ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DEPARTMENT OF PHYSICS

SECOND CYCLE DEGREE

LARGE LANGUAGE MODELS AND KNOWLEDGE IN PIECES AS EPISTEMOLOGICAL LABORATORIES: INSIGHTS FROM PHYSICS EDUCATION RESEARCH

Supervisor

Prof. Olivia Levrini

Defended by

Pietro Girani

Graduation Session / 07 / 2026

Academic Year 2025/2026

Abstract

This thesis is situated within the field of Physics Education Research and investigates the relationship between Large Language Models (LLMs) and Knowledge in Pieces (KiP) around the problem of knowledge. Rather than reducing the comparison to assimilation or opposition, the work constructs a theoretical space in which artificial and natural knowledge can be examined through shared dimensions while preserving their specific differences. The first part reconstructs the historical and epistemological genealogy of Artificial Intelligence, from early times to LLMs. It then analyzes Transformer-based models with regard to their specific architectural features. In parallel, the thesis reconstructs KiP within the broader horizon of Knowledge Analysis.

The central contribution lies in the bidirectional comparison between LLMs and KiP. On the one hand, KiP provides a vocabulary for interpreting LLMs as granular, sub-symbolic and context-sensitive systems whose coherent performances emerge without explicit rule-based organization. On the other hand, LLMs make more visible some theoretical tensions internal to KiP, especially the distinction between local activation and stable understanding, the role of language, and grounded experience in human knowledge. The thesis argues that LLMs do not possess understanding in the human sense, since they lack embodied experience and direct access to the world; nevertheless, they may function as epistemological tools for deepening into the analysis of human knowledge. Finally, the thesis outlines future concrete research directions in which LLMs could support Knowledge Analysis, particularly with respect to prediction and bootstrapping, large-scale linguistic validation, learning timescales, runnability and boundary crossing beyond the domain of intuitive physics.

CONTENTS

Introduction

This thesis is situated within the research field of *Physics Education Research* and aims to investigate the relationship between different theoretical frameworks, on the one hand *Knowledge in Pieces* (KiP) and on the other *Large Language Models* (LLMs), around the problem of knowledge.

In recent decades, *Artificial Intelligence* (AI) and, in particular, *Machine Learning* (ML) models (e.g., LLM) have undergone a particularly rapid evolution in terms of efficiency and applicability, with a consequent systematic diffusion throughout the entire social fabric, in sectors such as education, finance, and healthcare (Kim et al., 2024). The *Big Data* era is widely acknowledged in the literature as a catalyst of this evolutionary phenomenon (Reinhardt & Buchholz, 2025), an evolution which nevertheless also brings with it problematic aspects, all the more delicate if one considers that, in certain domains – defined as "high-stakes" – artificial decision-making systems are already required and used (Baron, 2025). These systems are also commonly regarded as *opaque*, showing a lack of transparency with respect to the internal mechanisms involved in the mediation process between input and output. Such opacity would seem to derive from the convergence of data complexity, the high number of parameters and their nonlinear interactions, as well as from the dynamics of the learning process itself (Chandrasekaran & Jordan, 2013). This widespread diffusion has also renewed – perhaps with greater urgency than in the past – certain questions concerning what it means for a machine, for an artificial system, to "produce knowledge". The current debate seems to revolve around the claim to a definitive categorization of AI's actual intelligence, and around the possibility that it produces knowledge that may be regarded as substantial and not merely apparent. Typically, this debate tends quickly to evade deeper analysis, sliding toward strongly polarized positions. At opposite sides stand, on the one hand, those who praise AI's performances, and, on the other, those who regard artificial models and algorithms as little more than «stochastic parrots» (Bender et al., 2021), capable at most of repeating or imitating human capacities, displaying only appearances of intelligence.

On another side stands Knowledge in Pieces (KiP), an approach to the study of knowledge concerned with intuitive knowledge within the domain of scientific knowledge in physics. This approach is situated not only within the research field of PER, but also in dialogue with other models of knowledge proper to the *learning sciences*. KiP has its roots in the

constructivist and cognitivist tradition of studies on learning, and is traced back to the work of Andrea diSessa (1988, 1993), who proposed it as an alternative to the then dominant model of *conceptual change*. KiP is generally considered part of the broader horizon of *Knowledge Analysis* (KA). Indeed, although KiP historically emerged first, as a specific epistemological approach concerning intuitive knowledge in physics, KA was later made explicit as a more general methodological framework within which KiP itself and research positions aligned with it operate (diSessa et al., 2016). In any case, the aim of this broad line of research remains to understand how human knowledge is constituted, what elements it is composed of, according to what dynamics it evolves, and how it interacts with the world.

These two sides, apparently independent of one another, are in fact intertwined both with respect to their origin and with respect to the research questions that both frameworks are called to address, especially today, and in particular the question concerning the very nature of knowledge. It is precisely this commonality of perspective and focus that this thesis takes as its primary objective in comparing these models of knowledge: LLMs with respect to artificial knowledge and learning; KiP with respect to intuitive scientific knowledge and natural learning.

The aim of this thesis, in turn, itself represents an alternative mode of analysis with respect to the partial reductions previously recalled. There is no interest here in making the two frameworks under consideration converge as identical or equivalent, which would amount precisely to superimposing artificial knowledge and natural knowledge, or, conversely, to further demarcating their differences. A methodological choice of this kind would risk bringing the comparison back within the same logics of polarization already mentioned. The contribution of this thesis is instead to understand whether it is reasonable to compare these two models, with the general objective of bringing to light *generative research questions* capable of fostering further reflection. In other words, the intention is to construct a theoretical space within which it is possible, legitimate, and productive to compare different models of knowledge and learning.

The objectives thus outlined have contributed to identifying the *specific research questions* of this work.

The first research question (RQ1, epistemological-methodological) asks whether *it is possible to construct a synthetic framework in which Large Language Models and Knowledge in Pieces illuminate one another as models of complex and non-symbolic*

knowledge systems. This question is articulated into two symmetrical and complementary sub-questions. The first (Sub-RQ1a) asks *how the vocabulary of Knowledge Analysis may allow us to map and make intelligible the trajectories of generation and the sub-symbolic learning of Transformers*, dimensions that remain opaque to date. The second (Sub-RQ1b) asks, conversely, *to what extent LLMs – as an object – may serve as a new access route for analyzing and better understanding the theoretical framework of KiP itself*.

The second research question (RQ2, applicative and oriented toward further work) asks instead *to what extent LLMs may be concretely employed as a research tool for developing the KiP paradigm, overcoming its current limits*. Unlike RQ1, in RQ2 the relationship between the two frameworks is deliberately asymmetrical and instrumental; the question is whether and how LLMs may become a working tool for future research in Knowledge Analysis.

It is important to specify that this thesis does not have the possibility of concretely realizing the possible applicative developments that emerge in relation to RQ2. Nevertheless, this research question shows its value precisely insofar as it allows the future application horizons of KiP to be brought to light.

