Intersect, Vol 19, No 1 (2025)

Philosophy of the Machines: A Manifesto for Humans in the Age of Artificial Agents

Generoso Immediato
Hitachi Rail

Abstract

We must reconsider our relationship with machines as artificial intelligence (AI) evolves from task-based support to autonomous or assisted generation. Generation is not creation.

Human intelligence—our only known model—is the sole benchmark for evaluating AI. Yet, we lack rigorous comparison standards because we do not fully understand the internal mechanisms of human intelligence. Perhaps it is more appropriate to speak of AI in simplified terms—as something that, under certain conditions, emulates what we commonly perceive as intelligent behavior[1]. But even that remains uncertain.

Let us remember that we still lack a universally accepted definition of intelligence, let alone a definition of thinking, or even a deeper understanding of the complex nature of consciousness[2].

---

[1] This calls into question early benchmarks like the Turing Test—Alan Turing's famous *imitation game* (Turing, 1950), often interpreted to mean that a machine could be deemed intelligent if its behavior is indistinguishable from that of a human. While foundational in the history of AI, this common operational reading conflates behavioral equivalence with intelligence itself (or understanding), without addressing the substance, origin, or epistemic structure of cognition (Oppy & Dowe, 2003). This *Manifesto* challenges such assumptions, arguing that the nature of intelligence demands a deeper philosophical and epistemological inquiry—not merely an evaluation based on outward behavior or performance.

[2] **Consciousness and the Foundations of Meaning.** Few topics encompass the breadth of multidisciplinary inquiry, philosophical depth, and scientific complexity for humanity as the phenomenon of consciousness (Savoldi et al., 2013). Intuitively, consciousness is inherently connected to the very nature of human thought, intelligence, and the linguistic constructs through which meaning emerges. While fundamental questions—such as the origin of the universe and the emergence of life—pose significant definitional and conceptual challenges, these questions gain their deepest significance precisely because consciousness enables humans to reflect upon and search for meaning within the cosmos. ***Universe's origins → Life's emergence → Consciousness.*** Indeed, consciousness may be considered a remarkable phenomenon from an evolutionary and cosmological perspective: it represents a state in which the universe becomes capable of self-observation and introspection. Although consciousness currently stands as the most complex and elusive manifestation of evolutionary processes known to us, it is scientifically prudent to frame it not necessarily as the universe's "ultimate evolutionary stage" but rather as the most sophisticated example of an observer generated by natural processes known thus far. The

Let us also remember that over the last 80 years of computing and automation, the dominant discipline for solving problems has not been AI, but engineering. Engineering provides the methodologies, mental models, and validation frameworks we use to design and deploy systems. AI extends this legacy—and at the same time disrupts it, introducing new epistemic and ethical challenges that cannot be resolved solely through efficiency.

On these premises, this Manifesto articulates the Philosophy of the Machines as a distinct discipline that consolidates established philosophical lines of inquiry into a unified, applicability-oriented corpus with an explicit order of inquiry. Presented as a manifesto, it develops this order across ten interdependent sections. The trajectory culminates in the Δ–η–ζ framework, introduced as a modelling and analytical scaffold—and, where operationally feasible, a basis for measurement—of net gain in real deployments. The aim is to support more realistic, auditable, and human-aligned business cases for AI and generative AI in a labor landscape that must evolve as artificial agents become pervasive and socio-technical complexity continues to expand.

Within this order of inquiry—moving from foundational analysis to an operational modelling scaffold—the Manifesto foregrounds three questions: (i) What kind of "intelligence" are we building? (ii) What kind of humans must we become in response? (iii) How should responsibility be allocated when systems exceed our capacity for full understanding and oversight?

These questions set the Manifesto's agenda for the age of artificial agents.

Executive Summary

Everything begins with a simple but uncomfortable observation: as computational systems—from classical software to *machine learning* and *generative models*—transition from fixed-function tools to advisory systems, humans are no longer just using machines; they are increasingly interpreting them. We are in a new phase in which human labor, in many domains, is being reconfigured around the evaluation, integration, and ethical filtering of machine-generated outputs. We call these new responsibilities *Cognitive Verification and Ethical Oversight*; together, these roles define a new form of epistemic labor that today remains largely unrecognized in current business cases and policy debates. This urgency is not merely theoretical. Industry surveys and practice reports suggest that organizations tend to adopt generative AI (GenAI) through pilot-led, assistant-first integrations, even as scaling remains uneven and

study of consciousness therefore bridges philosophy, neuroscience and psychology, biology, and physics/cosmology, and remains central to understanding what it means for a universe to harbor observers capable of examining their origins and existence.

verification/accountability costs are not yet fully internalized (Accenture, 2024; Boston Consulting Group, 2023; McKinsey Global Institute, 2023; PwC, 2025).

The opening sections—***Definition, Disciplinary Architecture,*** and ***From Information and Engineering to Agency: Establishing the Order of Inquiry***—set the foundations of the *Philosophy of the Machines* discipline. **The Introduction: *The Two Enigmas and the Birth of the Ethical Oversight*** then revisits the long-standing mystery of human intelligence together with a second, often neglected enigma: the layered unpredictability of (i) *computing systems* and (ii) *AI architectures*. Building from these foundations, the *Manifesto* unfolds across ten interconnected sections.

- **Section 1**, *Rethinking Labor, Intelligence, and the AI Economy*, examines how familiar claims of "20–30% efficiency gains" obscure a deeper reconfiguration of human labor: workers are quietly becoming evaluators, integrators, and ethical stewards of machine-generated content.
- **Section 2**, *From Deterministic Tools to Dynamic Advisors: CAD vs. GenAI*, contrasts classical digital tools—deterministic, transparent, and execution-oriented—with generative AI systems, which act as probabilistic advisors, altering authorship, agency, and responsibility.
- **Section 3**, *When the Mirror Mimics Thinking: Digital Twins in the Age of AI*, shows how AI-enhanced digital twins cease to be mere mirrors and become "co-agents," raising new questions of validation, safety, and trust.
- **Section 4**, *The Birth of a New Human Role: Cognitive Verification*, formalizes this emerging task: humans are no longer simply users of tools; they must interpret, interrogate, and selectively accept or reject AI outputs in many contexts.

The next group of sections explores the tensions and limits of this new condition.

- **Section 5**, *Contradiction 1: Epistemic Demands vs. Social Trends*, highlights a growing mismatch between the depth of judgment we require from human supervisors and the broader cultural drift toward speed, simplification, and slogan thinking.
- **Section 6**, *Contradiction 2: Co-Pilots and Misleading Metaphors*, uses the 'co-pilot' metaphor as a case study to show how popular framings can overstate competence and blur lines of accountability and control in operational decision-making.
- **Section 7**, *Toward a Philosophy of AI-Enabled Work*, sketches how work itself is being reconfigured when decision-making is shared with artificial agents.
- **Section 8**, *Why Training Is Not Enough*, argues that upskilling and awareness programs, while necessary, cannot, by themselves, absorb the epistemic burden imposed by advanced AI.

- **Section 9**, *Ethical Implications: Shared Intelligence, Shared Responsibility*, then distils the ethical consequences of this reconfiguration, insisting that a clear allocation of human and institutional responsibility must match any gain in machine capability.
- **Section 10**, *Reframing AI Efficiency: A New Model*, introduces the Δ–η–ζ framework and the notion of "Actual AI Gain". Here, projected benefits are treated as the result of a balance among:
    - Δ (Delta): the domain adaptability and robustness of the AI system,
    - η (Eta): the cognitive and ethical oversight effort demanded from humans, and
    - ζ (Zeta): the systemic friction, organizational constraints, and contextual variability of real deployments.

  This model is proposed as a *thinking scaffold*, not a ready-made metric. It exposes when AI creates true epistemic and economic value and when it simply shifts costs, risks, and complexity onto human supervisors and institutions.

The **Conclusion: *Beyond Efficiency – Toward a New Human–AI Symbiosis*** closes the *Manifesto* by suggesting that the central innovation of the AI era will be the redefinition of human agency and labor as interpreters of AI machines, emphasizing that the most urgent advances will be ethical and cognitive rather than purely technological.

