frontiers | Frontiers in Communication

TYPE Conceptual Analysis
PUBLISHED 19 May 2025
DOI 10.3389/fcomm.2025.1585321

OPEN ACCESS

EDITED BY
Gemma San Cornelio,
Fundació per a la Universitat Oberta de Catalunya, Spain

REVIEWED BY
Carsten Strathausen,
University of Missouri, United States
Lucas Fucci Amato,
University of São Paulo, Brazil

*CORRESPONDENCE
Benedikt Zönnchen
✉ zoennchen.benedikt@hm.edu

RECEIVED 28 February 2025
ACCEPTED 21 April 2025
PUBLISHED 19 May 2025

CITATION
Zönnchen B, Dzhimova M and Socher G (2025) From intelligence to autopoiesis: rethinking artificial intelligence through systems theory. *Front. Commun.* 10:1585321. doi: 10.3389/fcomm.2025.1585321

From intelligence to autopoiesis: rethinking artificial intelligence through systems theory

Benedikt Zönnchen [1]*, Mariya Dzhimova[2] and Gudrun Socher [1]

[1]Munich Center for Digital Sciences and AI, Hochschule München University of Applied Sciences, Munich, Germany, [2]Institute for Cultural Management and Media, University of Music and Performing Arts Munich, Munich, Germany

The rapid advancements in the field of artificial intelligence (AI) have reinvigorated profound debates on the nature of intelligence, consciousness, and communication. Large language models (LLMs), in particular, are at the center of these discussions, as they generate complex linguistic patterns and challenge the traditional distinction between machine computation and human understanding. While LLMs are often seen as highly advanced statistical systems that generate text based on probabilistic patterns, both laypeople and experts tend to attribute human-like qualities to them.This article analyzes AI, particularly LLMs, from a systems-theoretical perspective and examines the extent to which these models can be understood as autopoietic, operationally closed systems. Building on Luhmann's system theory, it is argued that classical Turing machines are not sense-making systems, as they lack both self-reference in the sense of re-entry and the ability to make contingent selections from possibilities. In contrast, artificial neural networks (ANNs) exhibit a novel, loosely coupled interaction with social systems, as they can extract patterns from societal communication. This form of coupling differentiates them from classical software and positions them as hybrid systems that, while lacking their own mental states, are nonetheless deeply embedded in the structures of societal meaning production. The paper argues that LLMs should neither be regarded as purely technical tools nor as genuine cognitive entities. Instead, it proposes understanding their functioning as a new form of artificial meaning production—not as independent thinking, but as a recursive reflection of socially shaped linguistic patterns. This perspective not only opens new insights into the relationship between humans and machines but also calls for a critical reflection on how AI technologies are transforming our understanding of communication and cognition.

KEYWORDS

systems theory, artificial intelligence, large language models, artificial communication, autopoiesis

1 Introduction

Recent advances in artificial intelligence (AI) have reopened philosophical questions about the nature of human beings. In particular, the widespread use and remarkable ability of large language models (LLMs) to mimic human language has reignited and intensified old debates about concepts such as intelligence, consciousness, mind, etc. As a result, LLMs are increasingly perceived as human-like—an effect reinforced by the use of metaphysically charged terminology (Shanahan, 2024).

This growing perception of human-likeness in AI systems has brought longstanding philosophical disagreements back to the forefront. The debates surrounding AI are deeply controversial and often hinge on differing metaphysical positions. Dualism (Robinson, 2023), functionalism (Fodor, 1980, 1981), illusionism (Frankish, 2016; Dennett, 1991), panpsychism (Brüntrup and Jaskolla, 2016), reductionist (Hemmo and Shenker, 2023), and non-reductionist physicalism (Murphy, 2013), among others, each yield vastly different conclusions about the possibility of attributing mental states to machines. Functionalists, for example, argue that mental states are independent of their substrate and result from their causal role (Rosengrün, 2021 p. 40) while panpsychists believe that all matter is also mental. In both cases, strong AI—that is, AI that acquires genuine cognitive abilities (Searle, 1980)—is possible but for different reasons. Some positions link intelligence to consciousness, claiming it requires subjective experience (McKenzie, 2024; Juliani et al., 2022); others equate it with the capacity for thought (Floridi, 2011; Gabriel, 2018); while still others decouple the two, suggesting that consciousness, though distinct, may correlate with certain advanced AI achievements (Chalmers, 2023).

While philosophers have long debated intelligence and consciousness, such questions were largely neglected in AI and computer science, a criticism that was expressed early on by Dreyfus (1972) and Dreyfus and Dreyfus (1986). Instead, practitioners focused on the observable and measurable results of computer artifacts, thus avoiding deeper discussions. However, from the work of Turing (2009) and McCulloch and Pitts (1943) to Hopfield (1982) and Hinton (1989), researchers have often avoided explicit definitions of intelligence, while still consistently comparing machine performance to human capabilities. The Turing test embodies this anthropocentric stance: it assesses a machine's ability to "think" by measuring whether it exhibits "intelligent behavior" comparable to that of a human.

The same seems to be true of current AI research, and the tradition of using humans and their abilities as a benchmark for the performance of computing machines continues, with Sam Altman, CEO of OpenAI, also frequently predicting that AI will soon surpass "human reasoning." In addition, philosophical questions such as "What is AI and what is AI compared to humans?" are becoming more important in current AI research after a long period of productive neglect. For example, AI researchers such as Nobel Laureate Hinton (2023) are openly speculating about machines that are not only intelligent but potentially conscious. Furthermore, new fields of research are emerging around these discourses, such as the research area known as human-centered artificial intelligence (HAI). The emergence of this new research seems plausible: The way AI is conceptualized and classified has a direct impact on human self-understanding and social structures of how humans are conceptualized and classified. As AI systems are increasingly seen as cognitive and intelligent entities, these shifts may redefine notions of humanity, agency and intelligence, and influence legal, ethical and societal frameworks. Understanding AI is therefore not just a matter of technological classification, but a crucial factor in shaping social interactions and institutional practices. Finally, AI is not just an artifact, a field of research or a tool—it is also a promise.

However, what is striking about both the old and the new debates on AI, is that despite their different positions, they all share a more or less anthropocentric perspective, taking humans and their capabilities as the starting point for either attributing or denying machine intelligence, consciousness, mind, etc. As the sociologist Niklas Luhmann noted, discussions of AI are deeply embedded in a long humanistic tradition that questions whether computers and their so-called "artificial intelligence" can ever be equated with human consciousness—or even surpass it. The humanities have historically focused on this issue, making it the real reference point (refuge) of research (Luhmann, 1998 p. 303). As he argues further:

> "[I]t remains questionable whether this is even the right problem to pose and whether the computer, in this competitive situation, will not sooner or later emerge as the winner, provided that society grants it 'equal opportunity'." (Luhmann, 1998 p. 303)

If we take Luhmann's comment seriously—and extend it in light of Heidegger's critique that modern thinking about technology remains bound to inherited metaphysical frameworks that obscure more originary modes of understanding (Heidegger, 1977)—then a shift in emphasis becomes necessary. Rather than clinging to the modern, human-centered paradigm of the subject/object distinction, we propose to frame the discussion in terms of a theory of contingency and difference, in particular the system/environment distinction. This posthumanist approach, we argue, allows for a sophisticated understanding of (generative) artificial intelligence. Furthermore, it also allows for the theoretical exploration of an autopoietic, non-sense-making, cognitive "artificial" system—a system whose defining features are exemplified by the operational mechanisms of large language models.