The path along which this thesis is structured has the precise aim of providing the content and material useful for answering the selected research questions. This path is articulated across four chapters.

The first chapter reconstructs a partial genealogy of AI, highlighting the historical, cultural, and epistemological premises that made possible, within a deeply interdisciplinary climate, the birth of this research field. The discussion traces cybernetics, information theory, cognitive theories, and Turing's epistemology up to the Dartmouth Conference, the site of the nominal birth of AI. Subsequently, an in-depth discussion is developed on the evolution of AI and its models, from symbolic AI to sub-symbolic AI, up to Deep Learning and Large Language Models.

Chapter 2 offers a close-up view of LLMs, particularly with regard to the computational architecture on which they are based. In this chapter, the constitutive elements, the relational structure among these elements, the learning process of the structure, the changes introduced by LLMs, and the limits generally attributed to them are discussed.

In Chapter 3, dedicated to KiP, the problem of natural knowledge and learning is addressed, with a specific focus on intuitive scientific knowledge, following a descriptive structure deliberately parallel to that adopted in the previous chapter.

The final chapter, which precedes the concluding discussions, represents the core contribution of this thesis. It is the place where the parallel structure constructed in Chapters 2 and 3 is put to use in order to concretely compare the two frameworks, thereby answering RQ1. From the comparison conducted in these terms, however, the first indications relevant to RQ2 also emerge: some of the generative questions brought to light in Chapter 4 already point to possible directions in which LLMs could, in future research, be employed as a tool for addressing the limits of KiP discussed in Chapter 3.

From a methodological standpoint, this thesis adopts a predominantly theoretical and comparative approach, grounded in an in-depth analysis of the literature concerning the two frameworks. This analysis is not limited to a separate descriptive reconstruction of the two models, but is oriented, from the outset, toward constructing the conditions of comparability of Chapter 4. To this end, as anticipated, Chapters 2 and 3 are structured according to a deliberately parallel exposition, articulated around common thematic blocks – the constitutive elements of knowledge, the relations among them, the mechanisms of learning, the theoretical revolutions, and the recognized limits of each framework – which make it possible to treat LLMs and KiP within shared dimensions, while safeguarding their respective specificities. In other words, this is a qualitative-conceptual methodology, which does not involve the collection of original empirical data, but is based on the critical reworking and relational analysis of distinct theoretical corpora.

The contribution of this thesis is situated within a broader context in which Physics Education Research is increasingly called to engage with the growing role of Artificial Intelligence in educational settings. The expected impact of this work can be considered on two levels, corresponding to the two research questions identified. At the theoretical and epistemological level, the comparison between LLMs and KiP aims to contribute to a better understanding of both frameworks, showing how each may help illuminate aspects of the other. At the more applicative level, although this thesis does not seek to develop concrete tools, the reflections emerging from RQ2 are intended to provide a starting point for future research exploring how LLMs might support work in Knowledge Analysis.

Chapter 1

1.1. Historical context: Cultural and Epistemological premises

In order to enter into the discussion and reflection around the theme of *Artificial Intelligence* (AI), it is appropriate to disclose the cultural and epistemological premises from which computer science was born and developed.

The term Artificial Intelligence is associated with the evolution and technological progress typical of the contemporary age, but in reality, it emerged almost seventy years ago. The term was coined by John McCarthy, who, together with other researchers (Marvin Minsky, Nathan Rochester and Claude Shannon), organized a summer conference on the subject at Dartmouth College in Hanover in 1956. Many were those who participated in the seminar, many of whom made a decisive contribution to the development of the disciplinary fields of computer science, psychology, economics, linguistics: Ray Solomonoff, Arthur Samuel, Oliver Selfridge, Trenchard More, Allen Newell, Herbert Simon, Clifford Shaw, George Miller, John Nash, John Holland, Julian Bigelow, Ross Ashby.

The profoundly interdisciplinary character of this working group constitutes the solid background of AI at its beginning. The aim of the summer conference at Dartmouth is thus declared:

> *«The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to know how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer»* (McCarthy et al., 2006).

The methodological direction of the conference appears evidently compatible with the cognitive sciences not only because of the themes stated (learning, language, concepts), but also because these entities are treated as describable and replicable processes, that is, as legitimate objects of modelling. Precisely here it becomes decisive to make explicit the cultural background of the period, because the availability (or otherwise) of using a mentalistic lexicon for machines depends on the passage, still underway in the 1940s and 1950s, from behaviorism to cognitivism.

Behaviorism is an approach developed by John Watson at the beginning of the twentieth century that tends to elude a priori the theoretical construct of "mind", categorized as a black box, while what is identified as *observable* – and therefore to be studied scientifically – is *manifest behavior*. The latter is defined as a process of mechanical causation between a *stimulus* (input) and *response* (output). If the mind is opaque, that is, inaccessible, learning is possible through prediction and control of behavior by manipulating stimuli and reinforcements (Pavlov's *classical conditioning* or Skinner's *operant conditioning*). The cognitivism of Jean Piaget et al. (Bruner, Vygotsky, Ausubel, Novak and Gowin) strongly opposes this paradigm and overturns its methodological choices. The mind is an active protagonist of learning processes; it elaborates incoming information by constructing representations and mental schemas (formalizable as internal states) and then returns an output. By rejecting the direct stimulus-response association, the "cognitive revolution" explicitly reintroduces processes and representations as mediating protagonists of the knowing process.

The passage from behaviorism to cognitivism does not occur by chance, but because the available conceptual infrastructure changes. In those years, in fact, two theories underwent major developments: *cybernetics* and *information theory*.

1.1.1. Cybernetics

Up to this point in history, science had seen its point of attention in the study of systems that are intrinsically simple or reducible to simple components (reductionism). In the passage from the nineteenth to the twentieth century there is a shift of attention toward all those types of systems (e.g. neuronal systems or economic systems) that have a large number of interconnected variables, such that the alteration of one factor produces immediate changes in many others. This shift does not yet coincide with the full constitution of the *science of complex systems*, but it prepares its epistemological ground. This field can be traced back to general systems theory, which seeks organizational principles common to biological, social and artificial systems (von Bertalanffy, 1968), to the distinction between disorganized complexity and organized complexity proposed by Weaver (1948), and to the later developments of non-equilibrium thermodynamics, where order and emergence are thought starting from open systems far from equilibrium (Nicolis & Prigogine, 1977; Prigogine & Stengers, 1984). On a more epistemological plane, Morin and Ceruti have shown that complexity does not indicate only a class of phenomena, but a different way of thinking organization, interdependence, constraints and emergence, beyond the limits of linear

reductionism (Bocchi & Ceruti, 1985; Ceruti, 1986; Morin, 1990). In this framework, cybernetics represents a decisive threshold, because it introduces categories such as system, feedback, control, communication and self-regulation. Thus, the beginning of the science of complex systems formalized at the epistemological level is marked.

