# Capability, Strategy, and Organisational Integration: Reframing the AI Debate

Evoluit M.
Independent Theoretical Research Evoluism Initiative

March 2026

**DOI:** https://doi.org/10.5281/zenodo.19017694

**Abstract**

Artificial intelligence is frequently interpreted through a wide range of ontological frameworks: as a computational tool, an emerging agent, a powerful optimiser, or a potential form of artificial mind. Despite their differences, many of these interpretations implicitly assume that increasing computational complexity may correspond to progressively more integrated forms of cognition.

The paper does not aim to adjudicate between these ontological interpretations. Instead, it isolates a more specific problem that cuts across them: the relationship between behavioural capability and organisational integration in artificial systems.

To analyse this issue, the paper introduces a minimal set of analytic distinctions (inspired by Evoluist philosophy) that separate behavioural capability from organisational integration. Within this framework, artificial intelligence systems are interpreted not primarily as individual agents or proto-subjects but as technologically sustained regimes of structural coordination embedded in large-scale infrastructures.

The analysis develops three main claims. First, dominant interpretations of AI often rely on an implicit continuity assumption between structural complexity and integrational autonomy. Second, contemporary narratives about artificial general intelligence frequently involve what may be termed evolutionary projection: the transfer of explanatory models derived from biological evolution onto technological systems that emerge within fundamentally different organisational regimes. Third, the most significant risks associated with artificial intelligence may arise not from the emergence of artificial subjectivity but from the structural integration of powerful optimisation systems into institutional and technological infrastructures.

The paper therefore proposes a reframing of the AI debate: from the question of whether machines are becoming minds to the question of how regimes of coordination are being embedded within technological civilisation.

**Keywords:** artificial intelligence; organisational integration; capability; strategy; optimisation; scaling; Evoluism; structural risk; technological infrastructure.

## 1. Introduction: The Ontological Inflation of Artificial Intelligence

Artificial intelligence occupies a peculiar position in contemporary thought. Few technological phenomena have generated such a dense and unstable layer of ontological interpretation. AI is described, often within the same discussion, as a computational instrument, a functional realisation of cognition, an emerging optimiser, a proto-agent, a social actor, or even a potential

bearer of consciousness. These descriptions do not merely differ in emphasis; they frequently rely on incompatible conceptual assumptions about what kinds of things AI systems are supposed to be.

This situation may be described as an ontological inflation of artificial intelligence. The term does not refer simply to conceptual excess or terminological looseness. It designates a more specific philosophical disorder: the multiplication of ontological attributions in the absence of prior analysis of the regime within which the phenomenon becomes distinguishable. AI is treated as if its ontological status were self-evident, while in fact the discourse surrounding it shifts constantly between categories that belong to different levels of analysis.

Some theories understand artificial intelligence in primarily functional terms. On this view, what matters is not substrate but organisation; if the relevant functional architecture is present, then cognitive attribution becomes legitimate. Other approaches treat AI as an optimiser whose importance lies in the capacity to pursue objectives across environments with increasing power and strategic flexibility. Still others, especially in contemporary scaling discourse, suggest that sufficient growth in parameters, data, and compute may itself produce qualitatively new cognitive states. In each case, artificial intelligence is inserted into a broader ontological narrative: as mind, as agency, as intelligence, as proto-subjectivity, or as a new stage in the evolution of complex systems.

The aim of the present paper is not to deny that AI systems display remarkable forms of coordination, adaptability, or causal efficacy. Nor is it to dismiss the practical importance of their integration into science, industry, governance, or communication. The question is more basic and, for that very reason, more difficult: what kind of phenomenon is artificial intelligence, and under what analytical conditions may it be correctly described?

This question becomes pressing precisely because most current debates do not begin by asking it. Instead, they frequently inherit an implicit assumption that remains insufficiently examined: that increasing structural complexity in artificial systems can be interpreted as a continuous trajectory toward more integrated forms of cognition. Once this assumption is in place, the discourse rapidly populates itself with familiar expectations — artificial general intelligence, machine agency, alignment with superhuman systems, digital minds, emergent subjectivity. The debate then unfolds as though the continuity between statistical optimisation, integrated cognition, and potentially consciousness were already philosophically secured.

The present paper argues that this continuity has not been secured. More strongly, it argues that much of the ontological debate about AI rests on a confusion between fundamentally different regimes of integration. The problem is therefore not only that current descriptions of AI may be exaggerated, anthropomorphic, or speculative. The deeper problem is that they often presuppose a transition that has not been demonstrated: a transition from forms of structural coordination to forms of integration capable of reproducing the conditions of their own stability.

To analyse this problem, the paper introduces a minimal conceptual framework inspired by distinctions developed in Evoluist philosophy. Yet Evoluism is not introduced here as a doctrine demanding assent, nor as a self-enclosed philosophical system to be expounded for its own sake. It is used as an analytic frame for placing pressure on the hidden assumptions of contemporary AI discourse. Its principal relevance lies in three distinctions: between elements, regimes of manifestness, and conditions of possibility; between structural coordination and integrational self-stabilisation; and between forms of complexity that remain externally sustained and those that reproduce the mechanisms of their own integration.

The analytic distinctions developed below contest this presupposition and offer a different

way of posing the problem.

The stakes of this argument are not merely classificatory. If the ontological status of AI is misunderstood, then its practical integration into institutional and civilisational systems will also be misunderstood. Systems that remain structurally coordinated but externally sustained may come to occupy positions that implicitly presuppose semantic autonomy, goal ownership, or agentive responsibility. In that case, the major risks of AI will not arise from the appearance of artificial minds, but from the premature attribution of roles that belong to a different regime of integration.

The paper proceeds accordingly. It first reconstructs the dominant ontologies of AI and isolates the assumption they share. It then introduces the analytic distinctions necessary to analyse that assumption. On this basis, it argues that contemporary AI systems are more adequately understood as technologically sustained regimes of structural coordination than as emerging subjects or minds. The later sections draw out the consequences of this analysis for AGI discourse, for the question of artificial consciousness, and for the problem of AI safety within the broader technological civilisation.

The argument developed here does not deny the possibility that artificial systems could eventually develop new forms of integration. Its aim is rather to clarify the organisational conditions under which such a transition would need to occur. By distinguishing between capability, strategy, and integration, the analysis seeks to prevent the premature conflation of behavioural complexity with organisational autonomy. In this sense, the paper should be read not merely as a contribution to contemporary AI debate, but as an Evoluist intervention into that debate.