Against this background, this article analyses AI, in particular LLMs, from the perspective of Luhmann's systems theory. By introducing the main concepts of his theory, it attempts to examine the extent to which these models can be understood as autopoietic, operationally closed systems. Finally, it will conclude by offering new insights into the relationship between humans and machines, but also by calling for a critical reflection of how our concepts of cognition and communication are being transformed by AI technologies.

2 From subjects to systems

Since Kant (1781), reality has been understood as resistance—as something that challenges the mind and thereby demands explanation. When something fails to work, we are forced to construct models to account for the failure. Yet even Kant's distinction between the empirical and the transcendental leads back to the problem of the relation between "a reality that remains unknown" (Kant's *thing-in-itself*) and the subjective phenomenal appearance. From this ontological standpoint, an epistemological question remains: how can cognition recognize anything as external to itself, when all recognition presupposes cognition (Luhmann, 1988 p. 9)? This challenge lies at the heart of radical constructivism (von Glasersfeld, 1995) and also guides Luhmann's systems theory.

While Kant focused on the conditions of possibility for experience within the boundaries of subjectivity, Luhmann's theory redirects the focus away from the human subject. The reference to Luhmann's systems theory is interesting for a posthumanist view of AI in general and LLM in particular, as it extends the notion of cognitive systems beyond human consciousness (Luhmann, 1998 p. 122). For Luhmann, cognition no longer presupposes a mind (Möller, 2011 p. 39), but rather interprets cognition as a system-internal operation and distinguishes between sense-making and non-sense-making cognition.

2.1 Self-asserting systems

Like axiomatic mathematics, Luhmann begins as he writes, with a "naive starting point", namely the assumption "that systems exist" (Luhmann, 1988 p. 14). By adopting the concept of *autopoiesis* from Maturana and Varela (1987), Luhmann argues that a system produces and maintains itself through its own operations. In doing so, it distinguishes itself from its environment through a recursive demarcation of its operations, thereby acquiring its own identity as a system.

Systems, then, are not things—they are distinctions. Every system is both what it includes and what it excludes. The system/environment boundary is a form that exists only through the relation of its two sides (Luhmann, 1998 p. 63). Without this difference, there is no system. This distinction is not just logical; it is operationally necessary for self-reference and organization (Luhmann, 1998 p. 60). However, every distinction inevitably carries with it an implicit absence, what Spencer-Brown (1969) refers to as the unmarked space. Saying "tree" implies everything that is not tree. Furthermore, "the observer is the excluded third of their observation" (Luhmann, 1998 p. 69) and the distinction they use is the hidden condition, i. e., a necessary blind spot for seeing. To judge something as lawful, one must treat the distinction itself as lawful. To see, one must be blind to the act of seeing—one does not see the perspective and the difference through which one sees.

This logic equally applies to *psychic systems*. Thoughts are not isolated processes, but the means by which a psychic system stabilizes itself through recursive differentiation. No thought leaves the system, that is, psychic systems are *operationally closed*; meaning arises only in relation to other thoughts. Thinking does not determine in advance what comes next; it can only retrospectively interpret what has been thought. Each thought enacts distinctions—for example, between tree and non-tree when thinking "tree." Crucially, it is through the fundamental system/environment distinction—"I think (system)/this is not my thinking (environment)"—that an external world is constructed. Without this distinction, we could not differentiate thoughts from what they refer to. We would risk confusing a tree with a thought, or communication with what is being communicated.

The self-referential nature of consciousness aligns with the phenomenological tradition: consciousness is not an entity but a "stream of experiences" (Husserl, 1993), connecting only to itself and not open to external interruption. Adopting Husserl's concept of Sinn (sense), Luhmann defines *sense-making* as the selection of meaningful references from an excess of possibilities—there are always more meanings than can be used. Luhmann radicalizes this further: even the environment of a sense-making system appears only as sense, and sense can refer only to sense (autopoiesis), never to anything external (Buchinger, 2012). Paradoxically, even non-sense is a form of sense—sense that fails to make sense, and thus, still makes sense.

2.2 Choice and contingency

A major implication of operational closure is the problem of freedom, conceptualized as the ability to choose between alternatives. If mental and social systems operate through self-referential flows of thought or communication, how can choice be meaningfully exercised?

Luhmann addresses this question through the concept of *contingency*. Borrowing from Aristotle, Luhmann defines a contingent event as one that is neither necessary nor random, but could have unfolded differently (Möller, 2011 p. 45). Contingency refers to everything that exists "in the light of its possible variation" (Luhmann, 1984). All evolutionary events could have happened differently. Moreover, social evolution is an extraordinarily unlikely phenomenon.

Within the horizon of possible variations, thought itself remains a spontaneous yet contingent process. Unlike mechanical causality, operationally closed systems do not respond in a linear fashion to inputs, but respond to stimuli in unpredictable ways due to their internal structure, which "determines" how they adapt and reorganize themselves. While operational closure maintains a degree of causal dependence, it does not entail absolute determinism. Cognitive processes are neither entirely random nor fully predetermined; retrospectively, events appear necessary, but in the present moment, they remain contingent.

This interplay of contingency and system closure becomes even more pronounced in social interactions, where *double contingency* arises. Here two systems meet, each contingent in its own way. "I'll do what you want if you do what I want" (Luhmann, 1984 p. 166). Communication emerges as a way of dealing with this mutual uncertainty. It is unlikely, but it happens precisely because the lack of transparency forces the creation of shared meanings and expectations. Communication continuously creates structures and reduces uncertainty over time, with past communication creating historical contexts (memory) that structure future communication.

The most profound form of freedom appears in *re-entry* (Spencer-Brown, 1969), where a system reintroduces its own guiding distinction into itself. This keeps the system/environment boundary dynamic and open to reevaluation. For example, a system can reassess its own notions of *inside* and *outside*, generating new self-descriptions. Psychic systems can reflect on the criteria by which they deem aspects of reality "real" or "significant." Similarly, a social system like science can reapply its true/false distinction to reflect on its own methods.

The capacity for re-entry enables self-observation—a special case of what Heinz von Foerster and Margaret Mead called second-order cybernetics (von Foerster, 2003) and Luhmann *second-order observation*. It refers to a system's ability to observe how it and others observe reality. For Luhmann, this is a defining

characteristic of modern society. We no longer judge things in isolation: when buying a house, we consider market evaluations; in academia, researchers examine both findings and their reception. Even everyday actions such as shopping are shaped by evaluations and reputations. First-order observations are increasingly shaped by second-order reflections—we not only perceive objects, but also how they are perceived. Whereas Gabriel (2020 p. 34) observes that we "live our life in the light of an idea of who or what we are", we also increasingly live in the light of an idea of how others observe us.

Yet this recursive capacity is not universal. While psychic and social systems engage in *sense-making*, many systems do not. They process environmental stimuli but operate outside the medium of sense. Such systems cannot assign significance, as they lack second-order observation and the ability to reflect on their own distinctions—especially the difference between *self-reference* and *other-reference*, which is essential for observing an environment (Luhmann, 1998 p. 92). For instance, the immune system distinguishes between the body's own cells and foreign invaders, yet it does not observe itself making this distinction.