Precisely the need to address and control complexity favors the birth of cybernetics, that is, the science of control and communication, in the animal and in the machine (Ashby, 1956), according to the definition of the period.

Whereas classical science directs its inquiry toward cause-effect links and physical magnitudes, cybernetics carries out a shift of investigative focus. As Bianchini argues, cybernetics is the science of the mechanisms of control and transmission of information in natural or artificial systems (Bianchini, 2007). This change is what Norbert Wiener (1948) and Gregory Bateson (1977) described as the passage from the "science of substances" to the "science of forms". In other words, cybernetics abstracts from the specific nature of the variables at stake in order to question the mechanisms of regulation that constrain the process toward a determined outcome (specification).

This reconfiguration of the fundamental questions is not a simple change of method but underlies and implements a new ontological structure.

Dynamic system. «Cybernetics […] is a "theory of machines", but it treats, not things but ways of behaving» (Ashby, 1956). The object of study are, therefore, systems that are not described by a point-like state but by a «trajectory or line of behaviour» (Ashby, 1956). In this sense they can be called dynamic systems, insofar as the temporal dimension is indispensable.

Organization of behavior. The behavior of a system is oriented ("organized") toward a purpose, toward a goal-state. How this orientation is concretely realized will be clarified further on, together with the concept of regulatory circuits and *feedback*.

Organization of behavior over time. If a cybernetic system is not defined by its point-like state in the instant but by a trajectory in time, what are sought are regularities (i.e. *pattern*, we would say today), that is, the recurring and recognizable (stable) configurations that the system produces and reproduces. In common terminology, however, the concept of stability is ambiguous. In this context, Ashby specifies that «through all the meanings runs the basic idea of an "invariant": that although the system is passing through a series of changes, there

is some aspect that is unchanging; so some statement can be made that, in spite of the incessant changing, is true unchangingly» (Ashby, 1956). Therefore, the sought stability of the system does not mean absence of variation, but the capacity to maintain a configuration within a set of configurations compatible with its goal-state: «a system can be said to be in stable equilibrium only if some sufficiently definite set of displacements D is specified» (Ashby, 1956). It is understood that given «a state of equilibrium A, D is usually defined to be a displacement, from A, to one of the states "near" A» (Ashby, 1956).

The system governs its own organization through *communication* and *regulation*.

Communication circuits. They are channels for the transmission of information. Following Wiener, information is a measure of order (*negentropy*, cf. Information Theory) and of the regularity of a pattern. A random sequence does not transmit information, while there is information when a message introduces a recognizable difference with respect to chance, that is, it reduces uncertainty. What circulates in a cybernetic system are signals that reduce uncertainty about the state of the system or of the environment, making its regulation possible. The circulation of this information is the structural condition without which no control is possible. Wiener goes so far as to maintain that information is information, neither matter nor energy (Wiener, 1948). Information is elevated to a fundamental ontological category, alongside matter and energy, configuring a new conceptual world within which to describe phenomena.

Regulatory circuits. The control of the evolution of the cybernetic system occurs substantially on two levels: a higher control level (or regulator) and a lower operational level (the regulated system). The lower level manages local perturbations through continuous feedback cycles. The feedback – or retroaction – is the heart of cybernetics: it designates the process by which the output of a system is reintroduced as input, creating a closed circuit or *loop*. The system acts on the environment or on itself, producing a measurable effect. This effect is configured as the current state. The current state is compared with a reference value (technically called *setpoint* or prescribed value). The difference between the two states is calculated, producing an error as output, namely the gap between what it is and what it should be. This error activates a correction, that is, the behavior of the system is modified in the direction that reduces the gap. For example, in a thermostat, the reference value is the set temperature, let us say 20°C. A sensor continuously detects the room temperature. If the actual temperature is 17°C, the error is -3°C and the system activates the heating; when the

difference becomes zero, the heating turns off. The system does not "know" how cold it is in an absolute sense, it only knows how far it deviates from the reference. The higher level, instead, monitors the global state, and when local feedback is not sufficient to keep the system within acceptable limits, it intervenes by modifying the parameters - that is, it issues commands (hierarchical control).

Purpose and self-regulation. The setpoint is what, in the mechanism of regulation, implements the purpose. It is not an intentional attribute, but a functional parameter internal to the system that defines the final state to be reached. It is therefore incorporated into the architecture of the system itself. The architecture of loop feedback produces a second decisive consequence: self-regulation. The system is capable of converging toward its own goal-state and of maintaining it without an external agent having to intervene at every step. Control is distributed within the cycle itself, thanks to the transmission of information. This makes reasonable Ashby's definition according to which *cybernetics could be defined as the study of systems that are open to energy but closed to information and control* (Ashby, 1956). Energetic openness guarantees that the system can act on the world; closure to information – or *information-tight* (Ashby, 1956) – guarantees that every effect of that action returns as input, maintaining the loop.

The framework thus outlined allows a non-ambiguous comparison with behaviorism. On the one hand, cybernetics is effectively behavioral in the sense indicated by Ashby. It privileges functional descriptions and observable regularities (behaviors), and can study a system even by treating it as a black box (manipulating input and recording output in order to infer constraints and transformations). On the other hand, it departs from the behaviorist stimulus response schema precisely because behavior in retroaction is not a linear chain: it requires internal states and communication circuits that mediate action, legitimizing a "mentalistic" vocabulary (purposes, control, decision) in an operational sense.

1.1.2. Information Theory

If cybernetics configured itself as the science of the mechanisms of control and transmission of information in natural or artificial systems, the birth of information theory appears closely connected with cybernetics, but at the same time begins to lay the foundations for a clearer passage from behaviorism to cognitivism that in those same years was one with the emergence of AI (Bianchini, 2007).

While redefining the way in which it is possible to describe the behavior of systems, cybernetics left open the question: what actually circulates in communication circuits? Is it possible to quantify information? It is precisely into this fissure that Claude Elwood Shannon's *information theory* enters. His work is not to be understood as a philosophical-psychological continuation of cybernetics, but as its mathematical-engineering formalization, starting from the identification of mathematical rules that describe this phenomenon. The fundamental problem of communication is that of reproducing at one point the same – or approximately the same – message selected at another point (Shannon, 1948). On the other hand, just as for cybernetics it is irrelevant to attend to the energetic or material parameters of the system – that is, the substance – for Shannon's engineering approach to information theory the meaning of the communicated message is not important. What is significant is that the actual message is the one selected from a *set* of possible messages (Shannon, 1948). This resistance to semantics is not a theoretical renunciation, but an ontological gain. Information, in fact, is traced back to the structure of possible alternatives and to their probabilistic distribution. Shannon considers first of all the source of information and asks how it can be described mathematically (Shannon, 1948). In telegraphy, for example, the messages to be transmitted consist of a series of letters that are not arranged in random order. The way in which a sentence is constructed necessarily follows a statistical structure proper to the language to which reference is made. Therefore, the succession of each symbol (letter) will have an occurrence (or probability) different from the others. Shannon argued that a physical system or a mathematical model of a system that produces a sequence of symbols governed by a set of probabilities is called a stochastic process (Shannon, 1948).