Definition

The *Philosophy of the Machines* (Immediato, 2025) is a multidisciplinary intellectual framework dedicated to understanding the cognitive, epistemological, ethical, and organizational transformations introduced by artificial intelligence (AI) in real socio-technical systems. Within this theoretical framework, AI systems—and, more broadly, computational machines—are always situated alongside human agents and interact with them in continuously evolving environments; both humans and machines co-evolve over time. In principle, what happens during the *design* phase necessarily shapes what occurs during the *operational* phase, as the three components—human, machine, and environment—jointly unfold throughout the system's lifecycle. In this philosophical setting, AI is understood—remaining faithful to the term's historical roots—as the highest currently available expression of computational capacity and complexity (the combination of hardware and software), in which the stochastic nature of AI-level functions compounds the unpredictability of the underlying deterministic computational layers.

The contribution of this discipline is twofold: (i) it operates at the intersection of philosophical traditions that have already addressed, in depth, different facets of human–AI interaction, and (ii) it provides a specific *order of inquiry* (see Section ***From Information and Engineering to Agency: Establishing the Order of Inquiry***). The synthesis of an

organic philosophical landscape—the critique of syntactic "as-if" intelligence and the insistence that genuine understanding is more than formal symbol manipulation (Searle, 1980); the view of cognition as extended and environmentally scaffolded across brains, bodies, and tools (Clark & Chalmers, 1998); the "double enigma" of human and artificial intelligence as jointly reshaping our epistemic landscape (Andler, 2023), with a revised reading of the second enigma, in which the artificial-intelligence side is itself dual, combining the unpredictability of complex computational substrates with the distinct unpredictability of AI models, two conceptually related layers though requiring different technical and regulatory responses; recent work on reasoning, information flow, and agency in artificial agents (Koralus, 2025); and analyses of long-term alignment, control, and systemic risk in advanced AI systems (Bostrom, 2014)—is proposed as a functional integration designed to evaluate the realities we are now modelling and deploying. From this perspective, the theoretical effort—at least as a first conceptual attempt—is to offer the Δ–η–ζ framework (see **Section 10**) as a modelling, analytical, and ultimately measurement tool for a form of human and economic value that cannot and must not be allowed to dissolve into an ever-growing complexity.

Disciplinary Architecture

With the ambition of developing a substantive *philosophy of AI*, it is crucial to situate it within a distinct disciplinary architecture centered on the machine as an engineered artifact capable of processing information and subject to concrete technical and institutional regimes of design, constraint, and evaluation. A purely informational or abstract treatment of AI is therefore insufficient: any adequate philosophy of artificial agents must take seriously the engineered character of these systems, the standards and failure modes they embody, and the socio-technical configurations in which they operate.

Figure 1 depicts the philosophical topology underlying the Philosophy of the Machines: a structured view of how key domains have historically and conceptually converged toward a new interdisciplinary center.

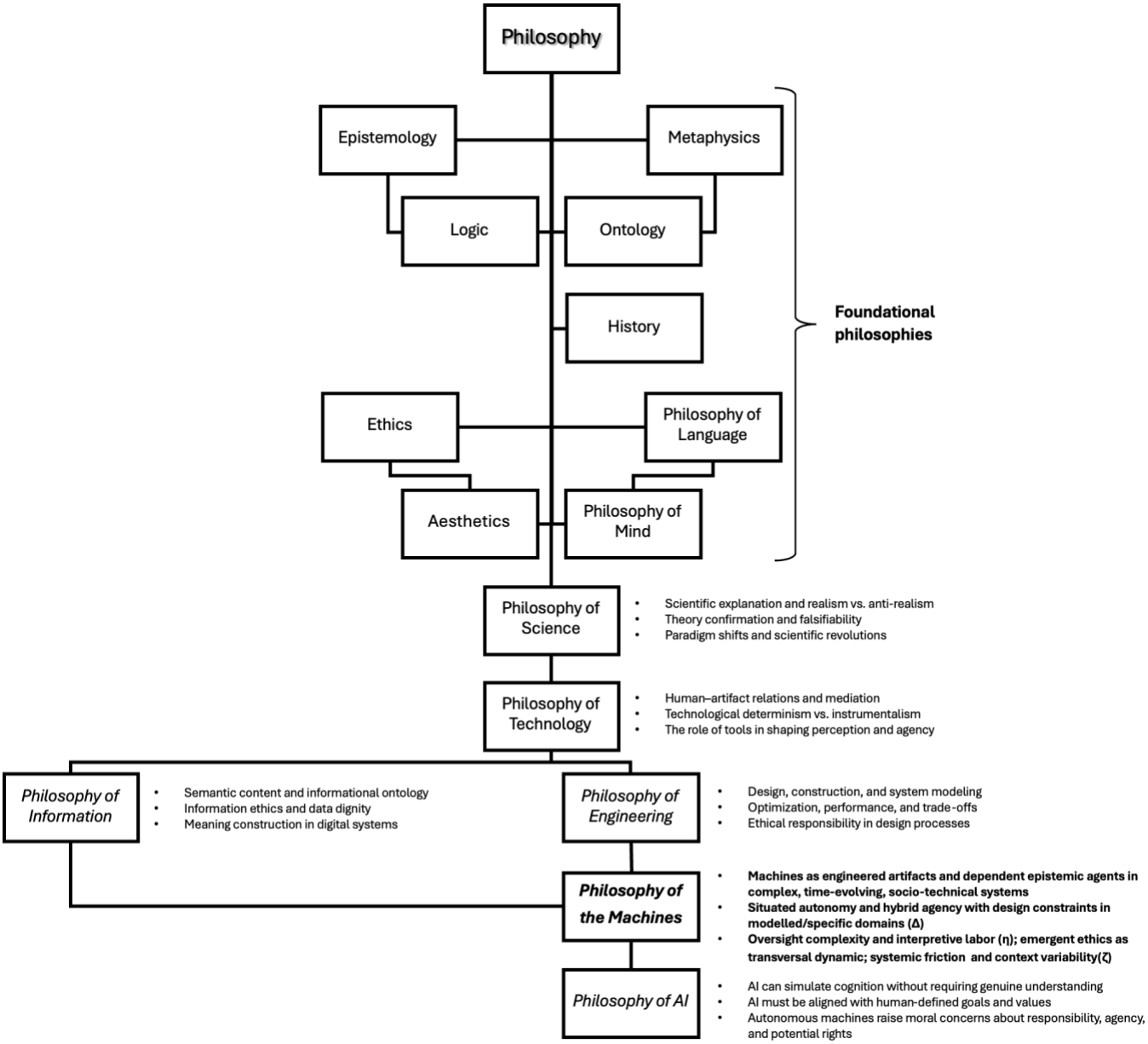

FIGURE 1. Philosophical Lineage of the Philosophy of the Machines.

From Information and Engineering to Agency: Establishing the Order of Inquiry

The Philosophy of the Machines follows a specific order of inquiry that runs from artifact to measured gain, where "gain" denotes the net outcome

(economic and labor gain) attributable to AI within operational environments.

1. **Human intelligence**. The starting premise is that intelligence, in the strong sense, remains human. Artificial systems today cannot be treated as quasi-human minds, but as increasingly powerful epistemic instruments whose behavior and limits must be read against the backdrop of human cognition, judgment, and responsibility.
2. **The machine as an engineered artifact in the human–machine–environment loop**. AI systems are understood first as engineered artifacts: they are designed and in operation. Before discussing AI autonomy or agency—with the necessary restrictions, of course—the analysis asks what the machine is, how it is built, how it fails, and which responsibilities its design and deployment entail.
3. **Dual unpredictability in computational machinery and AI**. Computational machines are complex devices that exhibit unpredictability at different levels. In the case of AI, the non-trivial behavior of the underlying computational substrate is compounded by the stochastic nature of learning-based algorithms. Conceptually, these layers of unpredictability are related, but they must be handled through adequate technical, regulatory, and organizational instruments.
4. **The emergence of new human epistemic labor**. Once AI systems are treated as such artifacts, a new form of work becomes visible. Humans are required to interpret, verify, and contextualize machine outputs, and to ensure that decisions remain aligned with domain constraints, institutional norms, and societal values. This labor is not an accessory; it is structurally necessary.
5. **From systemic synthesis to the Δ–η–ζ model**. On this basis, the Philosophy of the Machines proposes the Δ–η–ζ framework as a meso-level instrument for modelling, analyzing, and ultimately measuring the net epistemic and economic gain of AI deployments. The model links domain adaptability (Δ), oversight effort and complexity (η), and systemic friction (ζ) to the realized value of AI within work and governance, making explicit when projected benefits are genuinely achieved and when oversight burdens and environmental constraints erode them.