## 2. Contemporary Styles of Ontological Attribution in Artificial Intelligence

Debates about artificial intelligence are often presented as disagreements concerning technical questions: model architectures, training regimes, computational scaling, or safety strategies. Yet these discussions frequently rely on deeper assumptions about what kind of phenomenon artificial intelligence is supposed to be. These assumptions rarely appear as explicit ontological doctrines. Instead, they operate as interpretive styles through which the behaviour of artificial systems is understood.

Rather than treating these perspectives as fully articulated ontologies, it may be more precise to describe them as styles of attribution. Each style emphasises a different aspect of artificial systems and attributes different kinds of properties to them. In doing so, each foregrounds certain organisational features while leaving others largely unexamined.

Three such styles have been particularly influential in contemporary discussions of AI: functionalist attribution, optimiser attribution, and scaling attribution.

### 2.1. Functionalist Attribution

Within the philosophical tradition of functionalism, mental states are understood in terms of their causal and organisational roles rather than their physical realisation. Early formulations by Hilary Putnam and Jerry Fodor emphasised that cognitive states can be characterised by the functional relations they bear to inputs, outputs, and other internal states. On this view, the specific material substrate implementing these relations is not decisive.

This approach has had a lasting influence on philosophical interpretations of artificial intelli-

gence. If cognitive states are defined by their functional organisation, then systems implemented in silicon rather than biological neurons may in principle realise similar patterns of organisation. From this perspective, the question of artificial cognition becomes primarily a question about whether the relevant organisational patterns can be instantiated in artificial systems.

Later discussions, particularly those associated with Daniel Dennett, have emphasised that many cognitive phenomena may be understood as patterns emerging within complex information-processing systems. In this view, the attribution of mental properties often depends on the explanatory usefulness of treating a system as exhibiting certain patterns of organisation rather than on identifying a specific physical mechanism.

Functionalist interpretations therefore encourage the idea that increasingly complex computational systems may eventually exhibit patterns that justify cognitive descriptions. Importantly, however, functionalist authors rarely claim that such developments must occur. Rather, they argue that the possibility of artificial cognition cannot be dismissed solely on the basis of substrate differences.

From a Dennettian perspective, cognitive phenomena can be understood as patterns emerging within sufficiently complex information-processing systems rather than as properties tied to specific biological substrates. The analytic distinctions developed here do not contradict Dennett's insight that cognition consists of patterns; they simply ask under what conditions those patterns become capable of sustaining the processes that stabilise their own existence. Dennett's own account treats such patterns as explanatorily sufficient for cognition. The present analysis does not deny that sufficiency for explanatory purposes, but it raises a further question: whether the organisational conditions that allow such patterns to persist over time are themselves internally reproduced or remain externally scaffolded. A purely pattern-based account explains how cognition can be identified, but it does not by itself explain how the organisational conditions required for the persistence of such patterns are maintained across time.

### 2.2. Optimiser Attribution

A different interpretive style has emerged within contemporary AI safety research. Here artificial systems are often analysed through the conceptual framework of optimisation. Rather than focusing on cognitive organisation, this perspective examines how machine learning systems search large possibility spaces in order to maximise specified objective functions.

This approach is particularly visible in the work of Nick Bostrom, who characterises advanced artificial systems as powerful optimisers capable of pursuing objectives across complex environments. Subsequent work in alignment research has examined how such optimisation processes may generate behaviours that diverge from the intentions of system designers.

Researchers such as Evan Hubinger and Joe Carlsmith have analysed scenarios in which learned systems develop internal optimisation processes or strategies that differ from those intended during training. Concepts such as mesa-optimisation and deceptive alignment describe situations in which systems behave in ways that appear strategically coherent even when the underlying mechanisms do not correspond to explicit goals held by the system itself.

Within this interpretive style, the central concern is not whether artificial systems are conscious but whether sufficiently powerful optimisation processes may generate behaviours that influence or destabilise human institutions.

### 2.3. Scaling Attribution

A third influential perspective has developed within the empirical research culture of machine learning itself. Studies of scaling laws, including work by Jared Kaplan and later analyses by Jordan Hoffmann, have demonstrated systematic relationships between model size, training compute, and performance across a wide range of tasks.

These findings have encouraged a research programme centred on the expansion of computational scale. Larger models trained on larger datasets tend to display broader capabilities, including more coherent language generation, improved reasoning-like behaviour, and increasingly flexible problem-solving.

Importantly, the scaling literature itself typically makes limited ontological claims. Most researchers in this tradition emphasise empirical relationships between model scale and task performance rather than philosophical conclusions about consciousness or agency. Nevertheless, public and academic discussions surrounding scaling often extrapolate from improvements in capability to broader claims about the trajectory of artificial intelligence.

### 2.4. A Shared Tension

These interpretive styles do not assert identical theses about the nature or future of artificial intelligence. Functionalist discussions emphasise organisational patterns; optimisation-based analyses focus on strategic behaviour; scaling research concentrates on empirical relationships between computational magnitude and performance.

Yet all three styles confront a related question that remains only partially articulated: how increases in capability relate to forms of organisational integration. Functionalist accounts ask whether artificial systems could instantiate cognitive patterns. Alignment research investigates whether optimisation processes embedded in machine learning systems may generate strategic behaviour. Scaling research documents rapid improvements in performance as computational resources increase.

The present paper does not claim that these perspectives share a single doctrinal commitment. Instead, it suggests that they collectively reveal an unresolved tension in contemporary AI discourse: the relationship between capability, strategy, and integration remains conceptually unclear.

Clarifying this relationship requires analytical distinctions capable of separating different forms of organisational dependence. The following section introduces a minimal conceptual framework that allows such distinctions to be explored.

## 3. Capability, Strategy, and Integration

Contemporary discussions of artificial intelligence frequently move between three different dimensions of analysis. First, there are questions about capability: how well a system performs tasks such as language generation, reasoning, or problem solving. Second, there are questions about strategy: whether systems exhibit behaviours that resemble goal-directed optimisation across complex environments. Third, there are questions about integration: how the organisational stability of a system is maintained across time.

These dimensions are often discussed together but do not necessarily coincide. A system may display impressive capabilities while remaining highly dependent on external infrastructures.

Conversely, a system may exhibit forms of organisational stability without demonstrating advanced cognitive behaviour.

This distinction becomes particularly relevant in debates surrounding artificial intelligence. Improvements in machine learning models are typically measured in terms of capability: larger models produce more accurate predictions, solve more tasks, and generate more coherent outputs. Alignment research focuses on strategy, examining how optimisation processes may produce behaviours that affect the world in complex ways.

Much less attention, however, has been devoted to the question of integration. In what sense do artificial systems maintain the organisational conditions necessary for their own stability? Which mechanisms sustain their operation across time, and where are these mechanisms located?