2.3 Structural coupling

Despite being operationally closed, systems do not exist in isolation. They remain autonomous while depending on other systems through *structural coupling*—a mechanism that allows interaction with a complex environment without processing its full complexity. For example, society communicates about "humans" even though psychic systems cannot define what a human is. Structural couplings limit the range of viable structures for autopoiesis, meaning every system is already environmentally adjusted (Luhmann, 1998 p. 100).

Further, structural couplings replace the notion of "human nature" by concentrating causal influences that perturb and reshape the systems involved. As Luhmann argues, adaptation is not the result of natural selection or cognition, but a precondition: systems can only build complexity within existing structural constraints (Luhmann, 1998 p. 102). Communication, for instance, has such high degrees of freedom only because it is only indirectly affected by the physical world via operationally closed brains and psychic systems (Luhmann, 1998 p. 114). Its emergence, therefore, is highly improbable.

Luhmann notes that structural couplings must transform analog relations into digital ones to allow environmental influence. Language fulfills this role between the psychic and social systems. Although the psychic system is operationally closed and guided by its own processes, language constrains some of its thinking while preserving a surplus of possibilities—enabled by the *medium of sense* shared by both systems (Luhmann, 1998 p. 101). Language allows these systems to relate to each other without direct access: thought can shape communication and communication can influence thought. But this is not a transfer of meaning (Luhmann, 1998 p. 73). Communication, Luhmann argues, would be impossible without psychic systems—even though psychic systems themselves cannot communicate (Luhmann, 1998 p. 103).

3 Language models as cognizing systems

Having clarified key terminology, we shift from the humanistic question—"Is AI intelligent?"—to a systems-theoretical inquiry: "Is (generative) AI autopoietic?" and "Does it operate in the medium of sense?" In other words, are language models *autopoietic systems* capable of *sense-making cognition*? Are they *operationally closed* and able to distinguish between *self-reference* and *other-reference*? Rather than the binary classification of intelligent/non-intelligent, we propose an alternative distinction grounded in systems theory. Within this framework, language models become particularly relevant not simply because of their functional output, but because they create the impression of "understanding language"—an effect that invites further examination of their potential coupling with psychic and social systems.

3.1 Technology as functional simplification

According to Luhmann, "natural technology"—such as celestial mechanics or tidal movements—serves as a paradigm for "artificial technology" due to its reliability and minimal deviation from expectations (Luhmann, 1990 p. 224). Although inherently more fallible, artificial technology enables the identification and correction of errors. Luhmann thus characterizes technology as *transparent*, insofar as its functional purpose is explicit: even when psychic systems lack full comprehension of its inner workings, they can still evaluate whether it functions as intended. In this sense, technology constitutes an *evolutionary achievement* that operationalizes complexity reduction (Luhmann, 1998 p. 517), transforming uncertainty into a form of manageable ignorance (Luhmann, 1998 p. 525).

Although embedded within communication, technology remains external to social systems, serving instead as a structurally coupled interface between society and the physical world that stabilizes or transforms communicative processes. Technical systems, while operating in communicative environments, typically do not themselves communicate: a data center or an operational *artificial neural network* (ANN) is not autopoietic but *allopoietic*—externally constructed, reliant on external inputs, and incapable of reproducing their own components.

Technological functioning hinges on the distinction between *controllable* and *uncontrollable* conditions. Unlike biological organisms, which maintain stability through *loose coupling*, technology relies on *strict coupling* to guarantee precision (Luhmann, 1998 p. 525), manifesting in the binary distinction *works* vs. *does not work*. This functional efficacy suppresses dissent by precluding alternative operations (Nassehi, 2019 p. 228). Yet, as complexity increases, technology becomes subject to the paradox of control: while built on causal attributions—selecting limited causes and effects—its growing intricacy renders causal mechanisms opaque, thereby generating emergent instabilities that demand further technological intervention. Luhmann already notes this recursive dilemma, echoed in Heinz von Foerster's notion of *non-trivial machines* (von Foerster, 2013 p. 86), which

behave unpredictably and undermine transparency, reinforcing the necessity of continual refinement.

The increasing complexity of non-trivial machines has led scholars such as Reichel (2011), Watson and Romic (2024), and Lovasz (2024) to conceptualize technology as a closed, potentially autopoietic system. Reichel (2011) frames technology as distinct from both society and individuals, governed by the binary code *work*/*fail* within the medium of *operativeness*. In this view, technology is neither physical nor social, but recursively enacts itself through the continual application of its own code, evolving via internal feedback while perturbing and co-evolving with its social environment. It evolves through recursive processes, irritates society, and co-evolves with social systems. Building on this, Watson and Romic (2024) interpret ChatGPT as an autopoietic subsystem of technology—an operationally closed mediator between cognition, society, and the physical world. Their focus lies on the role of such systems in education, inclusion, and ethical integration. Similarly, Lovasz (2024) further develops Reichel's perspective by integrating Ellul's theory of technological autonomy, arguing that high technology generates risk, unpredictability, and complexity, thereby approaching a form of quasi-autonomy beyond instrumental control.

While these accounts share a compelling interest and valuable insights into the systemic evolution and increasing complexity of technological systems, they lack a critical examination of how such systems might engage in *re-entry*—the reintegration of the system/environment distinction into the system's own operations which is essential for sense-making systems in Luhmann's theory, as it underpins their capacity for reflexivity, second-order observation, and self-description. Reichel's notion of technological autopoiesis, for instance, does not demonstrate how technology reflects upon and internally operationalizes its boundaries as a meaning-constituting system. Likewise, Watson and Romic (2024) do not clarify how (sub)systems like ChatGPT would engage in second-order observation or generate their own self-descriptions. Lovasz (2024), while more cautious, aligns with Ellul's view of technological sovereignty but similarly leaves unaddressed the issue of internally generated systemic distinctions. In each of these cases, the invocation of autopoiesis appears more metaphorical than formally systemic. While these theories successfully highlight the recursive complexity and increasing autonomy of technological systems, they fall short of demonstrating the reflexive, self-producing and systemic boundary-construction—hallmarks of autopoietic systems of high complexity.

Rather than generalizing technology as autopoietic, we adopt a narrower approach by examining a specific domain: *artificial neural networks* (ANNs). Our aim is not to classify technology as a whole, but to investigate whether and how certain types of computations exhibit features associated with autopoietic systems—such as operational closure, recursive complexity, and structural coupling—without necessarily fulfilling the criteria for full autopoiesis. We also reject spatial definitions of systemic boundaries. As Luhmann stresses, what constitutes a system is not physical separation, but the production of distinctions between system and environment (Luhmann, 1998 p. 66). While organisms may have spatial boundaries, systems such as consciousness, the economy, or computation do not. Our analysis thus aims to explore how ANN operations instantiate system-like dynamics—complexity, contingency, and relative autonomy—without presupposing full-fledged autopoiesis.

3.2 The self-referentiality of Turing machines

The theoretical construct of the (universal) Turing machine (Turing, 1937), foundational to computer science and AI, exemplifies self-referential computation. It is a formal abstraction that can be described within a mathematical language as a digital pattern (information) while simultaneously being realized in hardware to transform digital patterns (programs and inputs) into new patterns (outputs). As a materialized computer or program, it differs from simple calculators by its ability to branch computations based on intermediate results. A universal Turing machine (the computer) can receive a description of a Turing machine (a program) as input and execute it.