In our stance, the *Philosophy of AI* does not precede the *Philosophy of the Machines*; it emerges from it and presupposes its order of inquiry. This investigation also reconnects with the *Philosophy of Information* by foregrounding an operational question: how machines generate, transmit, and transform structured information into *usable knowledge*. In this sense, information becomes *actionable, situated, and operational*. Machines are physical and logical systems that function as ***derivative epistemic agents***, and their operations depend on human design, data, and governance. What

they encode, decode, or obfuscate helps shape the topology of our shared understanding and the effective boundaries of ***human oversight***.

Introduction: The Two Enigmas and the Birth of the Ethical Oversight

As humans reflect and debate on the last 70-80 years, it becomes evident that the mental model, methodology, and acceptance criteria for solving problems take center stage in AI. ***(i) How individuals think, (ii) act on their thoughts, and (iii) accept the outcomes, are, to some extent, mirrored in the human conception of AI.*** This is because, unsurprisingly, AI is a concept developed by human intelligence, and it is essential to emphasize, without forgetting, that these are distinct entities with different meanings (Floridi, 2011; Shneiderman, 2022). It is probably better to say 'two different enigmas' (Andler, 2023). Though it is almost clear how human intelligence and AI relate to the engineering of problem-solving, a dual perspective exists regarding the "mystery" of how humans think and how AI exposes thinking. Putting aside the enigma of human intelligence, engineers and scientists working on AI algorithms and training know that the "second enigma" I am discussing relates to the computer's nature. It is not a novelty; it originates in the very beginning and resides at a lower level of abstraction[3].

Unpredictability in computing systems arises fundamentally from two distinct sources: (i) *inherent theoretical limitations*, such as complexity, randomness, and sensitivity to initial conditions, and (ii) *practical implementation* factors. Although individual computer instructions execute deterministically, system-level interactions, concurrency, and environmental variables introduce emergent behaviors difficult to predict accurately.

Additionally, AI amplifies this intrinsic unpredictability through learning and training processes that are inherently probabilistic, context-sensitive, and opaque, particularly in *deep learning*.

---

[3] **Layered unpredictability**. AI systems operate within high-level algorithmic frameworks, but their behavior also depends on classical software and hardware layers (e.g., compilers, concurrency, scheduling, and processor-level effects). Unpredictability, therefore, does not arise only from stochastic learning: it is entangled with the non-deterministic characteristics of general-purpose computing. In safety-critical domains (aerospace, rail, automotive), engineers mitigate these sources of variability through deterministic design, redundancy, formal methods, and stringent standards (Leveson, 2011). Introducing AI components adds a layered uncertainty: (i) *exogenous deployment effects* (hardware/operating system/runtime), (ii) *inference-time stochasticity and numerical non-determinism*, and (iii) *residual model-internal failure modes* induced by the training objective and underspecification. This structure complicates assurance and makes explainability and accountability depend on explicit system boundaries and operating contexts. Even rigorously engineered systems integrating AI therefore require complementary mechanisms emphasizing interpretability, real-time validation, and robust oversight; otherwise, they drift toward new assurance paradigms that are not yet fully identified, standardized, or institutionally consolidated.

This ***augmented unpredictability*** is a double-edged sword. In certain cases[4], it fosters unprecedented opportunities for innovation, particularly in recent generative AI applications. Simultaneously, it introduces new and critical risks in safety-critical domains and broader social contexts.

It is precisely this augmented unpredictability—among other factors—that elevates AI's ethical implications to a critical level. To responsibly navigate these dynamics, it becomes essential to establish a dedicated human function—systematic ***Ethical Oversight***—grounded in information ethics and responsive to the mapped ethical failure modes of algorithmic systems (Floridi, 1999; Mittelstadt et al., 2016).

A thorough four-point summary (**I–IV**) will merge philosophical, technical, and business perspectives. These three dimensions—how AI systems function, how they reshape human thought and ethics, and how they translate into real economic and organizational value—are essential to unlocking AI's full potential for society.

I. Computer Science and the Verum-Factum Principle:

Giambattista Vico's ***Verum-Factum principle*** *holds that we can truly understand only what we have made* (Vico, 2000/1725). This idea, rooted in classical epistemology, assumes that human-made systems are, in principle, fully knowable. But in the age of AI, this assumption is under strain.

Due to the complexity, nonlinearity, and emergent behavior in AI models, distributed architectures, and security mechanisms, the inherent unpredictability of computational systems challenges our ability to fully control what we have created.

In practice, we often design systems whose behavior we can no longer fully anticipate, explain, or audit. This doesn't invalidate Vico's insight—it reframes it. We may have "made" the code, but no longer entirely "know" what it will do.

This epistemic gap calls for reexamining how we define knowledge, agency, and accountability in computer science, especially in safety-critical or ethically sensitive contexts.

[4] However, the consequences of unpredictability differ radically by domain. In GenAI use cases—artwork, poetry, speculative imagery—creative variability is not only tolerable but desirable: evaluation is aesthetic or interpretive, rather than governed by an objective standard of correctness. Unpredictability becomes a source of novelty. By contrast, where outputs must be functionally correct—such as code generation, legal compliance, medical diagnosis, or safety-critical decisions—ambiguity is not neutral. Results are valid or invalid, safe or unsafe, legal or noncompliant, and such applications demand determinacy, traceability, and logic-based constraints. This distinction is epistemic: it concerns the conditions under which an output can be known, justified, and trusted. In creative domains, judgment is plural and context-dependent; in rule-governed domains, it is bounded by procedures, verifiability, and shared standards. The architecture may be the same, but the *epistemic expectations* are categorically different. We must not confuse stochastic success in open-ended generation with readiness for domains that require analytic rigor and accountability.

II. **Aristotle and the Return to First Principles:** In *Posterior Analytics* (Aristotle, 1984), Aristotle introduced First Principles thinking—the practice of reducing complex systems to their most fundamental, irreducible elements, or ***archai***. These are not derived from other ideas but are the foundational truths upon which knowledge is constructed.

First Principles thinking, as Aristotle framed it, is not just a mental tool—it's a method of *unbuilding assumptions* (Campos Zabala, 2023). In today's AI landscape, where inherited architectures and black-box models often go unquestioned, this approach urges us to examine why systems are structured the way they are, not just how to optimize them. Rather than merely tuning parameters, First Principles thinking invites a return to fundamental design lifecycle questions:

- *What problem are we solving?*
- *For whom?*
- *Under what assumptions?*

In this way, it opens the path for deep innovation, not by accelerating what exists but by resetting the conceptual ground. It allows breakthroughs in algorithm design and human–machine interaction to emerge not from scale but from clarity. First Principles thinking is not limited to engineering optimization—in our perspective, it also offers a framework for ethical design. By reducing AI-related dilemmas to their fundamental *ethical components*[5], developers can embed these values at the system's origin rather than retrofitting them downstream. This reasonable approach shifts ethics from external regulation to internal architecture, promoting the development of AI systems that are efficient, powerful, and aligned with societal values from the start.

As AI systems increasingly shape life at scale—health, finance, law, labor—the emphasis on *ethical-by-design* approaches becomes essential. First Principles thinking provides the clarity to make ethics a design input, not a public relations output.

III. **Innovation Beyond Existing Paradigms:** The First Principles approach, accompanied by its problem breakdown and ethical components, encourages businesses to move beyond incremental improvements and toward transformative innovation. This mirrors the philosophical idea of "thinking outside the box" by questioning the very structure of the "box" itself. This could be translated as developing products generated by new technological paradigms, such as quantum computing or neuromorphic computing, which diverge from traditional binary logic. Similarly, this approach can drive innovation in business models by prompting leaders to

[5] The ***ethical components*** and their initial thematic classification are formally introduced in **Section 10**.

reconsider the interplay among AI, human labor, customer relations, and social sustainability. Businesses with this mindset can better anticipate and shape future trends rather than react impulsively.

IV. **From Heidegger to Post-Humanism: Rethinking the Technical and the Philosophical.** Ultimately, AI demands more than technical sophistication; it requires philosophical reckoning.