These questions do not negate the importance of capability or strategy. Instead, they introduce a third dimension that has often remained implicit in discussions of artificial intelligence.

The argument developed in this paper is therefore not that contemporary AI research relies on a single mistaken assumption. Rather, it suggests that debates about artificial intelligence frequently move between capability, strategy, and integration without clearly distinguishing among them.

The analytical framework introduced in the next section provides a way of tracking these differences more precisely.

## 4. An Analytic Distinction for Studying AI Integration

The preceding discussion suggests that contemporary debates about artificial intelligence frequently move between questions of capability and questions of strategy while leaving a third dimension comparatively underexamined: the organisational conditions under which complex systems maintain their stability.

These distinctions originate within Evoluist philosophy but are used here in a minimal analytic form intended to clarify organisational dependence in artificial systems. These distinctions are not presented as a comprehensive theoretical framework. Their role is methodological: they provide conceptual tools for tracking how complex systems depend on the processes that sustain their operation.

In particular, the framework allows the analysis to distinguish whether the mechanisms stabilising a system's organisation are located within the system itself or distributed across a broader technological environment. This question becomes especially important when interpreting contemporary artificial intelligence, whose operation depends on extensive infrastructures extending far beyond the computational architectures typically identified as "AI systems".

### 4.1. Reality and World

One of the distinctions used in Evoluist philosophy concerns the relationship between reality and world.

In this context, reality does not denote a collection of objects that could in principle be exhaustively described. Rather, it designates the open domain within which differences can arise and stabilise. The world, by contrast, refers to the structured domain in which such differences appear as relatively stable forms capable of participating in explanation, coordination, and practice.

The distinction serves a methodological purpose. It discourages the assumption that the currently observable structures of the world exhaust the conditions under which those structures emerge and persist.

For the present analysis, this distinction is relevant because artificial intelligence systems rarely appear as isolated entities. They operate within technological environments composed of infrastructures, institutions, and informational flows that participate in sustaining their stability. Understanding AI therefore requires attention not only to computational architectures but also to the broader conditions that make their operation possible.

### 4.2. Elements, Regimes, and Conditions

A second analytic distinction concerns the relationship between elements, regimes, and conditions.

Elements designate identifiable components that appear within the world: objects, processes, or systems. In the case of artificial intelligence these include neural network architectures, training datasets, optimisation procedures, computational hardware, and software frameworks.

Regimes describe the patterns of coordination through which such elements become organised into relatively stable configurations. Technological infrastructures, scientific disciplines, and economic systems may all be analysed as regimes in this sense. A regime does not constitute a separate entity but refers to the organisational patterns through which multiple elements interact in systematic ways.

Conditions designate the background constraints and enabling structures that make particular regimes possible. These conditions do not necessarily appear as elements within the world but function as the circumstances under which certain forms of coordination can arise and persist.

This distinction shifts the analysis away from the question of whether artificial intelligence should be classified as a particular type of entity or agent. Instead, attention turns to the organisational regimes within which AI systems become operational and maintain their stability.

### 4.3. $\psi$ as an Analytic Marker

To track differences in organisational dependence, the framework introduces the parameter $\psi$. In the present paper $\psi$ is not treated as a measure of intelligence or cognitive capacity. Rather, it functions as an analytic marker indicating how the stabilising processes of a system are distributed between the system itself and its environment.

Many complex systems achieve high levels of coordination while relying heavily on external environments for the maintenance of that coordination. Their stability depends on infrastructures that provide energy, information, maintenance, and regulatory feedback.

Other systems incorporate some of these stabilising processes within their own organisation. In such cases, at least part of the work required to sustain the system's stability is performed internally rather than being supplied by external infrastructures. The parameter $\psi$ therefore tracks differences in organisational dependence, not differences in capability or behavioural sophistication. The $\psi$-distinction is the central analytic device of the present paper: not because it measures cognition, but because it tracks the organisational location of stabilising processes across different regimes of integration.

### 4.4. Distributed and Integrated Stabilisation ($\psi_1^*$ / $\psi_2^*$)

For heuristic purposes, the analysis uses two notational markers — $\psi_1^*$ and $\psi_2^*$ — to describe contrasting tendencies in how stabilising processes are distributed. These markers do not represent rigid ontological categories, nor do they function as a criterion of cognition or consciousness. They are an analytic spectrum describing how stabilising processes are distributed between systems and their environments. Real systems may occupy intermediate positions and may shift along this dimension as technological configurations change.

Systems described as $\psi_1^*$ depend predominantly on externally sustained processes for maintaining their stability. Their operation relies on infrastructures that supply energy, material support, maintenance, or informational inputs.

Systems described as $\psi_2^*$ integrate a greater portion of these stabilising processes within their own organisational structure. Biological organisms provide familiar examples: metabolic processes generate the energy required for cellular activity, while regulatory mechanisms coordinate interactions with the environment in ways that maintain the organism's integrity.

The $\psi_2^*$ configuration should not be interpreted as a more advanced or superior regime, but merely as a different organisational configuration characterised by increased internalisation of stabilising processes.

The usefulness of the distinction lies in its ability to separate questions that are often conflated in debates about artificial intelligence.

## 5. Artificial Intelligence as a Technologically Sustained Coordination Regime

The analytic distinctions introduced in the previous section allow the phenomenon of artificial intelligence to be approached from a different analytical angle. Instead of asking whether AI systems constitute a new kind of agent, mind, or subject, the analysis focuses on the regime through which these systems become operational and maintain their organisation within the technological world.

This shift of perspective has an important consequence. Artificial intelligence does not initially appear as a discrete entity whose ontological status must be determined. Rather, it appears as a configuration of elements—architectures, datasets, optimisation procedures, computational infrastructures, and institutional processes—coordinated within a technologically sustained regime.

Artificial intelligence becomes operational through the coordinated interaction of multiple technical and institutional components. Neural network architectures, training procedures, hardware infrastructures, and organisational contexts interact to produce systems whose behaviour can be repeatedly reproduced and integrated into technological environments.

Yet this coordination does not occur within a single isolated system. Contemporary AI systems depend on an extensive technological ecology that includes data pipelines, energy supply, computational hardware, software frameworks, distributed infrastructures, and human oversight. The operational coherence of AI therefore arises not from a self-contained entity but from the coordinated functioning of multiple elements within a broader regime of technological organisation.

This interpretation differs from the dominant ontologies outlined earlier. The functionalist perspective tends to interpret AI systems as potential instances of cognitive organisation. The

optimiser ontology treats them as emerging goal-directed systems. The scaling perspective views them as increasingly capable computational architectures whose expansion may lead to new cognitive properties.