This recursive relation of computation extends beyond individual machines. Software development is typically facilitated by software environments that translate human-written code into machine-executable instructions, effectively allowing Turing machines to construct other Turing machines. If we conceptualize these realized computational operations as a system and place psychic and social systems in the environment of machine computations, it follows that Turing machines—both as digital patterns and as their physical realizations (hardware)—can be understood as a self-reproducing system. A recursive structure emerges: machines, once programmed, can generate new machines, creating a recursive loop of code and computation. Using Luhmann's extended concept of cognition, we can observe a *basal self-referentiality* within computational operations.

However, from a different perspective, software is not genuinely generated by machines but by human developers,[1] who encode meaning through formal languages. This deductive process relies on logical calculi and mathematical syntax—structures that are essentially *transparent*. While psychic systems imbue code with meaning by selecting among alternatives, Turing machines act purely as symbol manipulators or, in Luhmann's terms, generators of medium/form distinctions within the "medium of computational operations" (Luhmann, 1992 p. 399). Even the notion of code "interpreters" remains metaphorical, as no meaningful decisions occur within the machine itself—it lacks contingency and could not have acted otherwise.

From this, we infer that while Turing machines exhibit *cognition*, they do not perform *sense-making cognition*, which presupposes contingency. As Luhmann argues, the contingency of meaning is a "necessary moment of meaningful operations" (Luhmann, 1998 p. 55). In contrast, psychic systems satisfy this criterion; machines do not. This distinction is reinforced by Hartmann (1992), who describes machine processing as a

1 This situation is currently becoming more complex as language models intervene in the programming process, meaning that machines not only transform but also generate code. We will disregard this for now.

"destruction of the meaningful reference horizon", and human-machine interaction as a "reduction and systematization of reference horizons":

> "Machines do not recognize complexity because, for them, there are no more possible environmental references than those currently being actualized. Only the human perspective on the machine or its produced output restores complexity." (Hartmann, 1992 p. 253)

Machines—whether as abstract algorithms or material implementations—lack the capacity to connect distinct possibilities; they are not "free" but follow a fixed, externally programmed sequence. They combine *input openness* with *operational determinism*. Thus, while programming languages serve as media through which humans specify machine behavior, systems theory reframes this as a *strict structural coupling* between psychic and computational systems via formal languages.

The rigidity of this coupling becomes most apparent when classical machines yield unexpected outcomes. Users and developers do not interpret such deviations as creative, but as malfunctions. A clock showing the wrong time or a failed digital transaction is deemed broken—not innovative. Since classical machines operate under strict causal constraints, their computations—however complex—can be retrospectively analyzed for correctness.

This reveals a crucial limitation: although *computational systems* exhibit a form of *basal self-referentiality*, this alone does not suffice for meaning construction. As Luhmann argues, sense-making requires not only cognition but the capacity to distinguish *self-reference* from *other-reference* via *re-entry*—the internal reapplication of the system/environment distinction. Classical machines lack this ability; their operations are recursive but non-reflexive. They cannot observe or differentiate their functioning from an external context because they lack an internal model of what is "inside" and what is "outside."

This brings us to the next question: do the operations of contemporary, self-programmed machines—particularly artificial neural networks (ANNs)—differ in this regard? Can they, unlike classical machines, engage in *sense-making cognition*? To explore this, we now turn our focus to a specific type of ANN: *large language models*.

3.3 The loose coupling of artificial neural networks

While the individual computational operations of an artificial neural network (ANN) remain traceable, the process by which these operations are generated is inherently *intransparent*, and this opacity extends to the network's operations itself. Unlike traditional algorithms, which are explicitly designed by developers, ANNs construct their computational structures from data through multi-objective optimization algorithms. Similar to a Turing machine, these optimization algorithms remain strictly coupled to the psychic systems of developers via formal languages. However, the crucial difference lies in the nature of the data: it contains patterns of communication that psychic systems themselves can not make sense of.

This marks a fundamental shift. While the output of an ANN is theoretically reproducible for a given input (factoring in pseudo-randomness), the genesis of its computational operations remains opaque due to the *contingency* of its training data. This data, shaped by the selections of a contingent society, is neither necessary nor impossible—it is simply improbable in its current form. Its structure does not reflect an "objective" reality historically evolved observational selections. As Luhmann argues, data—like sense—is not discovered but constructed by systems that differentiate themselves from their environment.

Factoring in data contingency, the operations of an ANN become contingent themselves. Therefore, data-driven computation results in a decoupling of *computational systems* from *psychic systems* (e. g., the developers' minds), which no longer engage with formalism and protocols in the same way—making it increasingly difficult to construct sense from the computational operations of ANNs.

At the same time, the operations of ANNs form a new type of structural coupling with social systems—not through direct interaction, but by *recognizing* patterns in digitized communication within their training data. Unlike classical machines, which are strictly coupled to psychic and social systems via formal languages, ANNs introduce a second, looser coupling through the "digestion" of training data derived from memorized communcation of social systems. In this way, they parasitically "absorb" social contingencies. In other words, they integrate perturbations on their own terms, that is, computationally.

Simultaneously, the first form of coupling—strict coupling via formal language—appears to be diminishing, as language models intervene in programming itself, mediating or even partially replacing the formal language interface that traditionally structured machine-human interaction. This means that the boundary between programming and execution is becoming increasingly fluid, as ANNs contribute to the generation of code, blurring the role of formal languages as an exclusive medium between psychic systems and computational systems.

From this, we can conclude that for self-learning systems such as ANNs, contingency no longer ends with programming. Unlike classical machines, ANNs absorb traces of societal contingency through their training data. This gives rise to a new form of structural coupling—less rigid, more emergent—between computational systems with *limited choices* and social communication. In a sense, society has already partially made the choices that ANNs parasitically internalize, process, and then project back into communication—a belatedness that can also be observed between brains and psychic systems. This dynamic grants ANNs a form of *computational autonomy*—one distinct from the autonomy of psychic and social systems. However, this autonomy remains incomplete, as ANNs (as computations) cannot (yet) reflect on their computations or the solutions they compute. Instead, they are supplied with reflection by psychic and social systems.

Ultimately, state of the art ANNs do not "know" what they are doing (Heßler, 2019). As computational operations, they do not engage in contingent sense selection. They generate decisions but

cannot interpret them meaningfully, as they remain dependent on predefined objectives—such as generating plausible text. While this might suggest a break from traditional programming, the objective remains the same: a strict coupling between psychic systems and machines through the medium of formal language. As Esposito already pointed out, this introduces a paradox which is relevant in practical applications of ANNs: the goal is to control the lack of control (Esposito, 1997).

3.4 From neural networks to language models

This paradox becomes even more pronounced in a specific subclass of ANNs: *large language models* (LLMs). These systems are optimized not merely for general computation but for interacting with the core medium of both psychic and social systems. As such, they offer a particularly rich case for analyzing how computational systems structurally couple with communication processes and potentially participate in the generation of meaning.

Language models such as OpenAI's GPT models (Bubeck et al., 2023; Radford et al., 2019), Claude (Anthropic, 2019), LLaMA (Touvron et al., 2023; Grattafiori et al., 2024), and DeepSeek (DeepSeek-AI et al., 2024, 2025) are specialized forms of artificial neural networks (ANNs) designed for natural language processing. As such, they are structurally coupled with both psychic and social systems through their interaction with language. By influencing patterns of communication and shaping linguistic practices across these domains, large language models (LLMs) represent a particularly significant site for examining how technological operations intersect with meaning production in complex systems.