Recalling what was said previously in Wienerian terms, there is information insofar as a message introduces a recognizable difference with respect to chance, that is, it reduces uncertainty. For Shannon, instead, there is information when the improbability of the observed pattern increases. This difference will become evident below.

Shannon grafts his formalization by defining information as a choice among possible messages, measurable through *Shannon entropy*:

$$H = -\sum_{i} p_i \log_2 p_i \qquad (1)$$

where p_i is the probability of the symbol (or message) i. The base 2 of the logarithm indicates that the unit is the *bit* (binary digits). The form of Shannon entropy is the only possible one for satisfying continuity, monotony and decomposition (additivity). The choice of the logarithm is dictated by practical reasons.

To understand the impact of this choice, it must be considered that:

- $H = 0$ if and only if we are certain of the outcome (one probability is 1 and the others are 0). In this case there is no uncertainty.
- H is maximum when all events are equally probable; this is the situation of maximum uncertainty.

The first case is associated with the concept of null information because if a system is completely predictable, each of its outputs transmits "nothing new". The message does not change the state of knowledge one has about the system. The second case, instead, is associated with the concept of maximum information, or rather maximum capacity of information. If uncertainty is maximum, each output represents new information about the system.

Comparing Wiener with Shannon, the former looks at the stability of systems in action. In this frame, information is what *reduces disorder*, what allows the system to correct deviations from its goal-state. Information is therefore associated with order, regularity, negentropy. Shannon looks instead at *transmission* and asks how much a receiver can learn from a message. The answer to this question depends on how much uncertainty there was before receiving it. Here information is what *produces novelty*, and it is measured with entropy.

The new notion of entropy introduced by Shannon (with respect to Boltzmann's thermodynamic entropy) makes it possible to pose quantitatively some decisive questions: what rate of transmission is possible without noise (channel capacity), what rate remains possible with noise while keeping the error arbitrarily small (coding theorems), how to compress a message (efficiency) and, for continuous signals, what sampling frequency makes the signal re-constructible. Here information theory introduces limits "of principle" (bandwidth, noise, capacity) that transform communication into an engineering problem, as it was in Shannon's idea.

It is at this level that welding with cybernetics becomes urgent. Cybernetics had placed the closed circuit of regulation at the center, but precisely that circuit requires efficient communication channels. From here one also understands why Wiener insists on information as an entity in itself and as a new brick of the description of reality ("the new world is made of matter, energy and information"). Shannon makes this thesis operational, and this operationality is what allows the notion of information to be transported, with credibility, from technical-engineering systems to models of the mind (models of the cognitive sciences). This is because information theory legitimizes a level of internal informational states (overcoming behaviorism) and of the transformations that govern them as a fully formalizable object, that is, describable as encoding, memory, transmission, reduction of uncertainty, without needing to resolve semantics.

Shannon's founding gesture – treating language as a statistical source of symbols, regardless of their meaning – will reveal itself to be an epistemological choice of far greater scope than his engineering intentions. What was born as a functional mechanism for the theory of information transmission contains the principle on which modern linguistic models of AI rest.

1.1.3. The Macy Conferences: the Interdisciplinarity that precedes Dartmouth

The Macy Conferences were a series of ten interdisciplinary meetings concerning the themes just proposed, held in New York under the direction of Frank Fremont-Smith at the Josiah Macy Jr. Foundation in the period 1946 –1953, with an organizational antecedent in 1942 that contributed to forming the main nucleus of the group. In 1943 the publication of the Rosenblueth-Wiener-Bigelow paper on *Behavior, Purpose and Teleology* further focused attention on the problems that would lead to the first Macy conference on cybernetics in March 1946, entitled *Feedback Mechanisms and Circular Causal Systems in Biological and Social Systems* (American Society for Cybernetics, n.d.).

The historical context of reference is that of the first half of the twentieth century, deeply marked by the two World Wars, only a few years apart. The twenty years that preceded the birth of AI are significant precisely because of the evolution that cybernetics underwent, because of the effective practical results it gave in the military field and that brought to it an ever-increasing number of financial resources (Bianchini, 2007). During the Second World War, scientists from different fields had already collaborated to solve complex strategic problems. For example, Alan Turing offered a decisive contribution to the decryption of Nazi

codes (De Caro & Giovanola, 2025). Among the participants of the first Macy Conferences there appear figures connected to military or technical-military institutions, for example John Stroud, indicated as belonging to the U.S. Naval Electronic Laboratory (Pias, 2003), a further sign that part of the competences and professional networks were consolidated within institutions that, during and after the war, had reasons to invest in communication, control and automation.

This period is significant not only for these developments and funding but also for the interconnections between cybernetics and the disciplines interested in the rational action of human beings, such as psychology, economics and some branches of philosophy, interconnections that seemed necessarily destined to flow into at least a partial fusion of methods, objectives and techniques by all those who dealt with such themes (Bianchini, 2007). In this perspective, cybernetics and information theory had produced a vocabulary of new concepts – control, feedback, information, entropy – that seemed applicable both to biological systems and to artificial ones, but that no existing discipline was equipped to treat alone. A meeting place was therefore needed where the major experts of different disciplines could be brought together, making possible the analysis of problems at the boundary.

Although the historical context suggests that the war acted as a catalyst in the development of cybernetics, it is necessary to correct the possible reading that reduces the Macy Conferences to a project primarily oriented to war aims. In the transcripts of these conferences, it is interesting to focus attention on what is the introduction to the Macy Conferences by the director Freemont-Smith. He felt a profound ethical urgency in the post-war period, convinced that the development of the natural sciences had created unprecedented risks for humanity. The construction of nuclear weapons is the most immediate reference in this regard. To counter this threat, he considered an immediate interdisciplinary dialogue indispensable. He was convinced that this unification and communication (meeting place) between physical and psychological disciplines was not only an academic goal, but a necessary condition for the achievement of peace in the world (Pias, 2003).

Therefore, the aim of the Macy Conferences is twofold: on the one hand, to create a meeting place for different disciplines so that the confrontation on the ethical questions of science could be effective; on the other hand, a common language was necessary to lay the foundations for a general science of the functioning of the human mind, that is, to examine

to what extent the analogies between organisms and machines could be legitimately pushed (Bianchini, 2007), a point that propelled the diffusion of what later became known as cognitive science.

