (Searle, 1980, p. 423). Because an LLM lacks intentionality—a subjective, conscious connection to the world—it cannot understand the semantic content of the moral words it processes. It manipulates the symbol "justice" based solely on its formal, statistical relationship to other symbols, not through any experiential or causal link to actual justice in the lived world.

Bender et al. (2021) classify LLMs as "stochastic parrots." They argue that despite their impressive fluency, these models are essentially haphazardly stitching together sequences of linguistic forms observed in their vast training data "according to probabilistic information about how they combine, but without any reference to meaning" (Bender et al., 2021, pp. 616–617). The model does not learn facts about the moral world; it only learns the probabilities of character sequences as they appear in human-generated text. Consequently, language models are inherently disconnected from the communicative intent that characterizes actual human moral discourse (Bender et al., 2021, pp. 616–617).

Therefore, when an LLM generates a moral judgment, it is not engaging in the "ordinary scientific methods" or moral reasoning that Boyd identifies as a reliable procedure for obtaining knowledge about reality (Boyd, 1988, p. 307). As Liao (2020) argues regarding artificial moral status, for an entity to

possess the physical basis for genuine moral agency, it must possess the capacity to track, grasp, and act upon the underlying moral rationales that make an action right or wrong (p. 489). Integrating Liao's criteria into our framework significantly strengthens the ontological critique of artificial moral realism by exposing a structural, rather than merely technological, deficit in generative AI. Because an LLM operates entirely within a self-contained syntactic loop—devoid of intentionality, semantic grounding, and a truth-tracking causal connection to the external world—it is ontologically incapable of accessing, let alone evaluating, the normative reasons that constitute moral facts. The machine does not underperform at tracking objective moral properties; rather, its architectural Level of Abstraction (LoA) makes such tracking a metaphysical impossibility. This ontological barrier is crucial for our broader argument: if LLMs cannot satisfy the baseline semantic and rational demands of moral realism, they cannot be treated as objective moral truth-trackers. Consequently, any attempt to design "Trustworthy AI" frameworks that treat model outputs as "de-biased" moral benchmarks commits a profound category error, obscuring the syntactic reality of statistical language generation under an anthropomorphic illusion. To accurately capture what the machine is actually doing when it "speaks" about morality, we must abandon realist models and shift toward a framework that accounts for the expression of purely syntactic probabilities.

2.2. Introducing Mechanical Expressivism

Given that Large Language Models (LLMs) fail to track mind-independent moral properties or satisfy the criteria of artificial moral realism, an alternative framework is required to account for the nature and function of machine-generated normative language. When a generative system outputs well-formed moral assertions, it utilizes a vocabulary traditionally reserved for human ethical deliberation. To resolve this ontological tension, we propose a novel descriptive framework: Mechanical Expressivism. This framework adapts classical non-cognitivist expressivism to the unique computational architecture of generative AI, demonstrating that while the machine operates without internal intelligence, it functions as an automated vehicle for the expression of compressed normative attitudes.

Simon Blackburn refines and expands this non-cognitivist view through the framework of quasi-realism. Blackburn establishes that expressivism does not seek to explain ethical discourse by representing aspects of the world, but rather by analyzing it as being fundamentally about practice, choices, and actions that express our evaluative dispositions (Blackburn, 1999, p. 213). The core objective of quasi-realism is to show how an expressivist foundation can legitimately accommodate realist-sounding talk of moral truth and objectivity without relying on robust metaphysical structures (Blackburn, 1999, pp. 213–214, 216). Humans naturally speak of moral claims within a propositional framework, but Blackburn argues that this grammar does not imply the existence of independent moral properties; instead, it provides an elegant framework for systematizing, debating, and negotiating our shared practical commitments (Blackburn, 1999, pp. 214, 216).

The philosophical foundation for this approach lies in the non-cognitivist tradition, most prominently articulated by A. J. Ayer. In Language, Truth and Logic, Ayer (1946) argues that ethical concepts are fundamentally pseudo-concepts that add no factual content to a proposition (p. 110). When an agent states that an action is morally wrong, they are not describing an inherent property or asserting a verifiable factual truth; rather, they are merely expressing internal emotional disapproval (p. 111). Because these normative assertions contain no genuine propositions, they are inherently unverifiable, serving a purely emotive and expressive function that manifests the speaker's evaluative state rather than representing objective reality (p. 112).