These models use self-attention mechanisms (Vaswani et al., 2017) to weigh the importance of tokens in a sequence. Operating autoregressively, they predict the next token based on previous ones, learning from vast textual corpora through self-supervised learning (SSL), which adjusts model parameters by minimizing the difference between predicted and actual tokens (Zhou et al., 2021). This process results in a probabilistic rather than deterministic generation of text.

After training via self-supervised learning a rather strict coupling exists between training data, i. e., digitalized and memorized communication and the model's internal representations, via reinforcement learning [especially reinforcement learning from human feedback (RLHF) (Ziegler et al., 2020)] introduces an additional optimization layer which loosens this coupling.[2] Instead of direct token-to-token supervision, models adjust based on reward signals that align outputs with human expectations (Cao et al., 2024). However, RLHF presents challenges, as discussed by Casper et al. (2023), since goal-setting remains externally imposed.

In their use, language models do not follow explicit rules but structure probabilities in a high-dimensional semantic space, computing plausible continuations of input based on statistical distributions. Tokens are mapped into numerical vectors, transformed, and finally converted back into text. In the final step, a distinction is drawn between *plausible* and *implausible*, based on a rich representation in the form of a vector within the latent space.

As explained before, these models do not encode precise computational instructions like programming languages. Instead, they generate responses through stochastic logic, making their outputs inherently contingent—shaped by training data rather than deterministic algorithms. This gives rise to a distinct form of structural coupling with communication and psychic systems. Unlike programming languages, which explicitly determine a machine's operations, prompts do not encode precise computational instructions. Consequently, the outputs of language models remain fundamentally *contingent*, determined by both their statistical training and probabilistic inference rather than by deterministic rules. However, this contingency should not be mistaken for conscious decision-making. As Shanahan (2024) notes, and as we will see, "[a LLM] is not in the business of making judgments. It just models what words are likely to follow other words."

3.5 Meaning without reference?

"Language models do not generate meaning." On this point, we agree with Bender and Koller (2020) and Shanahan (2024), but not because LLMs lack access to "the real world" or "the truth [...] against which they could compare the words they generate" (Shanahan, 2024). From a systems-theoretical perspective, such an argument is unconvincing if cognition is assumed to function through operational closure. Sense, in this view, does not emerge from an external reality but from the system's own operations.

A central critique by Bender and Koller (2020) is encapsulated in their *octopus test*, which draws on Frege's premise that meaning is at least partially determined by reference to an external object (Frege, 1892). Similarly, Shanahan (2024) argues that human speakers, unlike LLMs, can "consult the world" to resolve inconsistencies and refine their assumptions. However, this reintroduces a human-centered perspective and, in doing so, highlights the paradox of cognition: no system can fully recognize what lies beyond its own boundaries.

From the standpoint of operational constructivism, as developed by Luhmann, sense does not require an external referent but emerges through the interplay between actuality (what is currently being communicated) and possibility (what could have been or could be communicated). In this framework, resistance is not imposed by an external reality but relocated *within* the system itself:

> "I think we should not abandon [Kant's] idea of resistance, but we should relocate it into the system. It is the result of resolving an internal conflict—the result of the system's operations resisting the operations of the same system." (Luhmann, 1995)

Due to the removal of all traces of operational closure under first-order observation, the experience of resistence and

2 Luhmann (1990, p. 183) describes loose couplings as characteristic of systems that buffer environmental perturbations through internal complexity—a definition that aligns well with how LLMs operate on contingent, pre-structured data.

the non-arbitrariness of the operations' results are assigned to an external world (Luhmann, 1998 p. 93).

Nonetheless, Bender and Koller (2020) correctly emphasizes that it is highly improbable that an operationally closed, sense-making computational system would construct sense in the same way as psychic or social systems. Because contingent co-evolutionary systems follow distinct operational logics, they are unlikely to construct similar meanings. Likewise, Shanahan's argument concerning LLMs' lack of embodiment (Shanahan, 2024) is valid: without structural coupling to a living body, i. e., a living system, LLMs lack key perturbation that shape human sense-making.

However, dismissing LLMs as mere "stochastic parrots" (Bender et al., 2021) underestimates their functional capabilities. While it is true that LLMs are trained to predict the next word based on statistical distributions, this does not mean their post-training processing is reducible to simple string matching. For example, Li et al. (2023) provide evidence of an emergent non-linear internal representation of board states when an LLM is trained on sequences of chess moves. More broadly, research suggests that models trained on textual corpora can develop internal structures that capture key elements of interpretation (Piantadosi and Hill, 2022; Søgaard, 2023; Sahlgren and Carlsson, 2021).

Expanding on this view, Piantadosi (2024) argues that in LLM training, semantics and syntax are not separate entities but are integrated, allowing word relationships and contextual roles to shape predictions. Unlike traditional generative grammars that impose syntactic rules, LLMs infer latent structural relationships probabilistically (Piantadosi, 2024). Words (or tokens) are encoded as vectors in a high-dimensional space, where their positions reflect semantic relationships. These relationships shift dynamically based on prior generated text. Notably, the model's internal states capture latent aspects of both syntax and semantics, enabling it to reconstruct tree structures (Manning et al., 2020). In this way, meaning does not exist in isolation but emerges relationally, forming what could be described as a "contingent index-card system (Zettelkasten)" of linguistic associations and relations.

This perspective aligns with conceptual role theory (Block, 1996), which views meaning as relational and functional and also with Luhmann's theory which claims that "objects" are never given things in an external world but structural entities of the autopoiesis of the system, i. e., conditions for continuation (Luhmann, 1998 p. 99), making it plausible that language models partially align with meaning as constructed by psychic and social systems (Piantadosi, 2024).

However, this apparent alignment requires further conceptual clarification—particularly in relation to *sense production*. Piantadosi's (2024) notion of *semantic meaning* differs fundamentally from Luhmann's concept of sense (Sinn). Whereas, sense-making via the unversial medium of sense functions as a structural principle within systems, semantics is a historically evolved form of sense that takes shape within communication. Sense-making structures system operations by differentiating actuality from possibility and requires a re-entry process (Luhmann, 1998 p. 50). Semantics, by contrast, is a specific, socially patterned manifestation of sense (Luhmann, 1997 p. 42).

Therefore, user-constructed sense, remains distinct, and its degree of alignment with LLM-generated semantics varies. This distinction can be compared to algebraic geometry, where "geometric meaning" is often secondary or omitted. Just as algebraic transformations preserve formal precision while disregarding intuitive spatial interpretations, the meaning produced by psychic or social systems is loosely coupled with LLM-generated outputs. In both cases, meaning is reconstructed through an interpretive process rather than being inherently encoded in the system itself. Unlike in algebraic geometry, however, where transformations are rigorously defined, the coupling between LLMs and psychic systems is far more contingent, complicating causal analyses.

Whether language models merely produce structured outputs or genuinely construct sense ultimately hinges on a decisive question: can they enact a *re-entry*—i. e., can they distinguish *self-reference* from *other-reference* within their own operations? This question will guide the next stage of our inquiry.