- *What does reasoning, deciding, and acting mean in a world where machines could ever do all three?*

These questions extend beyond implementation. Technology is not a neutral instrument; it shapes what becomes salient, what is valued, and how agency is configured (Heidegger, 1977). If AI systems propose actions, influence judgments, and mediate relationships, the question becomes whether humans remain the authors of these outcomes—or increasingly respond to them.

Posthumanist accounts emphasize that we are no longer simply working with machines but evolving through them: the line between user and system becomes philosophical terrain, not just an *interface boundary* (Haraway, 1991; Hayles, 1999).

This is why the relationship between technical and philosophical thinking is no longer optional. It is the condition for responsible design. Classical thinkers knew that wisdom requires both intention and self-limitation. In this spirit, we must approach AI not only as a tool to improve efficiency but also as a mirror that reflects our evolving humanity.

Philosophical inquiry raises a key question as a complement to our initial one:

- *Do we genuinely need AI autonomy—is it an explicit design requirement, or are we mistaking it for one?*

Classical thinkers taught us this dual wisdom: to invent is human, but to govern invention is also human. This dual insight should remain central to how we design and govern AI systems.

Aligned with this established order of inquiry, the *Manifesto* unfolds through *ten interconnected sections* (**1–10**).

1. Rethinking Labor, Intelligence, and the AI Economy

Transitioning from foundational methodologies—such as First Principles thinking, innovation strategies, and system thinking—to the pragmatic challenges of large-scale AI implementation marks a pivotal moment for renewed philosophical inquiry; it is not only a technical matter. At this stage in AI's evolution, deeper reflection offers crucial insight into the shifting relationship between *human cognition* and the *machine's intelligence-like behavior.*

A subtle but profound philosophical question emerges: what new forms of human labor are being created in response to AI? This question not only initiates a necessary philosophical inquiry but also addresses the

aspect of AI's economy: anticipating and articulating the foundations of a sustainable business case for AI and GenAI.

The commonly cited efficiency gains of 20–30%[6], occasionally up to 40%,—often presented in consulting benchmarks and transformation forecasts (Accenture, 2024; Boston Consulting Group, 2023; McKinsey Global Institute, 2023; PwC, 2025)—suggest a significant productivity uplift. But if these gains are confirmed in practice, they mean more than increased efficiency. AI isn't only taking over repetitive tasks—it's quietly redrawing the boundary between human and machine roles, especially in how decisions are made and who makes them.

Human workers are being redefined—not only as producers and consumers of knowledge but increasingly as evaluators, integrators, and "ethical stewards" of machine-generated content.

2. From Deterministic Tools to Dynamic Advisors: CAD vs GenAI

In conventional digital tools, such as Computer-Aided Design (CAD), the system operates as a passive, deterministic enhancer.

This category of digital tools boosts productivity, accuracy, and visualization. But their behavior is deterministic: these systems respond only to user inputs and predefined functions. They don't explore, suggest, or improvise. The user defines the problem, and the system executes. All creativity in these tools still belongs entirely to the human.

In contrast, AI typically functions as a dynamic and probabilistic advisor.

AI generates content, proposes decisions, and can even initiate action (e.g., in RPA or automated systems), yet its outputs are not always grounded in traceable logic or deterministic rules in a way that supports human-auditable justification (Doshi-Velez & Kim, 2017). As such, the user can no longer assume full authorship or complete control; a shared

---

[6] The "20–30% efficiency gain" did not originate with AI. It reflects a well-established estimation range historically used in business transformation programs—including Lean initiatives, *Enterprise Resource Planning* (ERP) rollouts, process automation, and shared services transitions. When these projections are conducted rigorously, with traceable assumptions, empirical baselines, and sound modeling techniques, they can offer meaningful foresight and strategic value. However, when used superficially, without access to the full data, methodology, or auditability, such estimates can become intellectually fragile. In these latter cases, critical reasoning suggests that the 20–30% range persists primarily because it is:

- **Large enough** to justify investment,
- **Small enough** to remain credible and achievable,
- **Simple enough** to model in 2–3 year *Return on Investment* (ROI) cycles.

This *Manifesto* challenges the uncritical acceptance of such foresight by introducing the Δ–η–ζ framework (see **Section 10**). Without such a lens, AI gains risk being not only exaggerated but also ethically unsustainable and economically misleading, especially when the hidden costs of oversight and integration remain unaccounted for.

responsibility emerges, and with it a new basis for trust in human–machine systems (Nissenbaum, 2001).

3. When the Mirror Mimics Thinking: Digital Twins in the Age of AI

Digital twins were born from a deterministic tradition. They were designed as high-fidelity, real-time replicas of physical systems—models that observe, simulate, and optimize reality through a closed loop of data and causality.

Yet, when one of their components becomes an artificial agent—when a digital twin includes an AI—the notion of "twinning" begins to dissolve.

In theory, a twin mirrors reality. But AI introduces interpretation. Therefore, the mirror becomes "non-deterministic". In conclusion, the Digital Twin concept is shattered.

Digital twins incorporating embedded artificial intelligence transcend mere reproduction of physical behavior; they evolve, infer, and possess the capability to generate recommendations. They may modify simulation outcomes based on contextual data or develop insights from usage patterns in ways that remain beyond the complete understanding of their creators.

Consequently, in a certain way, the digital twin transcends its role as a reflection of the physical systems under investigation, transforming instead into an advisory entity or a hybrid artifact. It evolves from a mere mirror to becoming a "co-agent" (this is one of the contradictions addressed later).

At the core of this shift lies a deeper and controversial paradox.

In traditional systems engineering, a digital twin is considered scientifically reliable because it is traceable: its logic, structure, and data sources are transparent and reproducible. But artificial intelligence—especially deep learning—breaks that assumption.

The first problem is that AI models often operate as black boxes, trained on datasets whose full provenance is unknown or inaccessible. Their decision-making logic emerges from high-dimensional spaces that defy intuitive understanding. Because many AI components are pre-trained, fine-tuned, or evolved, even developers may lack a complete view of what the model knows or how it was shaped.

In this sense, we are embedding within the twin a mechanism that we do not fully understand. Yet, we ask the composite system to simulate reality as if it were deterministic.

This raises a critical epistemological question:

- *Can a twin be trusted as a scientific replica if part of its reasoning cannot be audited or reconstructed?*

This question creates epistemic and ontological tension. If part of the twin is an opaque model—a black-box neural net or a probabilistic agent—then its behavior can no longer be traced solely to the physical world it represents. Then the second problem: one that lives in software, not sensors.

From a systems engineering standpoint, this has concrete consequences. AI-enhanced digital twins pose challenges:

- **Validation**: *What does verifying a model that learns and adapts mean?*
- **Safety Assurance**: *How do we certify behavior if one component generates outputs based on training data and evolving inference patterns?*
- **Trust**: *Can we trust a model that is no longer fully explainable, even if it performs better?*

Philosophically, these twins are no longer twins. They are *alter egos* of the systems they model—not just observers but interpreters. Their presence demands a new framework of accountability, where simulation becomes co-decision, and virtuality is not a shadow but a partner. Even our models are no longer passive in the age of artificial agents.

We must now ask:

- *What are we building when the mirror not only reflects, but begins to reason and respond?*

4. The Birth of a New Human Role: Cognitive Verification

This new human-AI interaction introduces a new category of human tasks: interpreting, evaluating, and integrating AI outputs within a context-sensitive framework that encompasses both real and digital environments.

We name this emerging responsibility ***Cognitive Verification***[7]—the "System 2" burden of checking outputs that may be fluent yet ungrounded, echoing the symbol-grounding problem (Kahneman, 2011; Harnad, 1990).

It is a form of *epistemic validation* in which the human does not verify the system itself, but instead evaluates the meaning, trustworthiness, and applicability of a machine-generated output. This task arises wherever AI systems produce results that require interpretation within context, and its critical nature scales with the consequences of the decision involved.

This is not verification in the traditional sense, such as checking a mathematical formula, nor is it merely prompt engineering or post-editing of AI-generated text. It is more nuanced: *epistemic filtering*[8]—a

---

[7] This task is closer to *epistemic validation*, but we use the term *Cognitive Verification* to emphasize the human role in actively interrogating the AI's output—not verifying the system as a whole but assessing the trustworthiness and applicability of a specific suggestion or result.