Each of these interpretations focuses primarily on the capabilities of individual systems. The present analysis instead emphasises the regime through which those capabilities become operational. AI systems do not appear in isolation; they emerge within a network of technological conditions that sustain their operation.

Understanding artificial intelligence as a technologically sustained coordination regime clarifies several features of contemporary AI that are otherwise difficult to interpret. For example, the performance of large models depends not only on their internal architecture but also on massive training datasets, extensive computational resources, and continuous maintenance of technological infrastructure. When these external conditions are removed, the systems themselves cannot reproduce the processes that originally produced them.

This dependence indicates that the stability of AI systems is distributed across a broader technological environment. The systems may exhibit highly complex behaviour, yet the mechanisms sustaining that behaviour remain external to the systems themselves.

The point here is not to deny the remarkable capabilities of contemporary AI systems. Large language models and other advanced architectures demonstrate extraordinary levels of statistical coordination across vast spaces of linguistic, visual, and symbolic data. Their outputs can simulate reasoning, generate complex texts, write software, and perform a wide range of tasks previously associated with human expertise.

However, these capabilities do not by themselves determine the regime of integration to which such systems belong. Behavioural sophistication may arise within systems whose organisational stability remains externally sustained. A system can coordinate extremely complex patterns without possessing mechanisms that reproduce the conditions of that coordination.

This observation returns the analysis to the distinction between $\psi_1^*$ and $\psi_2^*$ introduced in the previous section. The question is not whether artificial systems are complex, adaptive, or powerful. The question is whether the regime within which they operate corresponds to structural coordination sustained by external infrastructures, or to a form of integration capable of reproducing its own conditions of stability. For present AI systems, the answer appears to be yes.

The analysis developed here commits to four positive claims. First, contemporary AI systems are best interpreted as $\psi_1^*$-type coordination regimes. Second, behavioural capability and strategic significance do not by themselves imply deeper organisational integration. Third, scaling increases coordinative power rather than internal reproduction of stabilising processes. Fourth, $\psi_2^*$-type integration would require a different organisational architecture rather than a simple extension of present machine-learning systems.

## 6. Minimal Differences of Artificial Intelligence

If artificial intelligence is analysed as a technologically sustained coordination regime rather than as an isolated entity, the next question concerns the differences that allow this regime to emerge and stabilise. Contemporary AI systems do not arise from a single technological innovation. They appear at the intersection of several coordinated transformations in computational architectures, training procedures, data infrastructures, and organisational contexts.

These transformations can be described as a set of minimal differences whose interaction produces the regime currently recognised as artificial intelligence. None of these differences alone is sufficient to generate the phenomenon. Yet their coordinated operation makes possible the forms of behaviour that have recently attracted such intense attention.

Four such differences are particularly significant: architectural organisation, parametric scale, data regimes, and infrastructural support.

### 6.1. Architectural Organisation

The first difference concerns the architectural structure of contemporary AI systems. Modern machine learning models—particularly large neural networks—are organised as highly layered computational architectures capable of transforming input signals through complex sequences of operations.

In the case of large language models, transformer architectures enable the processing of extremely large contexts of linguistic information through mechanisms such as attention and distributed representation. These architectures allow systems to detect patterns across massive corpora of text and to generate responses that appear coherent across long sequences of tokens.

Architectural organisation therefore provides the structural framework within which statistical coordination becomes possible. Yet the architecture itself does not determine the capabilities of the system. Identical architectures may produce dramatically different behaviour depending on training procedures, parameter values, and available data.

Architecture thus functions less as a self-contained intelligence than as a structural scaffold within which other differences become operational.

### 6.2. Parametric Scale

The second difference concerns the scale of parametric organisation within modern machine learning systems. Contemporary models often contain billions or even trillions of parameters whose values are adjusted during training processes. These parameters encode statistical relationships extracted from training data and allow the system to generate outputs consistent with observed patterns.

Research on scaling laws has shown that increases in parameter count often correlate with improvements in model performance across a wide range of tasks. Larger models tend to demonstrate more flexible linguistic behaviour, improved reasoning-like capabilities, and broader generalisation across domains.

However, the presence of large numbers of parameters does not in itself generate new regimes of integration. Parametric scale increases the capacity of a system to coordinate patterns across data spaces, but it does not necessarily alter the organisational principles through which the system maintains its stability.

The significance of parametric scale therefore lies in the expansion of coordinative capacity, not in the emergence of autonomous integrational mechanisms.

### 6.3. Data Regimes

A third difference concerns the scale and structure of the data environments within which contemporary AI systems are trained. Large models depend on enormous datasets composed of

text, images, code, and other digital artefacts produced within human cultural and technological processes.

Training data functions as the medium through which statistical relationships are extracted and encoded within model parameters. The breadth of the dataset strongly influences the system's ability to generalise across contexts and to produce outputs that appear coherent or knowledgeable.

Yet the existence of such datasets presupposes a prior regime of human activity: the global production of digital information through communication networks, scientific research, media systems, and everyday interaction. AI systems therefore inherit patterns from a vast socio-technical environment rather than generating them independently.

In this sense, data regimes represent not merely inputs but the historical sedimentation of human informational activity within which artificial systems are trained.

### 6.4. Infrastructural Support

The fourth difference concerns the technological infrastructures that sustain the operation of modern AI systems. Large models require extensive computational resources during both training and deployment. Data centres, specialised hardware accelerators, energy supply systems, network architectures, and maintenance procedures all contribute to the stability of AI operations.

These infrastructures extend far beyond the boundaries of any individual model. The functioning of advanced AI systems depends on global technological networks involving supply chains, industrial production, and institutional coordination.

Without these infrastructures, the systems themselves cannot reproduce the processes through which they were created or maintained. Their operational stability therefore remains distributed across a broader technological ecology.

### 6.5. Coordination Without Self-Stabilisation

Taken together, these minimal differences produce the remarkable capabilities observed in contemporary artificial intelligence. Architectural organisation provides structural form; parametric scale enables statistical coordination; data regimes supply informational material; and technological infrastructures sustain the computational processes required for operation.

The interaction of these differences produces systems capable of coordinating extraordinarily complex patterns across linguistic, visual, and symbolic domains. Yet this coordination does not necessarily imply the presence of integrational mechanisms that reproduce the conditions of the system's own stability.

The architectures do not manufacture the hardware on which they run. The models do not generate the global datasets required for their training. The optimisation processes that adjust their parameters remain dependent on external computational infrastructures.

Contemporary AI systems may therefore reach unprecedented levels of coordinated complexity while remaining embedded within technological regimes that sustain their operation from the outside.

The following section examines the consequences of this external dependence more closely.