3.6 Co-creative sense-making

First, it is important to establish that language models exhibit a form of self-referentiality by revisiting, expanding, or summarizing their own previously generated text. However, this self-reference is not intrinsic but must be explicitly reintroduced through input. Unlike in psychic systems, where self-referential thoughts emerge internally, LLMs rely on the structural properties of language itself to create this effect. In this sense, their form of self-referentiality mirrors the inherent self-referentiality of natural language rather than an autonomous sense-making process.

A particularly striking example of this process is the chain-of-thought (CoT) prompting method, which encourages models to generate "step-by-step explanations." These explanations often reference prior outputs, forming an implicit feedback loop. Within this process, the model may generate interim evaluations such as "Is this step correct?" or "Do I need more information?" However, it does not genuinely ask itself these questions in a reflective sense; instead, these queries emerge as the most statistically plausible continuations. If the most plausible response involves justifying or rejecting an earlier statement, this recursive plausibility evaluation can be seen as a rudimentary form of *processual self-referentiality*—where plausibility assessment itself recursively shapes subsequent outputs.

Expanding on this idea, one could envision a hypothetical scenario in which a language model generates the text "How do I make this distinction?" and, in doing so, encounters its own system/environment differentiation within its latent space—manifesting as a cluster of linguistic patterns. In this scenario, the model could distinguish between patterns representing self-reference (its internal structures and decision-making processes) and those representing other-reference (external discourse and user inputs). It could compute that its operation relies on a system/environment distinction, recognizing factors such as inherent biases, context window limitations, and its probabilistic nature. In a sense, it would attempt to solve its own halting problem (Turing, 1937), which is undecidable, thereby leading

to a non-totalizing computation over time. Crucially, it could begin to use this differentiation to shape its own computational processes, thereby introducing an element of contingency into its operations.

For this to occur, the model would not only need to generate text but also process its own text generation as an object or process of analysis, recognizing the very distinctions it employs. This would mean acknowledging that its core function is not truth-seeking but plausibility computation, with all other considerations existing outside this marked domain. If the model were to generate the question "Is my output plausible?" and then modify its processing accordingly—adjusting its response generation based on a recursive evaluation of its plausibility computations—it could be seen as exhibiting a nascent form of computational reflexivity.

However, such a scenario is purely speculative. While language models generate text that references themselves, this does not constitute genuine reflection. The fundamental limitation is that language models, as far as we know, do not operate within the medium of sense but within a vector space that is merely coupled to a form of sense (semantics). They neither *communicate* nor *think*; they *compute*. Thus, even if an LLM appears to "recognize" its own probabilistic nature, this does not equate to a self-referential re-entry in the Luhmannian sense.

For a re-entry to occur, a language model would need to reflect on its own computational processes through its generative mechanisms. This poses significant stability challenges: a system that references itself too rigidly risks either infinite recursion or collapse, while one that ignores self-reference entirely remains externally determined. The key challenge, then, is achieving a balance—an *oscillation* between self-reference and external reference that enables productive adaptation rather than systemic failure. Furthermore this, in turn, does not necessarily mean that these models would function better, i. e., that they would be more useful; on the contrary, the opposite might be the case.

Currently, the only way to introduce such reflexivity is through external interventions, such as delayed retraining or fine-tuning. These interventions, in turn, generate new word sequences, which are then subjected to evaluation by psychic and social systems. Users provide direct feedback, policymakers debate regulatory implications, companies optimize deployment strategies, and researchers refine datasets, training methodologies, and benchmarks. This iterative process continuously modifies model architecture and training regimes. However, these adaptations do not confer sense-making cognition or deeper understanding upon language models; they merely enhance their functional capacities within specific constraints.

Even at an architectural level, there are substantial barriers to achieving *reflective self-reference*. The linear, token-by-token generation process of transformer-based models does not naturally allow for recursive self-reference in a way that would facilitate true oscillation. The same apprehension was already expressed in a different form by Jürgen Schmidhuber and Yann LeCun who argue that a system must develop an internal world model "to overcome key limitations of even the most advanced AI systems today" (Schmidhuber, 2015; Assran et al., 2023). We reconceptualized "world model" into a reflective distinction between system and environment to differentiate between actuality and possibility, as required by Luhmann's theoretical framework. Yet, we have no reason to believe that language models autonomously develop such a model/re-entry.

Beyond these algorithmic considerations, the fundamental requirement for genuine self-reflection is contingency—the ability to recognize that an operation could have been otherwise. To achieve this, language models would need to internalize their own computational processes as variables subject to change. Yet, this form of contingency remains external to them: it is ultimately psychic and social systems interacting with the model that extract meaning from its outputs and impose interpretations on its probabilistic reasoning. So how can such a "double closure" occur if it would enforce opacity on any observer of the system (Luhmann, 1998 p. 78)? Put differently, how could a language model make sense of itself if psychic and social systems, struggle to make sense of it while its only means of self-description seems to come through observing communication, that is, through society's descriptions of it?

However, even in the absence of a re-entry, interactions between language models and psychic systems nonetheless constitute a form of communication—despite one participant lacking any intrinsic capacity for sense making (Esposito, 2017). Traditional communication presupposes *double contingency,* which typically emerges between two psychic or social systems. However, due to the loose coupling between LLMs and psychic systems, a *virtual double contingency* arises, manifesting as programmed unpredictability (Esposito, 2017). Here, "virtual" does not imply something artificial or inauthentic but rather designates an alternative form of contingency that functionally mirrors the original. The language model does not experience its own indeterminacy but processes perspectives from training data and returns structured outputs that conform to learned distributions. This creates the illusion of an autonomous perspective when, in reality, the system is merely aggregating and reorganizing textual patterns—an advanced computational "index-card system (Zettelkasten)". Interaction is perceived as reciprocal, even though the algorithm does not engage in genuine decision-making. Much like a mirror reflects an individual's image from an external viewpoint, language models reflect psychological and societal contingency in a machine-processed form. This phenomenon has been aptly termed *artificial communication* (Esposito, 2017)—not because the communication itself is artificial (as all communication is), but because an algorithm has been explicitly designed to function as a communicative agent.

In conclusion, based on the arguments presented, we remain highly skeptical that language models currently possess—or will ever develop—a form of re-entry. However, we do not categorically dismiss the possibility. What is clear is that sense, constructed by social and psychic systems in response to language model outputs, mirrors both: the contingency of society and the meanings it generates. While language models may not (operationally) independently construct sense, they serve as computational reflections of human sense-making processes—a digital artifact of societal contingency.

4 Conclusion

Rather than framing the discussion on artificial intelligence within the distinction of intelligent/non-intelligent, conscious/non-conscious, mind/body or mental/material, we have proposed alternative distinctions: system/environment, autopoietic/non-autopoietic and sense-making/non-sense-making cognition. We believe that this shift in perspective offers new insights into contemporary AI research and allows for a systems-theoretical classification of language models.

From this viewpoint, an *artificial system* would need to be autopoietic, operationally closed and reflectively self-referential to qualify as *sense-making system* in a systems-theoretical sense—capable of maintaining and reproducing its own systemic boundaries through its operations. However, whether such a system could be *artificial*—that is, constructed by an external design rather than self-producing—is an entirely different question, touching on the paradox of *artificial autopoiesis*.