Through which questions is the need to model the functioning of the human mind addressed? In summary, if cybernetics had already shifted the focus from constituents to behavior, that is, from what systems are internally made of to how they organize themselves, the question becomes more radical: if organization is sufficient, then are mind and machine of the same nature? And if so, at what level of description? The Macy Conferences do not solve this question, but they make it scientifically legitimate, within a specific theoretical framework.

The final summary of the Macy Conferences (cf. Pias, 2003), written by Warren McCulloch, outlines the epistemological path by which cybernetics defines itself through the overcoming of the purely physical datum in order to arrive at a logic of form and information. Following McCulloch's reflections and the minutes of the sessions, there are three principal points of agreement reached, or questions left open, at the end of the Macy Conferences: the dual nature of signal, the modelling of the mind and of nervous networks, the object-observer interaction.

Before presenting the cybernetic redefinition undergone by the concept of signal, McCulloch is keen to specify that, generally, in classical science the signal is understood as a measurable physical event that occurs at a determined point in space and time (following Einstein's treatment), therefore considered exclusively in its energetic and material dimension: for example, an electrical impulse, a sound wave or a pressure variation, etc., described by discrete or continuous variables. Following the works of Shannon and Wiener, it is established that every signal has two aspects: one physical, the other mental, formal or logical (Pias, 2003). The physical aspect concerns the material support (voltage, pressure, impulses), while the formal aspect concerns the information that the signal conveys, defined as entropy (or negentropy, considered two sides of the same coin). This formal nature is what allows the functioning of the system to be abstracted from its matter, to lay the foundations for treating the human mind as a processor of messages (or symbols), that is, to divert the gaze from biological matter in order to concentrate on "calculating machines". In this, the distinction between analog and digital computers (which follows the distinction between analog and digital signals) represented one of the most significant epistemological turning

points. According to McCulloch's synthesis, the difference lies in the way numbers are represented:

- Analog Computers: In these devices, the magnitude of a continuous variable is made directly proportional to a number that enters into the calculation.
- Digital Computers: These systems are based on a set of stable values (at least two, such as the "on" or "off" state) separated by regions of instability. The number is represented by the configuration of these stable states.

The group of cyberneticists reached the conclusion that the discrete action of nervous components was the only way in which the nervous system could manage the immense quantity of information that crosses it (see below). For this reason, the *Universal Turing Machine* was adopted as a "model" for brains.

In doing so, the Macy Conferences model human behavior and artificial behavior as comparable, since they can be described with the same vocabulary.

Finally, if a cybernetic system is defined by its control and information circuits, and if the one who studies such a system is in turn a cybernetic system, then the subject/observer-object separation that founds classical science becomes problematic. This question, already latent in the conferences, would be formalized by von Foerster in second-order cybernetics only in the 1970s.

From this it follows that the Macy Conferences constitute a direct antecedent of Dartmouth because they make legitimate and practicable what Dartmouth would presuppose, namely a common lexicon for speaking of mind and machine and the idea that analogies between organisms and machines can be legitimately practiced within a framework of control, communication and information.

1.1.4. Turing and the Epistemology of AI

Alan Turing can be considered the founding father of the epistemology of AI, even before AI proper, which would be born nominally only with the Dartmouth conference of 1956. In 1950, Turing published the article *Computing Machinery and Intelligence*. If cybernetics with the Macy Conferences had made scientifically practicable the comparison between organisms and machines, with the aim of developing a general modelling of the human mind, Turing follows an apparently complementary but intersectional approach, namely he asks whether it is possible for machines to think. Precisely with these words, «I propose to

consider the question, “Can machines think?”» (Turing, 1950), he opens his article and immediately argues about replacing this question with another, epistemologically different one. Following his reasoning, answering the question “can machines think?” requires the further definition of what the concepts of “machine” and “thinking” are. The path to take in order to carry out this task proves dangerous and inconclusive, since the answer would have to be sought through a statistical survey, but this is absurd for Turing. He continues, «the new form of the problem can be described in terms of a game which we call the “imitation game”» (Turing, 1950), later known as the *Turing test*.

Let us see in detail the original and integral formalization of the imitation game:

> *«It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart front the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either “X is A and Y is B” or “X is B and Y is A.” The interrogator is allowed to put questions to A and B thus: C: Will X please tell me the length of his or her hair? Now suppose X is actually A, then A must answer. It is A’s object in the game to try and cause C to make the wrong identification. His answer might therefore be: “My hair is shingled, and the longest strands are about nine inches long.” In order that tones of voice may not help the interrogator the answers should be written, or better still, typewritten. The ideal arrangement is to have a teleprinter communicating between the two rooms. Alternatively, the question and answers can be repeated by an intermediary. The object of the game for the third player (B) is to help the interrogator. The best strategy for her is probably to give truthful answers. She can add such things as “I am the woman, don’t listen to him!” to her answers, but it will avail nothing as the man can make similar remarks. We now ask the question, “What will happen when a machine takes the part of A in this game?” Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman? These questions replace our original, “Can machines think?”»* (Turing, 1950).

The conceptual structure of the test highlights peculiarities and necessary constraints for the correct performance of the game. First of all, it is required that the interrogator be completely isolated from the other two interlocutors, so that tracing the identity of the interlocutors (whether they are machines or human persons) is possible only through the analysis of the outputs that the interrogator receives in response to his questions. This detail contributes to outlining a theoretical framework in which the approach is totally functional to the processes, making irrelevant the nature of the support of the interlocutor (whether it is biological support or technological support) with respect to logical-linguistic performances. It is a sort of *de-corporealization* of the linguistic process from the nature of support. Secondly, Turing

himself in his article maintains that the method employed in the game – questions and answers only – is suitable for covering almost every field of human activity, «from mathematics to chess» (Turing, 1950). Reflecting information theory, the Turingian framework underlines the centrality of language as an instrument of cultural and cognitive synthesis. In other words, Turing elects natural language as the paradigm for testing whether a computer is or is not capable of conversing in a way indistinguishable from a human being. This would provide definitive operational proof of being able to construct a thinking machine. It can be said that Turing is also one of the first to address with a formal and non-semantic approach what would later be called *Natural Language Processing* (NLP).

Turing continues the article by specifying which types of computers are allowed access to the imitation game: digital computers. Questioning the reasonableness of this hypothesis requires taking a step back.

In 1937 Turing published his first scientific article, entitled *On Computable Numbers, with an Application to the Entscheidungsproblem*. In this work, Turing addresses a problem proposed by the mathematician David Hilbert in 1928, the *Entscheidungsproblem*, the "decision problem": determining whether there exists a mechanical method capable of establishing, for every possible mathematical statement, whether it is true or not. Already Kurt Gödel in 1931 with his *incompleteness theorems* had suggested the possibility of a negative answer. What remained open, however, was the possibility of mathematically formalizing the concept of "mechanical method", that is, of method of calculation (we are within computability theory). For example, multiplying two numbers in columns or calculating the derivative of a polynomial are considered mechanical tasks. In the 1937 article Turing succeeded in providing this definition, introducing what have since been known as *Turing Machines* (hence TM).