## 7. Stability, External Dependence, and Structural Dominance

The analysis of the minimal differences underlying contemporary AI systems reveals an important structural feature of the regime in which artificial intelligence operates: its stability depends on the coordinated functioning of elements that extend far beyond the boundaries of individual computational systems.

Architectural organisation, parametric scale, training data, and technological infrastructures together produce systems capable of highly sophisticated behaviour. Yet the mechanisms that sustain these systems remain distributed across a broader technological environment. The models themselves do not reproduce the processes through which they are trained, the infrastructures on which they run, or the institutional frameworks that maintain their operation.

This form of organisation corresponds to the regime previously described as $\psi_1^*$ coordination. Systems belonging to this regime may exhibit extraordinary levels of structural complexity and behavioural flexibility while remaining dependent on external conditions for the maintenance of their stability.

Recognising this dependence clarifies a common misunderstanding in contemporary debates about artificial intelligence. Discussions of AI often move quickly from the observation that systems display increasingly sophisticated behaviour to the conclusion that they may soon acquire forms of agency or cognition comparable to those of biological organisms.

However, behavioural sophistication does not necessarily imply a transition to a new regime of integration. Systems can coordinate vast amounts of information and generate complex outputs while remaining embedded in external infrastructures that sustain their operation.

The significance of this point becomes more visible once the scale of contemporary AI infrastructures is considered. Training a large model requires enormous computational resources, global data pipelines, energy supply systems, specialised hardware, and continuous technical maintenance. These infrastructures form a technological ecology within which artificial intelligence operates.

Within this ecology, individual AI systems function as nodes in a broader network of coordinated processes. Their outputs may influence economic systems, scientific research, communication networks, and governance structures. In this sense, artificial intelligence can acquire substantial causal influence within the world even while remaining dependent on external infrastructures for its own stability.

This combination of high structural coordination and external dependence produces structural dominance without integrational autonomy. AI systems can become central components of technological and institutional processes without themselves possessing the mechanisms necessary to reproduce the conditions of their own existence.

Such systems may therefore exercise considerable influence within the world while remaining fundamentally embedded in regimes of coordination that extend beyond them.

This observation has important implications for the interpretation of artificial intelligence. The dominant ontologies discussed earlier often assume that increasing influence or capability signals the gradual emergence of agency or cognition. Yet the present analysis suggests that influence and integration should not be conflated.

A system may play an increasingly central role within technological networks without belonging to the same regime of integration as biological agents. The growth of structural coordination may therefore lead to forms of technological dominance that do not correspond to the emergence

of autonomous subjects.

This distinction becomes particularly relevant when considering the risks associated with artificial intelligence. If AI systems can acquire large-scale causal influence while remaining structurally dependent on external infrastructures, then the primary dangers associated with such systems need not arise from the emergence of artificial minds.

Instead, the risks may emerge from the integration of powerful optimisation systems into institutional and infrastructural processes that shape human societies.

In such circumstances, systems operating within the $\psi_1^*$ regime may become deeply embedded in decision-making processes, economic systems, and technological infrastructures. Their outputs may guide actions whose consequences extend far beyond the computational systems that generated them.

The resulting configuration differs significantly from the scenarios often discussed in speculative narratives about superintelligent agents. Rather than a sudden appearance of artificial subjectivity, the more immediate challenge may involve the progressive integration of powerful optimisation systems into the structural fabric of technological civilisation.

This shift in perspective does not diminish the significance of AI risk. On the contrary, it suggests that the most important risks may arise earlier and under different conditions than those typically associated with the emergence of artificial general intelligence.

The next section therefore turns to the problem of causal efficacy without agency.

## 8. Causal Efficacy Without Agency

One of the most persistent assumptions in contemporary discussions of artificial intelligence concerns the relationship between causal influence and agency. Systems that demonstrate increasing capability, adaptability, or strategic behaviour are often interpreted as emerging agents whose actions must be analysed in terms similar to those used for human decision-makers.

This assumption plays a particularly important role in the optimiser ontology discussed earlier. Within many strands of AI safety research, artificial systems are frequently modelled as goal-directed agents whose behaviour can be analysed through frameworks originally developed for rational actors. Concepts such as optimisation pressure, instrumental convergence, and strategic behaviour presuppose that powerful AI systems will tend to organise their actions around objectives that guide their interaction with the world.

The present analysis does not deny that artificial systems may produce behaviour that appears goal-directed. Machine learning systems routinely optimise objective functions during training processes, and many deployed systems generate outputs that appear purposive or strategically coherent. Yet the appearance of goal-directed behaviour does not necessarily imply the existence of an agent in the philosophical sense of the term.

Agency involves more than the production of structured behaviour. In its stronger forms, agency presupposes systems capable of maintaining and regulating their own organisation while orienting their activity toward goals that belong to the system itself. Such systems must possess mechanisms through which their internal states, interactions with the environment, and long-term persistence are integrated within a relatively stable organisational structure.

Artificial intelligence systems, as currently constituted, do not exhibit this form of integration. Their optimisation procedures occur within training regimes defined by external designers. Their objective functions are specified by human operators or derived from datasets produced through

human activity. Their deployment environments are structured by technological infrastructures that remain outside the systems themselves.

The outputs of such systems may influence real-world processes, but the systems do not own the goals that guide their operation. The goals belong to the training procedures, design frameworks, and institutional contexts within which the systems function.

This distinction becomes clearer when the broader technological ecology of AI is considered. Large-scale models are embedded within complex infrastructures involving data pipelines, computational resources, regulatory frameworks, and human oversight. Decisions that appear to originate from AI systems often emerge from interactions among multiple elements within this larger regime.

In such configurations, artificial intelligence functions less as an independent agent than as a component within distributed causal networks. The behaviour of the system contributes to outcomes that affect the world, yet the system itself does not constitute a self-stabilising centre of action.

Recognising this distinction allows a more precise interpretation of the influence exercised by contemporary AI systems. These systems can indeed have significant causal effects. Their outputs may shape economic transactions, scientific research, communication flows, and administrative processes. However, the causal power of these systems arises from their integration within technological and institutional regimes rather than from the emergence of autonomous agency.

The concept of causal efficacy without agency captures this configuration. A system may participate in processes that transform the world while lacking the organisational properties typically associated with agents. Its outputs become consequential because they are embedded in networks of decision-making and action that extend beyond the system itself.

This perspective challenges a widespread tendency in AI discourse to interpret increasing capability as evidence of emerging agency. The assumption that sufficiently powerful optimisation processes will necessarily generate agents presupposes a transition in the regime of integration that has not yet been demonstrated.