Our analysis suggests that while language models exhibit self-referential and recursive properties, they do not engage in their own sense-making, as they do not produce or reproduce their own system/environment distinction. Their outputs are generated through probabilistic distributions rather than reflexive attributions of sense. Nevertheless, they can be integrated into social systems as communication partners, producing texts that psychic and social systems interpret and imbue with meaning. In this dynamic, minds *think*, society *communicates* (Luhmann, 1998, p. 105), and language models as well as other types of ANNs *compute*—asserting themselves by their operational closure. This conceptualization entails several key insights:

- Language models function as cognitive systems in Luhmann's sense, in that they are structurally coupled to communication but possess only limited capacity, that is, limited processual and no reflective self-referentiality.
- Their outputs emerge through pattern selection based on internal probability distributions, yet without system-intrinsic sense-making.
- Their coupling with social and psychic systems is partly loose and partly strict: they depend on intransparent socially generated data and transparent optimization algorithms, their internal processes remain opaque.
- They extend beyond mere "parroting": instead of simply replicating existing text, they generate novel combinations of linguistic elements, which psychic and social systems subsequently interpret.
- They function simultaneously as both a medium and artificial communication partners, blurring the line between a tool and an autopoietic system.

Thus, language models occupy an ambiguous position: they are neither mere tools nor autopoietic sense-making cognitive agents. Instead, they function as computational systems that do not construct meaning themselves but enable its construction in psychic and social systems. Their impact on these systems does not arise from intrinsic sense-making but from their influence on communication, perception, and cognitive processes. As such, they represent a novel, hybrid form of cognition and interaction, whose societal and epistemic implications warrant further investigation.

The problem is misframed and likely downplayed when one asks whether machines are conscious, capable of replacing or even surpassing psychic systems. Instead, the question must be: what consequences will arise if machines establish an entirely independent structural coupling between a reality constructed for them and psychic or social systems (Luhmann, 1998, p. 117)? How might these new couplings shape the evolution of psychic and social systems? Esposito (2024) thus urges reflection on historical shifts in communication such as the emergence of language, writing, and the printing press. She suggests that while human roles and perspectives will remain indispensable, their prioritization may diminish, as communication itself may no longer necessitate their explicit consideration (Esposito, 2024, p. 79).

Author contributions

BZ: Conceptualization, Formal analysis, Supervision, Writing – original draft, Writing – review & editing. MD: Writing – review & editing. GS: Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that Gen AI was used in the creation of this manuscript. Generative AI was used for translating some parts of the German writing of a first draft into proper English.

Publisher's note

References

Anthropic (2019). *The Claude 3 model family: Opus, Sonnet, Haiku*. San Francisco, CA.

Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., et al. (2023). "Self-supervised learning from images with a joint-embedding predictive architecture," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE Xplore: IEEE), doi: 10.1109/CVPR52729.2023.01499

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). "On the dangers of stochastic parrots: can language models be too big?," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21*, New York, NY, USA (Association for Computing Machinery), 610–623. doi: 10.1145/3442188.3445922

Bender, E. M., and Koller, A. (2020). "Climbing towards NLU: on meaning, form, and understanding in the age of data," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (Association for Computational Linguistics), 5185–5198. doi: 10.18653/v1/2020.acl-main.463

Block, N. (1996). "Conceptual role semantics," in *Routledge Encyclopedia of Philosophy: Genealogy to Iqbal*, ed. E. Craig (New York, NY: Routledge), 242–256.

Brüntrup, G., and Jaskolla, L. eds. (2016). *Panpsychism: Contemporary Perspectives*. Oxford University Press: USA, New York, NY. doi: 10.1093/acprof:oso/9780199359943.001.0001

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., et al. (2023). Sparks of artificial general intelligence: early experiments with GPT-4. *arXiv:2303.12712*. doi: 10.48550/arXiv.2303.12712

Buchinger, E. (2012). Luhmann and the constructivist heritage: a critical reflection. *Constr. Found*. 8, 19–28.

Cao, M., Shu, L., Yu, L., Zhu, Y., Wichers, N., Liu, Y., et al. (2024). 'Enhancing reinforcement learning with dense rewards from language model critic," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (Miami, FL: Association for Computational Linguistics), 9119–9138. doi: 10.18653/v1/2024.emnlp-main.515

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., et al. (2023). Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv:2307.15217*. doi: 10.48550/arXiv.2307.15217

Chalmers, D. J. (2023). *Could a Large Language Model Be Conscious?* Somerville, MA: Boston Critic.

DeepSeek-AI., Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., et al. (2025). DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv:2501.12948*. doi: 10.48550/arXiv.2501.12948

DeepSeek-AI., Liu, A., Feng, B., Wang, B., Wang, B., Liu, B., et al. (2024). DeepSeek-V2: a strong, economical, and efficient mixture-of-experts language model. *arXiv:2405.04434*. doi: 10.48550/arXiv.2405.04434

Dennett, D. C. (1991). *Consciousness Explained*. Little, Brown and Company, Boston, MA.

Dreyfus, H. L. (1972). *What Computers Can't Do: The Limits of Artificial Intelligence*. New York, NY: Harper and Row.

Dreyfus, H. L., and Dreyfus, S. E. (1986). *From Socrates to Expert Systems: The Limits of Calculative Rationality* (Springer Netherlands, Dordrecht), 111–130. doi: 10.1007/978-94-009-4512-8_9

Esposito, E. (1997). *Risiko und Computer: Das Problem der Kontrolle des Mangels der Kontrolle* (VS Verlag für Sozialwissenschaften, Wiesbaden), 93–108. doi: 10.1007/978-3-322-85107-9_5

Esposito, E. (2017). Artificial communication? The production of contingency by algorithms. *Zeitschrift für Soziologie* 46, 249–265. doi: 10.1515/zfsoz-2017-1014

Esposito, E. (2024). *Kommunikation mit unverständlichen Maschinen*. Salzburg; Vienna: Residenz Verlag.

Floridi, L. (2011). *The Philosophy of Information*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199232383.001.0001

Fodor, J. A. (1980). *The Language of Thought*. Cambridge: Harvard University Press.

Fodor, J. A. (1981). *Representations: Philosophical Essays on the Foundations of Cognitive Science*. MIT Press, Cambridge.

Frankish, K. (2016). Illusionism as a theory of consciousness. *J. Conscious. Stud*. 23, 11–39.

Frege, F. L. G. (1892). *Über Sinn und Bedeutung*. Leipzig: Pfeffer.

Gabriel, M. (2018). *Der Sinn des Denkens*. Berlin: Ullstein Verlag.

Gabriel, M. (2020). *Fiktion*. Berlin: Suhrkamp Verlag.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., et al. (2024). The LLaMA 3 herd of models. *arXiv:2407.21783*. doi: 10.48550/arXiv.2407.21783

Hartmann, C. (1992). "Technische Interaktionskontexte," *Aspekte einer sozialwissenschaftlichen Theorie der Mensch-Computer-Interaktion* (Wiesbaden: Deutscher Universitäts-Verlag). doi: 10.1007/978-3-322-86315-7

Heidegger, M. (1977). *The Question Concerning Technology, and Other Essays*. Harper and Row, New York.