If we think of a human person intent on performing mechanical calculations such as those cited in the example above, we can picture an individual who pedantically executes pre-established rules, having available only a sheet of paper and a pen. In addition, if the required task is, for example, the solution of an algebraic expression, the subject will engage gaze and memory in a circumscribed way, carrying out operations focused on a limited number of symbols at a time. From the identification of parentheses to the hierarchical precedence of operations, he proceeds with the processing of the internal operations and moves, in

successive phases, to the analysis of exponents and of the other operators. Human memory is, therefore, limited. From this, Turing constructs the following analogy:

> «we may compare a man in the process of computing a real number to a machine which is only capable of a finite number of conditions $q_1,\ q_2, \dots, q_R$ which will be called "[states]"[1]. The machine is supplied with a "tape", (the analogue of paper) running through it and divided into sections (called "squares") each capable of bearing a "symbol". At any moment there is just one square […] which is "in the machine". We may call this the "scanned square". The symbol on the scanned square is called the "scanned symbol". The scanned symbol is the only one of which the machine is, so to speak, "directly aware"» (Turing, 1937).

In detail, a TM is composed of four fundamental elements:

1. Tape: a potentially infinite memory support, divided into single squares. Each square can contain a symbol belonging to a finite alphabet, in our case reduced to 0/1/,/*blank square.*
2. Read/Write Head (RWH): a device that interacts with the tape and that can read the symbol present in a square, write a new one or erase the existing one. It moves on the tape one square at a time, to the right or to the left.
3. (Internal) States: a finite set of conditions (indicated as $q_1,\ q_2, \dots, q_R$) that represent the "memory" of the machine and allow it to know how to interpret the symbol it is reading.
4. Instruction Table: establishes what the machine must do for every possible configuration (the union of the current state and of the read symbol).

The instruction table is therefore the quintuple of elements:

s	Current state	State
i	Symbol read currently	Reading
$I(s;i)$	Overwritten symbol, function of the first two parameters	Writing
$V(s;i)$	Direction of movement of the RWH, function of the first two parameters	Movement
$S(s;i)$	Next state, function of the first two parameters	New State

Table 1: quintuple of elements of a TM.

[1] In the original text "m-configurations", here rendered as "states" to avoid ambiguity with the tape configurations cited later.

The TM thus conceived is distinguished from finite automata, while itself belonging to the broader family of discrete-state machines. Its discreteness depends on the fact that it operates through a finite set of internal states, a finite alphabet of symbols and an instruction table that establishes, for every possible configuration, which symbol to write, how to move the head and which new state to assume. However, the TM does not coincide with a simple finite automaton, because it does not exhaust its memory in the sole set of internal states. It in fact has a tape divided into squares, readable and rewritable, which functions as a potentially unlimited external memory. It is this element that makes the TM a digital computer in the Turingian sense, not the fact that it already operates according to the binarism of modern computers, but the fact that it manipulates discrete symbols according to defined rules, preserving and modifying information on a memory support not reducible to the sole internal state of the machine.

The set of problems solvable with the Turing machine coincides, according to a conjecture of Alonzo Church, with the set of computable functions. This conjecture then leads to the Church-Turing thesis (follows).

Precisely because the tape can contain not only data, but also the formal description of another machine, it becomes possible to think a universal machine capable of simulating any other TM. In fact, the decisive conceptual step occurs in chap. 5 of the 1937 article. Here the possibility of the existence of a universal machine is hypothesized, later called the *Universal Turing Machine* (UTM), by analyzing what is the numerical encoding of the machine itself, which transforms the logical structure of a device into a datum readable by another device. First of all, Turing translates every row of the instruction table (the quintuples) into a Standard Description (S.D.). That is, he replaces the names of the states and of the symbols (in reading or writing) with a sequence of conventional letters (A, C, D, L, R, N). The table of a machine, having become a continuous string of symbols, is then replaced with a series of digits, one specific for each letter (for example A=1, C=2, D=3, …). The description of the machine is compacted into a single number, called Description Number (D.N.).

At this point, here is the decisive intuition: if a machine can be described by a number, that number can be written on the tape of another machine exactly as if it were any datum. One arrives at the definition of the UTM:

> *«It is possible to invent a single machine which can be used to compute any computable sequence. If this machine U is supplied with a tape on the beginning of which is written the S.D. of some computing machine M, then it will compute the same sequence as that machine»* (Turing, 1937).

The UTM is the abstract construction of a machine capable of computing any computable sequence, that is, a machine that, given the S.D. of another, simulates it. This is the reason why Turing, in the 1950 article, restricts access to the *imitation game* only to digital computers.

It is appropriate to remember that the assumption that we have carried with us up to this point is that thought is a formalizable process. If this were so, a universal computer (that is, the UTM) is theoretically able to simulate it. It is the Church-Turing thesis, which states: «if a problem is humanly calculable, then there will exist a Turing machine capable of solving it (that is, of calculating it)». More formally we can say that the class of calculable functions coincides with that of the functions calculable by a Turing machine. In other words, the intuitive concept of *mechanical method* (the one from which Turing had started with regard to the decision problem) coincides exactly with what a TM can do.

Repositioning ourselves in the historical context described previously, if the Macy Conferences provided the interdisciplinary humus and the vocabulary that normalized terms such as "feedback" and "control", Turing provided in advance the mathematical anchor that made those concepts operational and calculable (Bianchini, 2007; Pias, 2003).

We have already said that the TM (or the UTM) is an abstract formalization. Its realization occurs thanks to John Von Neumann. He implements the structure that, still in Turing's idea, a digital computer must have:

1. Store: corresponds to the blank sheets of the human person, whether these are the sheets on which he performs his calculations or the one on which the instruction table is printed. Just as the human person carries out calculations in his mind, the store corresponds to the computer's memory.
2. Executive unit: it is the part that carries out the various individual operations involved in a calculation.
3. Control: its task is to verify whether the executive unit performs calculations following the instructions contained in memory.

Although Turing's abstract conceptualization makes it possible to imagine machines that have at their disposal an "infinite tape", in practice this is not possible. However, Turing

himself states that «digital computers fall within the class of discrete-state machines […] but the number of states of which such a machine is capable is usually enormously large» (Turing, 1950) precisely thanks to the store.

Von Neumann translates these abstract concepts concretely, devising the basic architecture that constitutes all modern computers.

In conclusion, Turing's epistemological revolution is threefold in nature: methodological (the test), ontological (functionalism as the overcoming of behaviorism, universality and language) and technical (the architecture of computers).