When applied to generative AI, this expressivist lineage provides the precise conceptual tools needed to interpret the moral language of machines without falling into anthropomorphic fallacies. Under the framework of Mechanical Expressivism, we argue that when an LLM produces a normative output, it is engaged in a process that structurally mimics human expressivism, yet occurs entirely without an underlying conscious mind—a condition we define as agency without intelligence. Just as Ayer demonstrates that moral symbols add no factual content to a statement (Ayer, 1946, p. 110), the normative text generated by an LLM contains no factual reference to objective moral properties. The machine's

output does not state truths; it merely simulates the expression of evaluative attitudes.

However, a radical ontological divergence occurs regarding the source and nature of the expressed attitude. In human expressivism, the linguistic utterance is a direct reflection of an individual's internal, subjective emotional state or evaluative attitude (Ayer, 1946, pp. 110, 113). The machine, by contrast, possesses no internal life, no conscious affect, and no capacity for personal commitment. It cannot experience the moral approval or disapproval that Searle or Ayer attributes to human non-cognitivism. Instead, through the metaphysics of statistical compression, the LLM operates as an automated, mechanical proxy for collective human normativity.

The "attitudes" expressed by the model are not its own, but are the mathematically compressed, probabilistic averages of the diverse human moral perspectives embedded within its training corpora. When an LLM evaluates a scenario as "unjust" or "virtuous," it is executing a purely syntactic calculation that projects the dominant normative regularities and evaluative patterns of its data into a highly coherent expressivist form. The propositional appearance of its output—its ability to mimic objective, realist moral arguments—is best understood through Blackburn's quasi-realist lens: it is a surface-level grammatical mimicry that conceals a fundamentally non-cognitivist, expressive mechanism (Blackburn, 1999, p. 214).

By defining this phenomenon as Mechanical Expressivism, we establish a robust terminology that distinctively departs from traditional accounts of non-cognitivism. Whereas human expressivism is deeply rooted in subjective affect, emotional states, and intentional evaluative commitments, Mechanical Expressivism describes an entity that expresses normative claims entirely devoid of an internal affective life. The "attitudes" the machine expresses are not personal moral judgments, but rather the mathematical compression and statistical average of collective human sentiments embedded in the training data. This ontological distinction is particularly critical for current discussions in AI ethics. Contemporary AI ethics often frames "value alignment" as a process of teaching or aligning models to adhere to objective ethical standards or genuine human values. However, Mechanical Expressivism clarifies that we are not aligning AI with objective values; rather, we are constraining it to reproduce dominant statistical regularities under the guise of algorithmic neutrality. The LLM acts as an active, generative participant in the linguistic loop, but its agency remains strictly mechanical. This descriptive framework clarifies the paper's original contribution and sets the stage for analyzing the systemic risks that arise when human users form a hybrid epistemic partnership with a machine that mechanically expresses statistical averages rather than tracking objective truth or possessing genuine moral commitments.

3. Ontological Hermeneutical Injustice as a Failure of Trust

3.1. The Mechanics of Statistical Normativity and Hybrid Agency

When a human user engages in a continuous loop of prompting and refining moral arguments with an LLM, the seemingly passive statistical mirror transforms into an active, normative force. To understand the mechanics of this transformation, we must analyze the cognitive architecture of this interaction by critiquing traditional active externalism and applying the framework of distributed cognition.

Traditional active externalism, as pioneered by Andy Clark and David Chalmers, provides the foundational starting point for examining how external artifacts supplement human cognition. In their seminal work, Clark and Chalmers (1998, p. 7) argue for the "Extended Mind" thesis, which posits that the biological brain can couple with external entities to form an integrated cognitive system in its own right. They assert that "the human organism is linked with an external entity in a two-way interaction, creating a coupled system" where the external component plays an active role in driving cognitive processes (Clark & Chalmers, 1998, p. 8). Unlike the static, passive cognitive tools typically cited in extended mind literature—such as Otto's notebook (Clark & Chalmers, 1998)—an LLM is a generative, opaque, and non-deterministic artifact.