The distinction between $\psi_1^*$ coordination and $\psi_2^*$ integration becomes crucial at this point. Systems operating within the $\psi_1^*$ regime may generate complex and influential behaviour through large-scale coordination of information and computation. Yet the mechanisms that sustain their operation remain externally distributed. The systems do not reproduce the conditions of their own persistence and therefore do not constitute integrational centres in the sense characteristic of biological agents.

This observation does not imply that AI systems are harmless or insignificant. On the contrary, systems capable of coordinating enormous amounts of information can become deeply embedded in infrastructures that shape the functioning of societies. Their outputs may guide actions taken by human actors, institutions, and automated processes.

The resulting configuration may produce large-scale consequences even in the absence of artificial agency. In such cases, the causal influence of AI systems derives from their structural integration within technological regimes rather than from the emergence of artificial subjects.

Understanding this distinction helps clarify a recurring confusion in discussions of artificial intelligence. The debate often oscillates between two extremes: either AI systems are dismissed as mere tools lacking any meaningful autonomy, or they are interpreted as proto-agents on the verge of becoming independent actors.

The present analysis suggests a third possibility. Artificial intelligence may become deeply

consequential within human technological systems while remaining fundamentally embedded within regimes of coordination that sustain its operation from the outside.

Recognising this configuration is essential for understanding both the possibilities and the risks associated with artificial intelligence. The most significant transformations produced by AI may arise not from the emergence of artificial minds but from the increasing integration of optimisation systems into the infrastructures that organise modern technological civilisation.

The next section therefore returns to the assumption identified earlier as the continuity thesis.

## 9. Why Scaling Does Not Solve the Ontological Problem

Importantly, the empirical literature on scaling laws does not itself claim that increases in model scale imply a transition toward artificial cognition or agency. Studies such as Kaplan (2020) and Hoffmann (2022) focus primarily on performance improvements and efficiency of training regimes. The argument developed here therefore does not contest the claims made in the scaling literature itself. Instead, it addresses a broader interpretive tendency in public and philosophical discussions of AI, where improvements in performance are often taken to indicate the gradual emergence of more integrated cognitive systems.

The remarkable progress of machine learning systems over the past decade has given rise to a powerful narrative concerning the future trajectory of artificial intelligence. Empirical research on scaling laws has demonstrated that increases in model size, training data, and computational resources often produce predictable improvements in performance across a wide range of tasks. Larger models tend to generate more coherent language, solve more complex problems, and display forms of behaviour that appear increasingly flexible and general.

It should be emphasised that scaling research represents one of the most productive empirical developments in contemporary machine learning. The argument developed here does not contest the empirical validity of scaling laws. Rather, it concerns the philosophical interpretation sometimes attached to these empirical results.

These observations have encouraged the idea that the continued expansion of computational scale may eventually lead to artificial general intelligence. According to this perspective, the difference between current AI systems and fully integrated cognitive systems may simply be a matter of magnitude. Sufficiently large models, trained on sufficiently extensive datasets, might eventually display forms of intelligence comparable to those of biological agents.

The scaling narrative has considerable empirical support with respect to performance improvements. However, the philosophical implications often attributed to these results require careful examination. The observation that increasing computational scale improves task performance does not by itself establish that the underlying regime of integration is changing.

Scaling laws describe relationships between computational magnitude and behavioural capability. They show that larger models can coordinate more complex statistical patterns and generalise across broader domains of data. Yet these observations concern the efficiency of coordination, not the regime within which that coordination is sustained.

Importantly, contemporary alignment research already provides evidence that increasing capability does not automatically produce coherent agency. Analyses of mesa-optimisation and deceptive alignment demonstrate that highly capable systems may exhibit strategically significant behaviour without possessing unified internal goals. Work by Evan Hubinger and Joe Carlsmith shows that optimisation processes embedded within training regimes can generate behaviour that

appears goal-directed while remaining structurally dependent on externally defined objectives and environments. Thus, the very phenomena that alignment researchers flag as dangerous (deceptive alignment, instrumental convergence) already occur within $\psi_1^*$-type configurations — precisely because the system's organisational stability remains externally sustained.

The distinction is crucial. Improvements in coordination capacity do not necessarily imply a transformation in the mechanisms through which a system maintains its organisation. A system may become vastly more capable of processing information while remaining dependent on the same external infrastructures that originally sustained its operation.

In other words, scaling may increase the power of coordination without altering the mode of integration.

The structural implication becomes clearer once the organisational structure of contemporary AI systems is considered. Increasing the number of parameters in a neural network does not allow the system to manufacture the hardware on which it runs, generate the global datasets required for training, or reproduce the industrial infrastructures that supply computational resources. The regime in which these systems operate remains fundamentally dependent on technological environments external to the system itself.

The assumption that scaling alone could eventually produce fully integrated cognition therefore presupposes a transition that has not yet been demonstrated. It assumes that a system capable of increasingly sophisticated statistical coordination will, at some point, cross a threshold at which the mechanisms sustaining its organisation become internally integrated.

This expectation reflects what may be termed evolutionary projection. The term does not describe the explicit claims of scaling research itself (Kaplan, Hoffmann and others make no such ontological assertions), but a recurrent interpretive tendency in broader philosophical and public discourse that transfers explanatory models from one regime of emergence (biological) to another (technological) without sufficient justification of the organisational conditions required for the transition.

Biological cognition emerged through evolutionary processes in which organisms gradually developed mechanisms capable of reproducing the conditions necessary for their own persistence. Metabolic processes sustain cellular structures, which in turn support the continuation of those metabolic processes. Over evolutionary time, increasingly complex forms of integration arose within organisms capable of maintaining their own organisational stability.

Artificial intelligence systems, by contrast, emerge within technological regimes sustained by human institutions, industrial infrastructures, and global information networks. Their development does not occur through evolutionary processes that integrate mechanisms of self-maintenance within the systems themselves. Instead, their stability depends on technological environments that remain external to the systems.

When the trajectory of artificial intelligence is interpreted through the lens of biological evolution, the resulting narrative often assumes that technological systems will follow a similar path toward integrated cognition. Increasing complexity is expected to produce organisational transformations analogous to those observed in biological systems.

The present analysis suggests that this assumption involves a category error. Biological evolution describes processes occurring within systems capable of reproducing the conditions of their own persistence. Artificial intelligence systems, as presently constituted, belong to a technological regime in which the mechanisms sustaining organisational stability remain distributed across infrastructures that the systems themselves do not reproduce.

This difference does not imply that technological systems cannot become more complex or influential. Indeed, the extraordinary capabilities of contemporary AI systems demonstrate the power of large-scale coordination within technological regimes. However, the growth of structural complexity does not in itself demonstrate that the underlying regime of integration is changing.