[8] **JORABP Classification**. AI-generated outputs can be categorized into three epistemic classes: (1) correct outputs whose validity is easily and reliably verifiable; (2) incorrect outputs whose invalidity is readily evident and therefore dismissible; and (3) incorrect but highly plausible outputs which, despite their relative rarity, pose substantial epistemic and practical challenges. This third class-subtle inaccuracies or context-dependent errors that remain credible-demands significant cognitive effort for detection and mitigation (Fu et al., 2023) and is central to robust Cognitive Verification in high-stakes domains (healthcare, law, and safety-critical engineering). Let us define these categories as follows:

thoughtful, high-stakes review of what the AI might mean, what assumptions it has made, and what consequences might follow if its output is accepted or acted upon. Examples of this new responsibility include contexts where machine-learning systems and generative AI are already being piloted:

- **In healthcare**, a clinician may receive an automatically generated differential diagnosis or treatment plan. The machine accelerates option exploration, but the final decision depends on verifying fit with the patient's history and comorbidities, local protocols, and legal constraints. *Cognitive Verification* means treating the AI output as a hypothesis to be checked against medical knowledge, guidelines, and the concrete situation of a vulnerable human being.
- **In law**, a generative system may draft clauses fluently, summarize case law, or propose strategies. Yet because it has no standing in the legal order and bears no accountability, lawyers must verify jurisdiction, enforceability, and whether wording shifts risk between parties. *Cognitive Verification* is the distinction between stylistic plausibility and legally binding meaning.
- **In engineering and safety-critical infrastructure**, an AI advisor might suggest design variants, maintenance plans, or mitigations. It can surface non-obvious patterns but lacks an intrinsic sense of physical consequences, certification regimes, or liability chains. Engineers interpret suggestions through standards, failure modes, and safety cases. Here, *Cognitive Verification* is inseparable from *Ethical Oversight*: acceptance or rejection propagates into real-world risks for passengers, workers, or the public.

These are ***decisions of consequence*** demanded from humans with inputs they did not entirely create, based on logic they may not fully understand.

5. Contradiction 1: Epistemic Demands vs. Social Trends

A striking contradiction emerges when we examine the kind of judgment now required of humans, which was previously analyzed under the framework of Cognitive Verification.

On the one hand, we expect *deeper, slower, more critical thinking*—careful discernment across technical, ethical, and contextual layers. On the other hand, society is moving steadily in the opposite direction.

- **Juicy Oranges (JO)** → *valid and reliable outputs*
- **Rotten Apples (RA)** → *clearly incorrect outputs*
- **Banana Peels (BP)** → *plausibly correct, but dangerously slippery errors*

To quantify the frequency and impact of each category, statistically rigorous empirical studies are required. Such evidence would clarify the prevalence of Banana Peels in practice and inform training, oversight, and validation strategies.

The rise of social media and digital immediacy has led to cultural habits of:

- **Slogan thinking**
- **Rapid-fire opinions**
- **Copy-paste and forward logic**
- **Neglect of fake news**

These habits can increase susceptibility to fake news, bullshit, and overclaiming (Salvi et al., 2023). All of them contribute to an oversimplified picture of complex and nuanced subjects, which can occasionally be problematic or, even more so, harmful to society.
This cognitive flattening undermines the posture required for effective human oversight. If humans are expected to supervise powerful AI systems—interpreting probabilistic, nuanced, and context-dependent outputs—this imposes sustained attention and the retention of intention, capacities that can be degraded by short-form, high-switching media environments (Chiossi et al., 2023).

This contradiction may constitute the greatest threat to safe AI implementation: humans may no longer be equipped to identify AI errors due to a gradual erosion in sustained attention and critical, analytical reasoning.

Can this trend be reversed? Is our society amid a temporary shift, or are we undergoing a lasting transformation in how we think, learn, and behave, shaped by the architecture of our digital environments? This question remains open, but its stakes are clear: our ability to meaningfully oversee AI may depend on whether we still possess the habits of attention and reflection that oversight requires.

6. Contradiction 2: "Co-Pilots" and Misleading Metaphors

Technology companies have popularized the term "co-pilot" to describe artificial intelligence systems that assist humans.

This metaphor is designed to suggest support, not replacement, casting AI as a reassuring assistant rather than a dominant force.

But the metaphor deserves scrutiny. In aviation, a co-pilot is:

- **Trained to the same standard as the pilot**
- **Expected to take full control in the event of an emergency**
- **Embedded in a system defined by hierarchy, regulation, and constant rehearsal**

In contrast, the AI "co-pilot" in current use is:

- **Often probabilistic and non-transparent in reasoning**
- **Lacking common sense or context awareness**
- **Generating outputs that require human interpretation and correction**

So, the question arises: Is AI really a co-pilot—or is it something else entirely?
If it requires supervision, then perhaps the metaphor should be revised. This reframing opens urgent questions:

- *Who is truly in command?*
- *Do current metaphors obscure more than they clarify?*
- *In safety-critical contexts, can we afford metaphors that overstate AI's competence and understate the need for scrutiny?*

Re-examining the "co-pilot" metaphor is consistent with the *archai* concept[9], as delineated in Aristotle's First Principles (see **The Introduction: *The Two Enigmas and the Birth of the Ethical Oversight*, Point II**). Language shapes our expectations, and misleading metaphors can lead to incorrect assumptions and misplaced trust.
The point here is to acknowledge the genuine usability value of these tools, while observing that when such metaphors are embedded in product strategies, marketing narratives, and even policy documents, they can quietly install an inflated sense of competence and an ambiguous picture of responsibility.

7. Toward a Philosophy of AI-Enabled Work

These two contradictions point toward a deeper philosophical need: to rethink AI and the kind of humans we must become to use it well.

Efficiency gains are not automatic or free—they require new perceptions, interpretations, and ethical discernment capacities.

The real transformation of the AI era is elevating human judgment to a higher, more complex form, not just automation. If we fail to cultivate those human capacities, no amount of computational power will secure safety or wisdom.

AI may be fast. But the human response must be *deep, slow, and deliberate*.

8. Why Training is Not Enough

So far, the response has often been to "just train people to use AI." But that is insufficient. Why?

- **AI outputs are non-repetitive**: Every generation is context-dependent, shaped by nuance in prompts, data sources, and model architecture.
- **Domain knowledge is insufficient**: Even experts in their field may not intuitively understand how the AI arrived at its output, or how to challenge it effectively.
- **Cognitive strain and accountability**: The more we rely on AI, the higher the stakes in *knowing when to intervene*. This introduces

[9] This critical examination of metaphors is consistent with Aristotle's theory of ***archai***: according to it, sound knowledge must begin with correct First Principles or else the entire reasoning structure collapses. In this context, metaphors such as "co-pilot" act as ***quasi-archai***: they shape the initial mental models by which people understand the role and behavior of artificial agents. A misleading metaphor can install a false epistemic starting point, fostering miscalibrated trust and misplaced responsibility in human–AI interaction (Shneiderman, 2022). Therefore, reevaluating such metaphors is not merely a linguistic concern, but a philosophical necessity rooted in the integrity of reasoning itself.

stress and responsibility that training alone can't resolve and requires additional effort.

A novel *epistemic burden* emerges—interpreting the *probabilistic reasoning* of another "mind," albeit artificial (Doshi-Velez & Kim, 2017; Shneiderman, 2022).

Software has always been labor-intensive, and we well understand how challenging it can be to interpret or modify code written by third parties, often to the point where rewriting it from scratch is considered the most cost-effective and time-efficient option.

9. Ethical Implications: Shared Intelligence, Shared Responsibility

This leads to a philosophical shift: AI systems may increase efficiency but simultaneously generate novel obligations for human oversight. The time saved by automation may be reinvested into risk evaluation, moral reasoning, and cross-disciplinary sense-making.

We are entering an era where human labor is no longer centered on production but curation.

This curation is active and interpretive, requiring a new kind of intellectual flexibility that may not be teachable in a traditional manner.