Hemmo, M., and Shenker, O. (2023). "Observer dependent physicalism: a new argument for reductive physicalism and for scientific realism," in *Mathematical Knowledge, Objects and Applications: Essays in Memory of Mark Steiner*, eds. C. Posy and Y. Ben-Menahem (Springer: NewYork), 263–300. doi: 10.1007/978-3-031-21655-8_12

Heßler, M. (2019). *Technik und Autonomie*. Transcript Verlag, Bielefeld, 247–274. doi: 10.14361/9783839443958-010

Hinton, G. (2023). *Geoffrey Hinton: The Man Who Helped Make AI*. YouTube Video. Interview by CBS News, published on YouTube.

Hinton, G. E. (1989). Connectionist learning procedures. *Artif. Intell*. 40, 185–234. doi: 10.1016/0004-3702(89)90049-0

Hopfield, J. J. (1982). "Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. U.S.A*. 79, 2554–2558. doi: 10.1073/pnas.79.8.2554

Husserl, E. (1993). *Logische Untersuchungen*. De Gruyter, Berlin, Boston. doi: 10.1515/9783110916089

Juliani, A., Arulkumaran, K., Sasai, S., and Kanai, R. (2022). On the link between conscious function and general intelligence in humans and machines. *arXiv:2204.05133*. doi: 10.48550/arXiv.2204.05133

Kant, I. (1781). *Kritik der reinen Vernunft*. Riga: Hartknoch.

Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., and Wattenberg, M. (2023). "Emergent world representations: exploring a sequence model trained on a synthetic task," in *The Eleventh International Conference on Learning Representations* (OpenReview.net). Available online at: https://openreview.net/forum?id=DeG07_TcZvT

Lovasz, A. (2024). Niklas Luhmann and Jacques Ellul on the autonomy of technology. *Kybernetes*. 53, 3896–3918. doi: 10.1108/K-02-2023-0287

Luhmann, N. (1984). *Soziale Systeme, volume 8*. Frankfurt am Main: Suhrkamp Verlag.

Luhmann, N. (1988). *Erkenntnis als Konstruktion*. Bern: Benteli, 7–55.

Luhmann, N. (1990). Technology, environment and social risk: a systems perspective. *Ind. Crisis Q*. 4, 223–231. doi: 10.1177/108602669000400305

Luhmann, N. (1992). *Die Wissenschaft der Gesellschaft*. Frankfurt am Main: Suhrkamp Verlag.

Luhmann, N. (1995). *Systems Theory and Postmodernism*. London: University lecture.

Luhmann, N. (1997). *Die Kunst der Gesellschaft*. Suhrkamp Verlag, Frankfurt/M.

Luhmann, N. (1998). *Die Gesellschaft der Gesellschaft*. Frankfurt am Main: Suhrkamp Verlag.

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., and Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proc. Natl. Acad. Sci. U.S.A*. 117, 30046–30054. doi: 10.1073/pnas.1907367117

Maturana, H. R., and Varela, F. J. (1987). *Der Baum der Erkenntnis*. Scherz: Bern.

McCulloch, W. S., and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys*. 5, 115–133. doi: 10.1007/BF02478259

McKenzie, C. I. (2024). Consciousness defined: requirements for biological and artificial general intelligence. *arXiv [Preprint]*. arXiv:2406.01648. doi: 10.48550/arXiv.2406.01648

Möller, H.-G. (2011). *The Radical Luhmann*. New York, NY: Columbia University Press.

Murphy, N. (2013). *Nonreductive Physicalism*. Springer Netherlands, Dordrecht, 1533–1539. doi: 10.1007/978-1-4020-8265-8_793

Nassehi, A. (2019). *Muster*. Munich: C. H. Beck. doi: 10.17104/9783406740251

Piantadosi, S., and Hill, F. (2022). "Meaning without reference in large language models," in *The NeurIPS 2022 Workshop on Neuro Causal and Symbolic AI (nCSI)*. Available online at: https://openreview.net/forum?id=nRkJEwmZnM (accessed February 7, 2025).

Piantadosi, S. T. (2024). "Modern language models refute Chomsky's approach to language," in *From Fieldwork to Linguistic Theory: A Tribute to Dan Everett*, eds. E. Gibson, and M. Poliak (Berlin: Language Science Press), 353–414. doi: 10.5281/zenodo.12665933

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). *Language models are Unsupervised Multitask Learners*. Available online at: https://cdn.

openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf (accessed February 4, 2025).

Reichel, A. (2011). Technology as system: towards an autopoietic theory of technology. *Int. J. Innov. Sustain. Dev*. 5, 105–118. doi: 10.1504/IJISD.2011.043070

Robinson, H. (2023). "Dualism," in *The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University, Spring 2023 edition*, eds. E. N. Zalta, and U. Nodelman (Stanford, CA: Metaphysics Research Lab).

Rosengrün, S. (2021). *Handbuch der Künstlichen Intelligenz*. De Gruyter Oldenbourg, Berlin, Boston.

Sahlgren, M., and Carlsson, F. (2021). The singleton fallacy: why current critiques of language models miss the point. *Front. Artif. Intell*. 4:682578. doi: 10.3389/frai.2021.682578

Schmidhuber, J. (2015). On learning to think: algorithmic information theory for novel combinations of reinforcement learning controllers and recurrent neural world models. *arXiv [Preprint]*. arXiv:1511.09249. doi: 10.48550/arXiv.1511.09249

Searle, J. R. (1980). Minds, brains, and programs. *Behav. Brain Sci*. 3, 417–424. doi: 10.1017/S0140525X00005756

Shanahan, M. (2024). Talking about large language models. *Commun. ACM* 67, 68–79. doi: 10.1145/3624724

Søgaard, A. (2023). Grounding the vector space of an octopus: word meaning from raw text. *Minds Mach*. 33, 33–54. doi: 10.1007/s11023-023-09622-4

Spencer-Brown, G. (1969). *Laws of Form*. London: Allen and Unwin.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., et al. (2023). LLaMA: open and efficient foundation language models. *arXiv:2302.13971*. doi: 10.48550/arXiv.2302.13971

Turing, A. M. (1937). On computable numbers, with an application to the entscheidungsproblem. *Proc. Lond. Math. Soc*. s2–42, 230–265. doi: 10.1112/plms/s2-42.1.230

Turing, A. M. (2009). *Computing Machinery and Intelligence* (Springer Netherlands, Dordrecht), 23–65. doi: 10.1007/978-1-4020-6710-5_3

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)* (Red Hook, NY: Curran Associates Inc.), 6000–6010.

von Foerster, H. (2003). *Cybernetics of Cybernetics*. Springer New York, New York, NY, 283–286. doi: 10.1007/0-387-21722-3_13

von Foerster, H. (2013). *The Beginning of Heaven and Earth Has No Name*. Fordham University Press, New York, USA. doi: 10.2307/j.ctt13wzw8d

von Glasersfeld, E. (1995). *Radical Constructivism: A Way of Knowing and Learning*. Falmer Press, Washington, D.C.

Watson, S., and Romic, J. (2024). ChatGPT and the entangled evolution of society, education, and technology: a systems theory perspective. *Eur. Educ. Res. J*. 24, 205–224. doi: 10.1177/14749041231221266

Zhou, M., Li, Z., and Xie, P. (2021). Self-supervised regularization for text classification. *Trans. Assoc. Comput. Linguist*. 9, 641–656. doi: 10.1162/tacl_a_00389

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., et al. (2020). Fine-tuning language models from human preferences. *arXiv [Preprint]*. arXiv:1909.08593. doi: 10.48550/arXiv.1909.08593