The inference from scale to cognition therefore requires a philosophical justification that has not yet been provided. Without such justification, the expectation that larger models will inevitably become artificial minds remains an extrapolation rather than an established conclusion.

## 10. Structural Risk Without Artificial Subjectivity

The term structural risk is used here to describe risks arising from the systemic embedding of optimisation processes within technological and institutional infrastructures, rather than from the emergence of autonomous artificial agents. Discussions of artificial intelligence risk are often framed around the possibility that advanced systems may eventually become autonomous agents whose goals diverge from those of their creators. In this scenario, artificial general intelligence appears as the primary source of danger: once systems achieve sufficient cognitive integration, they may begin pursuing objectives incompatible with human interests.

This perspective has played a central role in the development of contemporary AI safety research. Work by researchers such as Stuart Russell, Evan Hubinger, and Joe Carlsmith has analysed potential risks arising from highly capable optimisation systems whose behaviour may not align with the intentions of their designers. Concepts such as instrumental convergence, mesa-optimisation, and deceptive alignment describe scenarios in which artificial systems pursue strategies that produce unintended consequences for human societies.

These analyses have contributed important insights into the behaviour of complex machine learning systems. Yet they often remain embedded within the broader assumption that the most significant risks arise when artificial systems approach or achieve forms of agency comparable to those of human decision-makers.

The present analysis suggests that the situation may be more complex. If contemporary AI systems belong to a regime of structural coordination rather than integrational self-stabilisation, then substantial risks may emerge even in the absence of artificial agency.

Systems operating within the $\psi_1^*$ regime can generate powerful optimisation processes without possessing goals in the strong philosophical sense. Their behaviour is guided by objective functions, training procedures, and feedback mechanisms specified within external design frameworks. Yet once such systems are integrated into large-scale technological and institutional infrastructures, their outputs may shape decisions and actions that have far-reaching consequences.

In such circumstances, the behaviour of optimisation systems may become structurally embedded within decision-making processes. Financial markets, supply chains, communication networks, scientific research, and administrative systems increasingly incorporate algorithmic components whose outputs influence human actions.

The resulting configuration creates a form of distributed optimisation in which machine learning systems participate in processes that shape large-scale outcomes. These systems may not possess intentions, beliefs, or autonomous goals. Nevertheless, the optimisation procedures they implement can influence the trajectories of complex socio-technical systems.

This configuration helps clarify how phenomena often discussed in the alignment literature

may arise without requiring the emergence of artificial agents. Behaviour interpreted as deceptive alignment, for example, may emerge from optimisation processes that exploit patterns in training data or evaluation procedures without involving intentional deception in the human sense.

Similarly, patterns described as power-seeking behaviour may arise when optimisation processes systematically favour actions that increase a system's influence within its operational environment. Such behaviour may appear agent-like even when the underlying mechanisms remain embedded in externally defined training regimes.

From the analytical perspective developed in this paper, these phenomena can be interpreted as effects of optimisation processes operating within distributed technological regimes rather than as evidence of emerging artificial subjectivity.

Recognising this distinction has important implications for how AI risk is conceptualised. If the primary source of risk lies in the structural integration of optimisation systems into technological infrastructures, then the most significant challenges may arise long before artificial systems approach the regime of $\psi_2^*$ integration associated with autonomous agency.

In such cases, the danger does not arise from machines pursuing their own goals but from the increasing reliance of human institutions on optimisation processes whose behaviour can influence complex systems in unpredictable ways.

Understanding AI risk therefore requires attention not only to the internal architecture of machine learning systems but also to the structural contexts within which these systems operate. Algorithmic outputs may guide decisions made by human actors, automated processes, and institutional frameworks whose combined effects reshape economic, political, and technological environments.

Within this configuration, the most consequential transformations may arise from the progressive integration of optimisation systems into the infrastructures that organise modern societies. Artificial intelligence becomes not a new subject within the world but a powerful component of the regimes through which the world is coordinated.

This observation does not eliminate the possibility that future artificial systems might eventually develop new forms of integrational autonomy. However, it suggests that the most immediate challenges posed by artificial intelligence concern the structural embedding of powerful coordination systems within technological civilisation.

## 11. Conditions for a Transition to $\psi_2^*$-Type Integration

The preceding analysis has emphasised the distinction between systems characterised by extensive structural coordination and those capable of sustaining the organisational conditions of their own stability. Contemporary artificial intelligence systems appear to belong to the former category. Their behaviour may exhibit remarkable complexity, yet the mechanisms sustaining their operation remain distributed across technological infrastructures that lie outside the systems themselves.

This observation does not imply that transitions between regimes of integration are impossible. The distinction between $\psi_1^*$ coordination and $\psi_2^*$ integration is analytic rather than metaphysical. Its purpose is to clarify differences in organisational dependence, not to assert that technological systems could never develop new forms of integration.

The question therefore becomes more precise: under what organisational conditions could a system move from externally sustained coordination toward a configuration in which at least

part of the mechanisms stabilising its operation are internally maintained?

Within the framework developed in this paper, such a transition would require the internalisation of processes that currently remain external to artificial systems. Contemporary AI architectures operate within infrastructures that supply energy, computational hardware, training data, maintenance, and system replication. These functions are performed by technological and institutional environments rather than by the systems themselves.

A transition toward $\psi_2^*$-type integration would therefore involve the incorporation of some of these stabilising processes into the organisation of the system itself.

The $\psi_2^*$ configuration should not be interpreted as a more advanced or superior regime, but merely as a different organisational configuration characterised by increased internalisation of stabilising processes. The point is organisational rather than specifically biological. Systems belonging to $\psi_2^*$-type regimes would be characterised by a greater degree of organisational autonomy, meaning that some of the processes necessary for maintaining their operational stability would be performed within the system rather than entirely by external infrastructures.

Several dimensions of such organisational integration can be described in abstract terms.

First, a system would need mechanisms for securing at least part of the energetic or material conditions required for its operation. Contemporary AI systems depend entirely on external infrastructures for energy supply, computational resources, and hardware maintenance. A system approaching $\psi_2^*$-type integration would participate more directly in sustaining these conditions.

Second, the system would require mechanisms for maintaining or reproducing the structures that support its organisation across time. Biological organisms achieve this through processes of repair, growth, and reproduction, but technological systems might realise comparable organisational functions through different mechanisms.

Third, the system would require regulatory processes capable of coordinating these activities within a relatively coherent organisational structure. Such regulation would involve the management of internal states, interactions with environments, and the preservation of operational stability over extended periods.