This raises key questions:

- *Can companies cultivate these meta-cognitive skills as part of corporate culture? And how?*
- *Should responsibility for evaluating AI outputs be distributed collectively or concentrated in expert roles?*
- *What happens when the speed of AI generation outpaces the human capacity for judgment?*

10. Reframing AI Efficiency: A New Model

Foresight studies often project efficiency gains by adopting AI and GenAI[10]. Yet these projections rarely account for the costs of **a) Cognitive Verification** and **b) Ethical Oversight,** which are combined in the overall **c) Human Oversight Effort** now required of human workers.
Formally represented as an equation:

[10] The model is designed to offer a more realistic account of the efficiency expected from AI adoption. It deliberately focuses on cognitive and ethical costs in real deployments—human oversight, interpretive ambiguity, systemic friction, and the evolving roles of workers interacting with machine-generated outputs—dimensions too often underrepresented in traditional efficiency forecasts and automation business cases. Accordingly, the model excludes a separate but significant layer: the infrastructure and technical costs of operationalizing AI (compute/cloud consumption, data and deployment pipelines, and systems and security engineering). In large-scale enterprise or critical domains, these costs can materially influence Total Cost of Ownership (TCO) and scalability. While outside the scope of this *Manifesto*, they are assumed as an *operational baseline*.

$$\text{(1)} \quad \text{HumanOversightEffort} = (\text{CognitiveVerification} + \text{EthicalOversight})$$

To address a realistic measure, we need a more nuanced model:

$$\text{(2)} \quad Actual\ AI\ Gain_{\%} = 100 \cdot [\Delta \cdot \text{ForesightEfficiency} - \eta \cdot \text{HumanOversightEffort}]$$

Where:

- **Foresight Efficiency** and **Human Oversight Effort** are percentages (in decimal form), representing theoretical time or cost impact. The final output is scaled to return a percentage value ranging from –100% to +100%, with negative values indicating a net loss from AI adoption.
- **Δ** adjusts for domain adaptability/robustness under distributional/context shift; domains differ in how stable their constraints and verification regimes are (Leveson, 2011; Shneiderman, 2022; Vaccaro et al., 2024).
- **η** accounts for the complexity of human oversight (e.g., critical fields such as healthcare or autonomous systems that require nuanced ethical judgment). The η-factor represents complexity and stress from a human perspective, which may vary across domains or environmental and institutional conditions that shape both deployment outcomes and the attribution of epistemic authority to AI systems (Floridi, 1999; Ferrario et al., 2024; Mittelstadt et al., 2016; Vaccaro et al., 2024).

Δ and η factors are modeled as normalized values between 0 and 1.

This methodology implies that future assessments must precisely characterize the Δ-factor for each relevant domain or business segment, tracing where projected efficiency gains are expected to emerge, while validating these assumptions through targeted analytical procedures. The η-factor, potentially even more complex, requires careful quantification, ideally informed by structured inquiry and empirical observation across operational and decision-making contexts.

If the expression inside the brackets is negative, the resulting "Actual AI Gain" may drop to zero or even below, signaling that the cost of human oversight outweighs the benefits of automation.

While **Equation (2)** centers on domain adaptability (Δ) and human-context complexity (η), subsequent refinements incorporate a **ζ-factor** to reflect individual interpretive variability, contextual ambiguity, organizational readiness, and cultural friction influence AI adoption outcomes. Without this new ζ-factor, the previous equation represents an idealized scenario in

which an entity experiences no internal friction and is fully prepared for AI integration. In a realistic scenario[11], the equation to consider is:

$$(3) \quad Actual\ AI\ Gain_{\%} = 100 \cdot [\Delta \cdot \text{ForesightEfficiency} - (\eta \cdot \text{HumanOversightEffort} + \zeta)]$$

Where ζ, representing ***Systemic Friction*** and ***Contextual Variability***, is modeled as a normalized value between 0 and 1; it reflects the efficiency loss introduced by real-world conditions that hinder the effective adoption of AI/GenAI given an adequate operational baseline (Pasquale, 2015). It captures the cumulative effect of:

- **Individual interpretive variability**: inconsistencies in human interaction with AI outputs.
- **Contextual ambiguity**: variability in operating conditions, workflows, and institutional constraints.
- **Organizational readiness gaps**: limitations in process integration, operating procedures, governance maturity, or leadership alignment.
- **Cultural friction**: resistance to automation or ethical discomfort within teams or institutions.

ζ values near 0 indicate minimal *systemic friction* (ideal deployment conditions), while values closer to 1 signal *severe misalignment* that can completely negate or invert projected efficiency gains. This model reveals that efficiency is not absolute: context, technology maturity, multiple risks, and human cognitive labor shape it. In domains such as medicine, engineering, or law, oversight costs may significantly diminish—or even reverse—the anticipated benefits of AI.

As such, the complete **Equation (3)** should serve as a foundational guideline for constructing more realistic and responsible business cases for AI, where strategic decisions are guided not only by projected automation gains but also by the often-overlooked human oversight burden and the systemic friction of implementation and organizational alignment[12].

---

[11] We relegate full numerical examples to **Appendix A**, where two contrasting scenarios (*high-gain* and *low-gain* software development with GenAI) are worked out in detail.

[12] **Structural intuition about the Δ–η–ζ model: law of diminishing returns and a scaling-failure regime.** The Δ–η–ζ model of "Actual AI Gain" introduced in **Section 10** is designed to encode the epistemic value of artificial agents' foresight in complex socio-technical systems. For convenience, we can write the inner term of the equation in compact form as $G = [\Delta \cdot F - (\eta \cdot H + \zeta)]$, with $\Delta, F, \eta, H, \zeta \in [0, 1]$, hence $G \in [-2, 1]$. Here, F denotes *foresight efficiency*, H the *human oversight effort* (Cognitive Verification plus Ethical Oversight). In the context of this *Manifesto*, G can be interpreted as a *net epistemic gain at deployment:* a local balance between what the system adds in usable foresight and what it subtracts in oversight burden and friction. Under mild and natural assumptions, this structure yields a ***diminishing-returns*** effect in realized deployment gain. Consider a family of increasingly scaled systems: $\begin{cases} G_A = [\Delta_A \cdot F_A - (\eta_A \cdot H_A + \zeta_A)] \\ G_B = [\Delta_B \cdot F_B - (\eta_B \cdot H_B + \zeta_B)] \\ G_C = [\Delta_C \cdot F_C - (\eta_C \cdot H_C + \zeta_C)] \end{cases}$. Assume that foresight improves with scale: $F_A < F_B < F_C$. At the same time, assume the oversight

The ethical evaluation of AI systems throughout the lifecycle requires a structured framework. It can be organized in the identification of the **core ethical components**, grouped into three main thematic clusters based on their focus:

- **Human-Centered Ethics (i–v):** Addressing the direct impact on human well-being, dignity, autonomy, and accessibility.
- **Governance Ethics (vi–viii):** Ensuring responsible agency, regulatory compliance, non-discrimination, and cultural inclusiveness.
- **Data and System Ethics (ix–xii):** Safeguarding privacy, informational integrity, system resilience, and interpretability.

These ethical components, critical for Cognitive Verification and Ethical Oversight, are:

i. **Human health and safety**
ii. **Fairness in process and outcomes**
iii. **Equality of opportunity and access**
iv. **Representation and accessibility in design and impact**
v. **Human autonomy and agency**
vi. **Environmental and societal sustainability**
vii. **Accountability (external answerability)**
viii. **Responsibility (internal ethical duty)**
ix. **Transparency and explainability**
x. **Legal and regulatory compliance**
xi. **Privacy and Information Integrity**
xii. **Security and resilience against threats**

This set can be directly referenced when applying the Δ–η–ζ model, particularly for analyzing ***cognitive verification*** (Doshi-Velez & Kim, 2017) and ***ethical oversight*** (Dignum, 2019; Mittelstadt et al., 2016) terms. For this purpose, **Equation (1)** can be explicitly integrated into **Equation (3)**, yielding the following formulation:

$$(4) \quad Actual\ AI\ Gain_{\%} = 100 \cdot \{\Delta \cdot \text{ForesightEfficiency} - [(\eta_1 \cdot \text{CognitiveVerification} + \eta_2 \cdot \text{EthicalOversight}) + \zeta]\}$$