The overcoming of behaviorism carried out by Turing does not consist in the rejection of manifest observables as criteria for evaluating the learning process, but in its radical reinterpretation. Classical behaviorism traces every response back to a mechanical stimulus response chain, without admitting the scientific legitimacy of internal states: what counts is only the observable output, independently of any intermediate processing. Turing, on the contrary, constructs a device whose internal structure is perfectly known, and the test serves to verify whether that internal structure, by virtue of its functional organization, is capable of producing behavior indistinguishable from intelligent behavior. Ontologically, intelligence is not a property of the material substrate (biological flesh or silicon), nor a simple reflection of physical structure, but a property of the organization of internal states. The imitation test becomes the method for ascertaining that a certain internal functional organization has effectively replicated the function of thinking - a sufficient condition, for Turing, to attribute intelligence to it. From the technical point of view, Turing's contribution lies not only in the formalization of the universal machine (UTM), but in the fact that the program of the machine is no longer incorporated into the physical structure of the machine, but separated from it and treated as manipulable data. Here the distinction between hardware and software is born, and with it the architecture that, through von Neumann, becomes the model of all modern computers.

Learning machines

In the last chapter of the 1950 article, Turing discusses the topic of *learning machines*, starting by responding to the criticism made by Lady Lovelace in 1843 of the *Analytical Machine* (first prototype of a mechanical computer developed to execute generic tasks) of Charles Babbage. Lovelace translated and commented on an article written in French by the

Italian engineer and mathematician Luigi Federico Menabrea (who later became Prime Minister of the Kingdom of Italy) that described the mechanical functioning of the Analytical Machine, based on a lecture given by Charles Babbage himself in Turin in 1840. The Lady did not limit herself to translating Menabrea's article but added a series of comments in an appendix that constitutes two thirds of the work. In "note G"[2] Lovelace states that the machine «has no pretensions whatever to originate anything. It can do *whatever we know how to order it to perform*» (her italics) (Menabrea, Lovelace, 1843).

According to Turing, the incapacity of a machine to create or "originate something new" is not an insurmountable limit; at most it depends on the complexity and reactivity of the machine. Turing divides machines, and by analogy also minds, into two categories: subcritical machines and supercritical machines. The former are able to provide a limited response, without ever generating more than they receive as a stimulus, as a piano string returns to rest after being struck by the hammer. The latter trigger a chain reaction of secondary and tertiary thoughts, in an exponential process, like atomic piles. Most animal minds are subcritical, but a brilliant human mind is reasonably supercritical. Turing's question becomes: can a machine be made supercritical? (Turing, 1950). This question leads to reflection on the process that brings a human mind to the state in which it finds itself. In this Turing highlights three resolving components:

a. The initial state of the mind, say at birth.
b. The education to which it has been subjected.
c. Other experiences, not to be described as education, to which it has been subjected.

Moreover, it is reasonable to think that it is more difficult to arrive at the UTM by trying to program an "adult machine" directly, given its level of complexity and the fact that its current state depends on a previous educational process. For this reason, Turing proposes programming a "child machine", that is, a machine that simulates the state at birth, and then subjecting it to an adequate educational process. The problem thus decomposed - the child machine and the educational process - represents a radical turning point for computer science and for the evolutionary history of AI, anticipating what will be the central mechanisms of *symbolic AI* and *Machine Learning* models.

[2] Lovelace's "Note G" also became famous because it contains a description of a detailed method for calculating Bernoulli numbers, a method recognized as the first computer program in history.

How to constitute the complexity of the child machine in the initial state? In this regard Turing proposes two programming strategies: the first is the "child machine" by definition (anticipation of *Machine Learning*), that is, one follows the idea that the child mind is imagined as a notebook with «little mechanism and lots of blank sheets» (Turing, 1950). Alternatively, it is possible to equip the machine with a complete system of logical inference. In this case the store of the machine will be partially occupied by definitions, axioms and propositions ready for use (anticipation of *symbolic AI* models).

The risk of adopting the logical system is that of combinatorial explosion. The inference system (i.e. the instruction table) does not tell the machine what is "efficient" to do, in terms of computational costs, but only what is "lawful" to do. In this sense, at every step of a certain calculation mechanism, the machine finds itself faced with a large number of possible alternatives. If the machine had to explore them all, the number of combinations would rapidly explode to infinity. To avoid this, what Turing calls "imperatives" are necessary, higher-order instructions distinct from the rules proper to the system. The imperatives have these characteristics:

a. They do not necessarily have a hierarchy of type.
b. They are contextual and empirical, that is, experiential. For example, an imperative might suggest not using a slow method if a faster one has been tried.
c. They are heuristic because they indicate the most promising path, not the most certain one.

Therefore, in one sentence, the imperatives are higher-level instructions for "regulating the order" of the logical rules (or choices). The difference between a "brilliant" machine or a "silly" (*footling*) one lies precisely in the presence of these imperatives.

This intuition of Turing will find its first practical application in the *Logic Theorist*, discussed at the Dartmouth conference, as we will see later.

1.2. Dartmouth Conference: birth of AI

As we have already seen at the beginning of Chapter 1, the objectives of the summer conference at Dartmouth – the nominal birthplace of AI – follow three particular lines of research. Attention is concentrated on making «machines use language, form abstractions and concepts, solve kinds of problems now reserved for human beings and improve themselves» (McCarthy, J. & Minsky, M. & Rochester, N. & Shannon, C.E., 1955/2006).

The seminar, in addition to taking up the Turingian conjecture of the total "simulatability" of human intelligence and learning by machines, concentrates its effort on the study of language. Inheriting the conceptual framework – mathematical from Shannon and operational from Turing – it is stated that a large part of human thought consists in manipulating words [or symbols][3] according to rules of reasoning and conjecture (McCarthy, J. & Minsky, M. & Rochester, N. & Shannon, C.E., 1955, 2006). Already here an element of novelty can be noted: the passage to heuristic programming.

In the previous section we said that Turing elects language as an operational and testing criterion (i.e. *imitation game* or Turing test), that is, as a means for verifying whether a machine is capable of replicating (simulating) the function of thinking. Turing does not explain *how* the machine must generate language, but concentrates on the functional structure (the internal states) that mediates between input and output and also introduces the concept of "imperatives", higher-order instructions that regulate the machine's logical choices in order to avoid combinatorial explosion. Dartmouth sets itself to implement the concept of "imperatives" by seeking the empirical and practical rules (*rules of thumb*) for solving combinatorial explosion in practice and efficiently, that is, the "heuristics". The term, which indicates precisely the intuitive mental shortcuts that the brain uses with respect to complex problems, does not explicitly appear in the proposal, but several elements suggest it.