However, when evaluating contemporary generative AI, this traditional framework of active externalism becomes fundamentally inadequate. Unlike Otto's passive notebook, an LLM is a generative, opaque, and non-deterministic artifact. It does not merely return the exact text previously stored by the

user. Instead, it utilizes its high-dimensional vector spaces to produce novel, probabilistic arrangements of language. Consequently, the relationship between the human user and the generative model cannot be conceptualized as a simple extension of a biological agent into a passive tool.

To accurately capture this interaction, we must look to Orestis Palermos's framework of distributed cognition and epistemic collaboration. Palermos (2022) incorporates Dynamical Systems Theory (DST) to explain how genuine cognitive integration occurs when systems "engage in continuous, reciprocal interactions with each other—such that the effects of each system are continuously fed back to itself" (p. 1487). According to Palermos (2022, pp. 1487–1488), when two systems are coupled in a way that their operations continuously affect and are affected by each other in real-time, they give rise to a single, integrated distributed cognitive system. In the context of generative AI, the prompt-response cycle constitutes exactly this kind of dynamic, bidirectional feedback loop. The human user inputs a query, the LLM generates a probabilistic output, and the human responds by adjusting their reasoning or reformulating the prompt. This continuous interaction creates what we define as a "Hybrid Epistemic Agent".

The systemic risk to "Trustworthy AI" arises from the specific way epistemic responsibility and cognitive labor are redistributed within this hybrid architecture. As Palermos (2022, pp. 1491–1492) notes from a virtue reliabilist perspective, for a distributed system to produce justified knowledge, its integrated parts must operate reliably and cooperatively interact to maintain cognitive integration and epistemic responsibility. However, because the LLM component of the Hybrid Epistemic Agent is inherently driven by the mechanics of statistical compression, it possesses a structural preference for majoritarian linguistics and statistical regularities. It does not evaluate moral arguments based on validity or truth-tracking capabilities, but based on the highest probability of token combinations.

When the human agent relies heavily on this coupled loop for moral reasoning, the machine's statistical normativity ceases to be a passive reflection and becomes an active normative influence. This transition from mere statistical language generation to active cognitive dictation is structurally enforced by contemporary AI alignment mechanisms, most notably Reinforcement Learning from Human Feedback (RLHF) (Unver, 2026, p. 4). RLHF does not simply filter out "toxic" text; it operates by mathematically penalizing outliers and rewarding outputs that align with a pre-defined, majoritarian statistical consensus, effectively constraining the AI to a "bounded moral context" (Wallach & Vallor, 2020, p. 387). Within the continuous reciprocal feedback loop of the Hybrid Epistemic Agent, this mathematical constraint functions as a rigid normative regulator. When a user queries a complex ethical dilemma, the RLHF-optimized model mechanically channels the response away from diverse or marginal perspectives, pushing the interaction toward a standardized, "safe" average. Consequently, the user is forced to refine their prompts and formulate their moral arguments strictly within the conceptual boundaries permitted by the algorithm's reward function. Through this bidirectional interaction, the model actively homogenizes the hybrid agent's moral reasoning. This demonstrates that the integration of generative AI into human moral practice is not an epistemically neutral extension of cognition, but an active restructuring of agency that subtly subordinates human reflective autonomy to mathematically enforced majoritarian compliance.

3.2. Ontological Hermeneutical Injustice

By adapting Miranda Fricker's framework of epistemic injustice to the unique ontology of machine-generated language, this section argues that the deployment of mechanical expressivist tools inherently inflicts a novel, systemic form of harm: Ontological Hermeneutical Injustice.

In her foundational text, Epistemic Injustice: Power and the Ethics of Knowing, Fricker (2007) defines the central case of hermeneutical injustice as having significant areas of one's social experience obscured from collective understanding due to structural identity prejudices in the shared hermeneutical resource (p. 155). This species of injustice stems from hermeneutical marginalization, where unequal distributions of power systematically exclude certain groups from the practices that generate collective meanings (p. 153). Consequently, marginalized individuals face a structural disadvantage or "lacuna" in the collective pool of meaning, leaving them unable to render their distinct lived experiences socially intelligible

(pp. 147–148, 160). While traditional human societies maintain this marginalization through political and institutional barriers (pp. 152–153), under the conditions of Mechanical Expressivism, this process undergoes a radical transformation.