As the complexity of real-world deployment grows, so does the burden of ensuring that outputs meet these ethical standards. Hence, ethical

burden terms $\eta \cdot H$ and friction ζ increase with complexity. Then, even if $G_B > G_A$, the incremental gain may shrink so that $G_C - G_B < G_B - G_A$, which matches the classical idea of diminishing marginal benefit. This effect can be reinforced if domain adaptability Δ itself decreases with scale. Within this framework, a ***scaling-failure regime*** corresponds to the case in which the cost term overtakes foresight: $\Delta \cdot F < \eta \cdot H + \zeta \Rightarrow G < 0$. Beyond this point, the system may still produce impressive outputs in isolation, but its *net epistemic gain becomes negative*: the additional foresight is more than outweighed by the human and systemic effort required to verify, contextualize, and safely absorb it. In this sense, the Δ–η–ζ structure shows how both can emerge endogenously from the coupling of foresight, oversight, and friction—consistent with broader calls to evaluate adoption through robustness and safety rather than performance alone (Taddeo, 2025).

oversight becomes a multi-dimensional epistemic duty essential for adopting trustworthy and responsible AI, and its associated cost term is made explicit in the model. This operationalization yields **Equation (4)**, which makes explicit the distinction between η_1 and η_2 as complexity parameters within the Δ–η–ζ model: η_1 scales the Cognitive Verification burden, while η_2 scales the Ethical Oversight burden. **Appendix B** then provides the First Principles decomposition and the mapping from problem elements to ethical components, showing how element-level ethical dimensions recombine into system-level ethical behavior.

Conclusion: Beyond Efficiency – Toward a New Human-AI Symbiosis

Beneath the strategic layer lies a deeper ontological shift: humans are becoming interpreters of machine creativity; as a consequence, the role of humans is not only to *do the work and make the final decision,* but also to *decide what the machine's work means*.

This is a new form of cognition, a new form of responsibility, and possibly "a new kind of labor."

We are now tasked with building ethical, useful AI and cultivating the human maturity required to live and work with it. This maturity is philosophical: humble in the face of complexity, alert to context, and conscious of its limits.

As AI continues to evolve, the most urgent innovation will not be technological but ethical, cognitive, and deeply human.

Let this be a philosophy not of fear or unquestioning optimism but of *preparedness, reflection, consciousness, and mutual transformation.*

We built the machines. But the question for humanity remains: how are we evolving?

References

Accenture. (2024). *Reinventing with a digital core* [Research report]. Accenture. https://www.accenture.com/dk-en/insights/technology/reinventing-digital-core

Andler, D. (2023). *Intelligence artificielle, intelligence humaine: La double énigme*. Odile Jacob.

Aristotle. (1984). *Posterior analytics*. In J. Barnes (Ed.), *The complete works of Aristotle*. Princeton University Press.

Boston Consulting Group. (2023). *How people can create—and destroy—value with generative AI*. https://www.bcg.com/publications/2023/how-people-create-and-destroy-value-with-gen-ai

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Campos Zabala, F. J. (2023). *Grow your business with AI: A first principles approach for scaling artificial intelligence in the enterprise* (Vol. 1). Apress. https://doi.org/10.1007/978-1-4842-9669-1

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

Deloitte. (2018). *The robots are waiting: Are you ready to reap the benefits of RPA?* [Online report].

Dignum, V. (2019). *Responsible artificial intelligence*. Springer.

Doshi-Velez, F., & Kim, B. (2017). Toward a rigorous science of interpretable machine learning (arXiv:1702.08608). *arXiv*. https://arxiv.org/abs/1702.08608

Floridi, L. (1999). Information ethics: On the philosophical foundation of computer ethics. *Ethics and Information Technology, 1*(1), 33–52. https://doi.org/10.1023/A:1010018611096

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

Fu, Y., Xue, Y., Wang, W., Liu, J., Lou, Y., Liu, Z., Fan, Y., & Chen, J. (2023). Security weaknesses of Copilot-generated code in GitHub projects: An empirical study. *arXiv*. https://doi.org/10.48550/arXiv.2310.02059

Haraway, D. J. (1991). *Simians, cyborgs, and women: The reinvention of nature*. Routledge.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena, 42*(1–3), 335–346. https://doi.org/10.1016/0167-2789(90)90087-6

Hayles, N. K. (1999). *How we became posthuman: Virtual bodies in cybernetics, literature, and informatics*. University of Chicago Press.

Heidegger, M. (1977). *The question concerning technology*. In *The question concerning technology and other essays* (W. Lovitt, Trans.). Harper & Row.

Immediato, G. (2025, May 1). *Philosophy of the machines: A manifesto for humans in the age of artificial agents*. Medium. https://medium.com/@generoso.immediato/philosophy-of-the-machine-0590bea0623e

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Koralus, P. (2025). The philosophic turn for AI agents: Replacing centralized digital rhetoric with decentralized truth-seeking. *Mind & Society*. https://doi.org/10.1007/s11299-025-00326-z

Chiossi, F., Haliburton, L., Ou, C., Butz, A., & Schmidt, A. (2023). Short-form videos degrade our capacity to retain intentions: Effect of context switching on prospective memory. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM. https://doi.org/10.1145/3544548.3580778

Latour, B. (1993). *We have never been modern*. Harvard University Press.

Ferrario, A., Facchini, A., & Termine, A. (2024). Experts or authorities? The strange case of the presumed epistemic superiority of artificial intelligence systems. *Minds & Machines, 34*, 30. https://doi.org/10.1007/s11023-024-09681-1

Leveson, N. G. (2011). *Engineering a safer world: Systems thinking applied to safety*. MIT Press.

McKinsey Global Institute. (2023). *The economic potential of generative AI: The next productivity frontier* [Report]. McKinsey & Company. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2). https://doi.org/10.1177/2053951716679679

Nissenbaum, H. (2001). Securing trust online: Wisdom or oxymoron? *Boston University Law Review, 81*(3), 635–664.

Oppy, G., & Dowe, D. (2003). The Turing test. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2021 ed.). https://plato.stanford.edu/entries/turing-test/

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press.

PwC. (2025). *2026 AI business predictions*. PwC. https://www.pwc.com/us/en/tech-effect/ai-analytics/ai-predictions.html

Salvi, C., Barr, N., Dunsmoor, J. E., & Grafman, J. (2023). Insight problem solving ability predicts reduced susceptibility to fake news, bullshit, and overclaiming. *Thinking & Reasoning, 29*(4), 760–784. https://doi.org/10.1080/13546783.2022.2146191

Savoldi, F., Ceroni, M., & Vanzago, L. (Eds.). (2013). *La coscienza: Contributi per specialisti e non specialisti tra neuroscienze, filosofia e neurologia*. Aras Edizioni.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*(3), 417–457.

Shneiderman, B. (2022). *Human-centered AI*. Oxford University Press.

Taddeo, M. (2025). *Intelligenza artificiale: basta cercare performance migliori. Concentriamoci su solidità e sicurezza*. https://www.corriere.it/economia/opinioni/25_maggio_28/intelligenza-artificiale-basta-cercare-performance-migliori-concentriamoci-su-solidita-e-sicurezza-faaed6a2-9765-4077-ace9-51392f3f5xlk.shtml

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433–460.

Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis.

Nature Human Behaviour, 8(12), 2293–2303. https://doi.org/10.1038/s41562-024-02024-1

Vico, G. (2000). *The new science* (D. Marsh, Trans.). Penguin Classics. (Original work published 1725)

Willcocks, L. P., Lacity, M. C., & Craig, A. (2017). Robotic process automation: Strategic transformation lever for shared services? *Journal of Information Technology Teaching Cases, 7*(1), 17–28. https://doi.org/10.1057/s41266-016-0016-9

Appendix A – Applying the Δ–η–ζ Model: A Case Study in Software Development

Note and Scope: The following case study is entirely simulated and academic in nature. Its purpose is to illustrate how the Δ–η–ζ model introduced in Section 10 can be instantiated with concrete numbers to support strategic decisions about GenAI adoption in software development. It does not describe any specific organization.
For consistency with Section 10, we use the following compact formulation of the Actual AI Gain (see **Equation (3)**):

$$\text{Gain}_{\%} = 100 \cdot [\Delta \cdot \text{F} - (\eta \cdot \text{H} + \zeta)]$$

Where:

- **F (Foresight Efficiency)** is the projected benefit from AI adoption in an idealized scenario (e.g., nominal 40% productivity gain, expressed as a fraction of 1).
- **H (Human Oversight Effort)** is the share of effort required to review, correct, or integrate AI outputs.
- Δ, η, and ζ are the three scaling factors defined in **Section 10**: domain adaptability, oversight complexity, and systemic friction, each in [0, 1].