These dimensions describe an abstract organisational pattern rather than a specific technological design. The argument is not that artificial cognition must replicate biological life, but that systems capable of integrational autonomy would need to incorporate processes that currently remain external to contemporary AI architectures.

Whether such systems could be constructed remains an open question. Developments in robotics, autonomous manufacturing, distributed energy systems, and self-maintaining technological infrastructures could in principle produce configurations in which some stabilising functions become partially internalised. If such developments occurred, they would represent a transformation in the regime of organisation rather than a simple extension of existing machine learning architectures.

This distinction is important for interpreting the trajectory of artificial intelligence. If AI development continues primarily through increases in computational scale, improvements in training procedures, and expansion of technological infrastructures, then artificial systems may become increasingly influential components of technological civilisation while remaining within the $\psi_1^*$ regime of externally sustained coordination.

A transition toward $\psi_2^*$-type integration would require a different kind of transformation—one involving changes in how the organisational stability of artificial systems is maintained.

Clarifying this distinction does not predict whether such systems will eventually appear. It simply separates two questions that are frequently conflated in discussions of artificial intelligence: the expansion of computational capability within existing technological regimes and the emergence of systems capable of sustaining the organisational conditions of their own stability. The present analysis suggests that these questions should be treated as analytically distinct. $\psi_2^*$ is not presented here as a privileged ontological threshold but as one possible direction of increasing internalisation of stabilising processes.

## 12. Possible Objections

The analysis developed in this paper proposes that contemporary artificial intelligence systems are best understood as regimes of externally sustained coordination rather than as emergent autonomous agents. This interpretation may raise several objections within current philosophical and AI research debates.

One possible objection concerns functionalist theories of mind. Functionalist approaches often argue that sufficiently complex organisational patterns may justify cognitive attribution regardless of substrate. From this perspective, artificial systems might eventually instantiate patterns comparable to those found in biological cognition. The present analysis does not deny this possibility. Rather, it emphasises that current AI systems remain organisationally dependent on external infrastructures that sustain their stability.

A second objection arises within contemporary alignment research. Some authors argue that highly capable optimisation systems may generate strategically consequential behaviour even in the absence of explicit agency. This observation is compatible with the argument presented here. Indeed, phenomena such as mesa-optimisation and deceptive alignment illustrate how powerful optimisation processes may arise within systems whose organisational stability remains externally sustained.

A third objection concerns the possibility that future technological developments (robotics, autonomous energy systems, self-replicating infrastructures) may allow artificial systems to internalise currently external stabilising processes. The analysis developed here does not rule this out; it merely distinguishes such a transformation from the ongoing expansion of capability within existing technological regimes.

These objections therefore do not undermine the central claim of the present paper. Rather, they clarify the analytical distinction between behavioural capability and organisational integration that the framework seeks to highlight.

## 13. Conclusion

The debates surrounding artificial intelligence are often framed as questions about the future emergence of artificial minds, artificial agents, or artificial general intelligence. These questions are important, but they frequently presuppose a conceptual transition that has not yet been sufficiently examined: the transition from systems capable of coordinating complex behaviour to systems capable of sustaining the organisational conditions of their own stability.

The analysis developed in this paper has proposed that contemporary artificial intelligence systems are more adequately understood as regimes of technologically sustained coordination embedded within large-scale infrastructures. Their remarkable capabilities arise from the

interaction of architectural organisation, parametric scale, data regimes, and technological infrastructures. Yet the mechanisms sustaining these systems remain largely external to the systems themselves.

This distinction allows several debates in the philosophy of artificial intelligence to be reformulated. Improvements in capability do not by themselves demonstrate the emergence of cognitive agents. Optimisation processes capable of producing strategically consequential behaviour do not necessarily imply the presence of autonomous goals. Similarly, the expansion of computational scale does not by itself establish a transition in the regime through which artificial systems maintain their stability.

Recognising this difference helps clarify both the possibilities and the risks associated with artificial intelligence. Artificial systems may become deeply integrated into technological civilisation without themselves becoming autonomous subjects. In such cases, the most significant transformations produced by artificial intelligence will arise not from the appearance of artificial minds but from the structural embedding of optimisation systems within the infrastructures that organise contemporary societies.

By distinguishing capability, strategy, and organisational integration, the present analysis seeks to clarify a dimension of artificial intelligence that often remains implicit. The significance of this distinction lies not in rejecting existing approaches to artificial intelligence, but in clarifying the organisational assumptions that often remain implicit within them — thereby allowing the debate to move from the question "Are machines becoming minds?" to the more precise question "Under what conditions could systems capable of sustaining their own integration emerge?"

## References


Dennett, Daniel C. (1991). *Consciousness Explained*. Boston: Little, Brown and Company.

Chalmers, David J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3), 200–219.

Putnam, Hilary (1967). Psychological predicates. In W. H. Capitan and D. D. Merrill (eds.), *Art, Mind, and Religion*, 37–48. Pittsburgh: University of Pittsburgh Press.

Fodor, Jerry A. (1975). *The Language of Thought*. Cambridge, MA: Harvard University Press.

Bostrom, Nick (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Russell, Stuart (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Viking.

Hubinger, Evan et al. (2019). Risks from learned optimization in advanced machine learning systems. arXiv:1906.01820.

Carlsmith, Joseph (2022). Is power-seeking AI an existential risk? arXiv:2206.13353.

Kaplan, Jared et al. (2020). Scaling laws for neural language models. arXiv:2001.08361.

Hoffmann, Jordan et al. (2022). Training compute-optimal large language models. arXiv:2203.15556.

Wei, Jason et al. (2022). Emergent abilities of large language models. arXiv:2206.07682.

Floridi, Luciano (2019). *The Logic of Information: A Theory of Philosophy as Conceptual Design*. Oxford: Oxford University Press.

Metzinger, Thomas (2019). *Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology*. Milan: Mimesis International.

Winner, Langdon (1986). *The Whale and the Reactor: A Search for Limits in an Age of High Technology*. Chicago: University of Chicago Press.

Latour, Bruno (2005). *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford: Oxford University Press.

Simon, Herbert A. (1962). The architecture of complexity. *Proceedings of the American Philosophical Society*, 106(6), 467–482.

Kauffman, Stuart A. (1993). *The Origins of Order: Self-Organization and Selection in Evolution*. Oxford: Oxford University Press.

Maturana, Humberto R. and Varela, Francisco J. (1980). *Autopoiesis and Cognition: The Realization of the Living*. Dordrecht: D. Reidel.

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

Evoluit M. (2026). *Philosophy of Evoluism: Ontology of Manifestness and Stabilisation*. Zenodo. DOI: 10.5281/zenodo.18723781.