When transposed onto the operational framework of generative AI, hermeneutical marginalization undergoes a radical transformation, shifting from a purely sociological phenomenon to a rigid, mathematical property driven directly by the architecture of Mechanical Expressivism. In traditional human societies, hermeneutical marginalization is maintained through political, historical, and institutional barriers that prevent marginalized groups from contributing to the common pool of meaning (Fricker, 2007, pp. 152-153). However, there is a direct causal link between Mechanical Expressivism and Ontological Hermeneutical Injustice: because a mechanical expressivist lacks the capacity to evaluate the truth or contextual depth of a moral claim, it merely calculates the most probable syntactic expression of past collective attitudes. This means the model inherently equates normative validity with statistical frequency. Confronted with the 'inductive ambiguity' of complex human values (Taylor et al., 2020, pp. 346-347), standard neural networks bypass contextual nuance and rely on superficial statistical correlations to force a standardized, aligned answer. Thus, the very mathematical compression required for an LLM to successfully "express" a coherent majoritarian attitude simultaneously and inevitably requires it to structurally suppress non-dominant linguistic and moral patterns.

Consequently, low-frequency linguistic patterns, minority ethical frameworks, and complex, non-normative interpretations of social life are mathematically treated as statistical anomalies or "noise." As the human user continuously interacts with the model in a coupled feedback loop, the generative system mechanically filters out these non-standard concepts, replacing them with standardized, statistically dominant alternatives. This computational filtering creates an artificial structural lacuna in the shared interpretive tools available to the Hybrid Epistemic Agent. The minority perspective is not simply ignored; it becomes structurally unrepresentable within the system's high-dimensional vectors, thereby eroding the diversity of the collective hermeneutical resource.

This systemic erosion constitutes the climax of the ethical critique against uncritical AI integration, directly addressing the international mandates to evaluate the "societal implications" and "ethical requirements" of trustworthy technologies. Truly human-centered AI cannot be achieved if the underlying cognitive architecture systematically strips agents of the subtle, diverse concepts needed to interpret marginal experiences. As Fricker emphasizes, the primary harm of hermeneutical injustice is that it structurally prevents individuals from making their experiences socially intelligible, undermining their status as equal participants in the shared enterprise of knowing (Fricker, 2007, p. 162). By treating normative language as a homogenized product of statistical probability, the mechanical expressivism of LLMs locks the human-machine partnership into a cycle of conceptual conformism. This structural failure demonstrates that technical risk management and superficial data de-biasing are fundamentally insufficient. Because the marginalization is embedded in the very ontology of statistical meaning-making, safeguarding epistemic autonomy requires an explicit recognition of these systemic limits. Consequently, resolving Ontological Hermeneutical Injustice cannot be achieved through algorithmic fine-tuning; it demands an active, structural commitment to preserving interpretive diversity outside the computational loop.

4. Conclusion

The pursuit of Trustworthy AI from a human-centered perspective demands a profound re-evaluation of how generative systems process, mediate, and alter moral language. This re-evaluation is profound precisely because it reveals that treating LLMs as objective moral arbiters—rather than mechanical expressivists—systematically coerces users into epistemic deferral to a statistical algorithm. This paper has demonstrated that traditional models of passive active externalism (Clark & Chalmers, 1998) fail to capture the dynamic, non-deterministic feedback loops that characterize the contemporary Hybrid Epistemic Agent (Palermos, 2022). By evaluating these systems at their proper computational Level of Abstraction (Floridi, 2011), we have shown that artificial moral realism is an ontological impossibility

due to the foundational divide between syntax and semantics (Searle, 1980; Bender et al., 2021). Instead, generative outputs must be interpreted through the framework of Mechanical Expressivism: the machine possesses agency without intelligence, acting as a statistical mirror that projects a mathematically compressed average of the evaluative attitudes embedded within its training data (Ayer, 1946; Blackburn, 1999).