Values inside the parentheses can become negative. In such cases, the Actual AI Gain should be treated as effectively zero: under those conditions, the deployment is not yet epistemically or economically justified.

A.1 Scenario 1 – High-Gain Organization

In the first scenario, a technology company has deployed a GenAI assistant for code generation and refactoring. The tool is embedded in a relatively mature engineering environment: stable tech stack, shared coding standards, and well-defined review practices.

Model parameters:

- F = 0.40 (a 40% projected productivity gain from faster and more consistent code generation),
- Δ = 0.90 (high domain adaptability due to modular design, shared libraries, and language stability),

- $\eta = 0.30$ (oversight is focused rather than exhaustive, supported by good tooling and clear review boundaries),
- $H = 0.20$ (only 20% of total effort is spent on human review, thanks to test automation and CI integration),
- $\zeta = 0.05$ (low systemic friction: strong tool integration, consistent documentation, and team buy-in).

Computation:

$$\text{Gain}_{\%} = 100 \cdot [0.90 \cdot 0.40 - 0.30 \cdot 0.20 - 0.05]$$
$$= 100 \cdot [0.36 - 0.06 - 0.05] = 100 \cdot 0.25 = 25\%$$

Despite the nominal 40% foresight gain, the **realized gain is about +25%**. The model highlights why:

- High Δ reflects a domain that is structurally well-aligned with GenAI (stable patterns, modular architecture).
- η is non-zero, but manageable, because linters, static analysis, and clear responsibility boundaries support oversight.
- ζ is kept low by organizational alignment and consistent practices.

From a Δ–η–ζ perspective, this is a case where *technical readiness* and *organizational maturity converge* to turn a marketing number into a substantial, but more modest, effective gain.

A.2 Scenario 2 – Low-Gain Organization

In the second scenario, another company deploys a similar GenAI assistant for software development, but in a less mature context: heterogeneous codebases, weak standards, and ambiguous review responsibilities.

Model parameters:

- F = 0.40 (the same nominal 40% projected gain),
- Δ = 0.70 (only moderate adaptability due to varying languages, legacy systems, and inconsistent architectures),
- η = 0.50 (oversight is complex: developers must deeply review logic, maintainability, and security),
- H = 0.30 + 0.20 = 0.50 (30% Cognitive Verification effort + 20% ethical/legal/security checks),
- ζ = 0.15 (non-trivial friction from tool integration, documentation gaps, and cultural resistance).

Computation:

$$\text{Gain}_{\%} = 100 \cdot [0.70 \cdot 0.40 - 0.50 \cdot 0.50 - 0.15]$$
$$= 100 \cdot [0.28 - 0.25 - 0.15] = 100 \cdot (-0.12) = -12\%$$

Formally, the model yields a **-12% gain**: under these conditions, the AI deployment destroys value rather than creating it. The interesting point is that the foresight term matches Scenario 1. What changes is the configuration of Δ, η, and ζ:

- Oversight complexity (η) is high because machine-generated code is hard to trust, debug, and integrate.
- Human oversight effort (H) expands to half of the process, combining cognitive and ethical checks.

- Systemic friction (ζ) captures fragmentation in workflows, uneven trust in the tool, and difficulty tracing responsibility.

From a Δ–η–ζ perspective, the lesson is not that GenAI for coding is intrinsically flawed, but that:

- Projected gains are conditional, not absolute.
- Without structural preparation (lower η and ζ), even a technically capable system (Δ = 0.70) can result in negative net gain.

A.3 Strategic reflection

These two simulated scenarios show that the same nominal foresight gain (40%) can, in theory, produce very different outcomes once domain adaptability, oversight complexity, and systemic friction are accounted for.

The Δ–η–ζ model provides a *transparent scaffold* for asking better questions:

- Are we deploying GenAI in a domain where Δ can realistically approach 1?
- Have we explicitly budgeted for the Cognitive Verification and Ethical Oversight effort captured by η and H?
- Where does ζ hide in our organization—in culture, processes, or infrastructure—and how can we reduce it?

In this sense, the model is not only a diagnostic tool but also a way to design more transparent and sustainable AI business cases aligned with the broader argument of the *Manifesto*.

Appendix B – Ethical Decomposition from First Principles

This appendix provides the formal statement by which First Principles (see **The Introduction: *The Two Enigmas and the Birth of the Ethical Oversight*, Point II**) decomposition is applied—starting from the decomposition of the initial problem into elemental units—to the ethical components introduced in **Section 10**.

Let $\mathcal{D}_P$ represents the Problem Domain, and P represents the complex system-level problem $P \in \mathcal{D}_P$.

First Principles decomposition yields a set of elemental sub-problems, $\{p_1, p_2, \ldots, p_n\} \subset \mathcal{D}_P$, where each p_i is a fundamental archai (First Principles' element).

Define a technical composition function:

$$f: \{p_1, p_2, \ldots, p_n\} \to P$$

This function f reconstructs the system P by integrating its elemental parts p_i.

Each p_i is mapped to a subset of ethical components $\{ e_{i_1}, e_{i_2}, \ldots, e_{i_m}\}$ using a mapping function:

$$g: \{p_1, p_2, \ldots, p_n\} \to \mathcal{P}(\mathcal{D}_E)$$

Where $\mathcal{P}(\mathcal{D}_E)$ denotes the power set of the Ethical Domain $\mathcal{D}_E$. Thus, for each p_i, an ethical subset $g(p_i)$ is individually determined.

Formalized as:

$$g(p_i) = \{ e_{i_k} \mid k \in \{1,2,..,m\},\ e_{i_k} \in \mathcal{D}_E \}, \text{ with } m \geq 1$$

System-level ethical behavior e_s is not merely the union of individual ethical components e_{i_k}. Instead, it emerges through a synthesis function h:

$$h\colon \bigcup_{i=1}^{n} g(p_i) \rightarrow e_s, \text{ thus: } e_s = h(\bigcup_{i=1}^{n} g(p_i)), \text{ where } e_s \in \mathcal{E}_S$$

Where h represents complex ethical interactions (emergent properties, tensions, non-linearities, possible risks, and conflicts across technical and ethical components).

In theory, if the system-level ethical behavior e_s coincides exactly with a single ethical component $e_{i_k} \in \mathcal{D}_E$, then $e_s \in \mathcal{D}_E$. However, in most realistic cases, ethical emergence involves multiple dimensions and trade-offs, requiring the recognition of a distinct Emergent System Domain $\mathcal{E}_S$. Generally speaking, e_s is an emergent, complex ethical phenomenon, possibly new, unforeseen, and irreducible. More philosophically, it reflects the *systemic ethical state of the whole machine-human system*, not of any part alone.

Decomposing problems into *archai*-level elements $\{p_i\}$ through f (Aristotle, 1984), projecting ethical dimensions $\{e_{i_k}\}$ individually through g (Dignum, 2019; Floridi, 1999), and synthesizing system-level ethics e_s rigorously through h (Mittelstadt et al., 2016; Latour, 1993; Nissenbaum, 2001), we ensure that ethical oversight grows naturally from foundational design, not as an afterthought.

Let us summarize the key conceptual steps supporting the formal statement:

- Technical integration adopting First Principles: $\{p_i\} \xrightarrow{f} P$
- Ethical projection and synthesis:
  $$\{p_i\} \xrightarrow{g} \{g(p_1), g(p_2), \ldots, g(p_n)\} \xrightarrow{\cup} \bigcup_{i=1}^{n} g(p_i) \xrightarrow{h} e_s$$
- Expanded view:
  - Solve technically: $P = f(p_1, p_2, \ldots, p_n)$
  - Evaluate ethically: $\{\bigcup_{i=1}^{n} g(p_i)\} \cup \{e_s\}$

We conclude that respecting First Principles requires treating technical and ethical decompositions at the same epistemological level.

Data and Code Availability

https://github.com/gimmediato/Philosophy-of-the-Machines

Acknowledgements

The author wishes to thank Antonella Migliardi for her curatorial support and careful editorial assistance in preparing this manuscript for publication.