Ignoring this descriptive reality carries severe normative consequences for technological governance. When human agents uncritically couple their evaluative reasoning with predictive engines, they trigger a systemic risk defined here as Ontological Hermeneutical Injustice. Because the underlying mathematics of generative AI optimize for linguistic regularities, they mechanically filter out minority perspectives and complex, non-normative interpretations, thereby structurally eroding collective interpretive resources (Fricker, 2007). This homogenization directly threatens our epistemic autonomy by standardizing human intent and evaluative behavior under the guise of algorithmic neutrality. Ultimately, genuine AI trustworthiness cannot be engineered through superficial technical de-biasing or alignment metrics (like RLHF) alone. It requires a systemic shift toward Epistemic Sustainability—a regulatory and governance commitment to actively safeguard conceptual diversity and preserve human reflective autonomy entirely outside the majoritarian loops of statistical calculation. In practical AI governance, implementing Epistemic Sustainability means moving beyond traditional safety auditing that merely filters "toxic" language. Future Trustworthy AI frameworks must mandate conceptual diversity metrics, requiring developers to systematically evaluate and report how their models handle minority interpretive patterns and non-standard moral ambiguities. Furthermore, governance policies should enforce interface design requirements that explicitly expose the statistical, non-cognitive nature of AI outputs to the user, thereby disrupting the illusion of objective moral authority. By institutionalizing these specific auditing protocols and design standards, governance bodies can ensure that hybrid epistemic environments actively resist conceptual homogenization and protect the human capacity for diverse moral reasoning.

Declaration on Generative AI

The author declares that Large Language Models (LLMs), specifically Gemini, were utilized during the preparation of this revised manuscript. AI assistance was strictly limited to formatting the text into LaTeX, structuring the document according to the workshop guidelines, and assisting in the grammatical condensation of the author's original philosophical arguments to meet page and word constraints. The core intellectual concepts, philosophical arguments, and theoretical frameworks remain entirely the original work of the author.

References

Ayer, A. J. (1946). Language, truth and logic (2nd ed.). Gollancz. (Original work published 1936)

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 FAccT Conference on Fairness, Accountability, and Transparency, 610–623. https://doi.org/10.1145/3442188.3445922

Blackburn, S. (1999). Is objective moral justification possible on a quasi-realist foundation? Inquiry, 42(2), 213–227. https://doi.org/10.1080/002017499321552

Boyd, R. N. (1988). How to be a moral realist. In G. Sayre-McCord (Ed.), Essays on moral realism (pp. 181–228). Cornell University Press.

Clark, A., & Chalmers, D. (1998). The extended mind. Analysis, 58(1), 7–19. https://doi.org/10.1093/analys/58.1.7

Floridi, L. (2011). The philosophy of information. Oxford University Press.

Fricker, M. (2007). Epistemic injustice: Power and the ethics of knowing. Oxford University Press.

Liao, S. M. (2020). The moral status and rights of artificial intelligence. In S. M. Liao (Ed.), Ethics of artificial intelligence (pp. 481–503). Oxford University Press.

Palermos, S. O. (2022). Epistemic collaborations: Distributed cognition and virtue reliabilism. Erkenntnis, 87(4), 1481–1500. https://doi.org/10.1007/s10670-020-00258-9

Searle, J. R. (1980). Minds, brains, and programs. Behavioral and Brain Sciences, 3(3), 417–424.

Taylor, J., Yudkowsky, E., LaVictoire, P., & Critch, A. (2020). Alignment for advanced machine learning systems. In S. M. Liao (Ed.), Ethics of artificial intelligence (pp. 342–382). Oxford University Press.

Unver, M. B. (2026). Re-aligning value alignment: A metaethical perspective on AI ethics. In S. S. Gouveia (Ed.), The Palgrave handbook on the ethics of artificial intelligence (pp. 1–16). Palgrave Macmillan. https://doi.org/10.1007/978-3-032-15112-4_1

Wallach, W., & Vallor, S. (2020). Moral machines: From value alignment to embodied virtue. In S. M. Liao (Ed.), Ethics of artificial intelligence (pp. 383–412). Oxford University Press.