The attention placed on heuristics is supported by the fact that from Turing to Dartmouth one passes from *universality* (cf. UTM) to *specificity*. The researchers at Dartmouth did not question the *universality* of the Turingian model, but hypothesized that every aspect of learning or any other characteristic of intelligence can in principle be described so precisely as to be able to make a machine to simulate it (McCarthy, J. & Minsky, M. & Rochester, N. & Shannon, C.E., 1955, 2006). Even though this does not emerge from McCarthy's article, it is attested that at the Dartmouth conference the *Logic Theorist* was discussed, a program developed by Allen Newell, Herbert Simon and Clifford Shaw (cf. sec. 1.3) that employed heuristic methods to solve the theorems stated by Russell and Whitehead in the Principia Mathematica (Bianchini, 2007). Moreover, if attention is directed to specific processes and not to universal (global) functional processes, the architecture to be implemented must also change. We pass, nominally, in fact from simple and tripartite architecture (store, executive

[3] The addition of this word serves to make it clear that the Dartmouth conference ushers in the era of symbolic AI.

unit and control) to a complex and capillary structure: the neural network. The conference does not invent or definitively develop this architecture, but proposes to continue the works of Uttley, Rashevsky, Farley, Clark, Pitts, McCulloch, Minsky, Rochester, Holland and others.

The objectives that the Dartmouth conference proposed to realize – those we have described so far – were reasonably optimistic, suffice it to consider that their achievement was expected to be completed in the two months during which the summer conference was to take place. The strong optimism of the *Dartmouthians*, for what has been said, also attracted distrustful opinions toward the new field of research (AI), so much so that funding (and therefore progress) in AI has gone through numerous, very evident cycles of expansion and braking, with the occurrence of the so-called "AI winters", during which the field of studies was substantially out of favor with the government and industrialists (Kaplan, 2018).

Moreover, it is not clear what the actual results of the conference were, since the expected final report was never produced. It can, however, be affirmed that AI begins, at least nominally, at the Dartmouth conference, which marks the main research directions of this new disciplinary field.

1.3. Developments: from GOFAI to LLMs

From then on, how were the problems raised addressed? We have seen previously that Turing proposed two different strategies for programming a "child machine", the strategy of the pseudo-tabula rasa or the endowment of a system of basic logical inferences incorporated into the machine at the "state of birth". The latter (the logical approach) became prevalent in the field of AI both at Dartmouth and in the 1970s and 1980s of the twentieth century. This approach contributed to forming the paradigm of symbolic AI, also called *Good Old Fashioned AI* (GOFAI, to distinguish it from post-symbolic AI or sub-symbolic AI).

1.3.1. Symbolic AI (GOFAI)

We now trace and discuss the evolution of the most important symbolic AI models.

THE LOGIC THEORIST (1956)

We know that at Dartmouth the program *The Logic Theorist* (LT) by Newell and Simon was also discussed. This program embodied the ideal of solving mathematical theorems, in particular those found in the fundamental work *Principia Mathematica* by Alfred North Whitehead and Bertrand Russell. The program successfully proved 38 of the first 52

theorems in chapter two of the Principia. Surprisingly, it also found new proofs, sometimes shorter than those previously obtained by human hand.

On receiving the Turing Award in 1976, Newell and Simon gave an explanation of what they themselves called the "physical symbol system hypothesis". A physical symbol system has the necessary and sufficient means for general intelligent action (Newell, Simon, 1976).

Then they argue that:

> *«By "necessary" we mean that any system that exhibits general intelligence will prove upon analysis to be a physical symbol system. By "sufficient" we mean that any physical symbol system of sufficient size can be organized further to exhibit general intelligence. By "general intelligent action" we wish to indicate the same scope of intelligence as we see in human action: that in any real situation behavior appropriate to the ends of the system and adaptive to the demands of the environment can occur, within some limits of speed and complexity»* (Newell, Simon, 1976).

The minimal hypothesis enucleated with these arguments is that symbols are at the root of intelligent action (Newell, Simon, 1976).

Following this idea, the structural architecture of LT provides: (1) a knowledge base containing the axioms and theorems, (2) a set of inference rules and (3) a control unit.

Every logical entity of the axioms and theorems (whether variables or also logical connectives) provided as base (or memory) to the machine is reduced to a *token*. These tokens are combined into *symbolic structures* (expressions), so that at any instant of time the system will contain a collection of these symbolic structures (Newell, Simon, 1976). The inference system operates on the expressions to produce other expressions, with processes of creation, modification, reproduction and destruction (manipulation). The third structural element of LT, the control unit, is what allows promising paths to be prioritized by exploiting heuristics, precisely in order to avoid the combinatorial explosion of which we have spoken so far.

How does the manipulation of symbols occur practically in the heuristic search process?

In this model, it is based on the transformation of *problem solving* into a cyclic process of generation and testing of symbolic structures. The control unit is therefore divided into a generator and into a test, so that:

> *«To state a problem is to designate (1) a test for a class of symbol structures (solutions of the problem) and (2) a generator of symbol structures (potential solutions). To solve a problem is to generate a structure, using a generator, that satisfies the test. We have a problem if we know what we want to do*

(the test) and if we do not know immediately how to do it (our generator does not immediately produce a symbol structure that satisfies the test). A symbol system can state and solve problems (sometimes) because it can generate and test» (Newell, Simon, 1976).

The key to intelligence lies in making the generator selective, that is, making the generation of new expressions not stochastic but heuristic.

If the whole process resides in symbolic manipulation, and if the objective is the demonstration of a thesis, the generator starts from an axiom and produces a modified version of it that is symbolically more similar to the symbolic structure of the thesis. The generator does not know a priori which symbolic structure is most similar: it makes an attempt (some, not all), then the test unit compares the differences between the current symbolic structure and the target one. If the test notes a difference, it selects (heuristically) the logical operator (contained in the inference system) specifically designed to reduce that particular difference, ignoring the other operators. An example reported directly by the authors consists in the solution of an algebraic equation:

«Consider the problem of solving a simple algebraic equation: AX + B = CX + D. The test defines a solution as any expression of the form X = E, such that AE + B = CE + D. Now one could use as a generator any process that produced numbers which could then be tested by substituting them in the preceding equation. We would not call this an intelligent generator. Alternatively, one could use generators that exploit the fact that the original equation can be modified by adding or subtracting equal quantities from both sides, or multiplying or dividing both sides by the same quantity, without changing its solutions. But, of course, we can get still more information to guide the generator by comparing the original expression with the form of the solution and making precisely those modifications to the equation that leave its solution unchanged, while at the same time bringing it into the desired form. Such a generator might notice that there was an unwanted CX on the right side of the original equation, subtract it from both sides, and collect terms again. It might then notice that there was an unwanted B on the left side and subtract it. Finally, it might eliminate the unwanted coefficient (A - C) on the left side by dividing. Thus, with this procedure, which now shows considerable intelligence, the generator produces successive symbolic structures, each obtained by modifying the preceding one; and the modifications aim to reduce the differences between the form of the input structure and the form of the test expression, while at the same time maintaining the other conditions for a solution» (Newell, Simon, 1976).

LT implements a rudiment[4] of what Newell and Simon would later define as *means-ends analysis*. By updating the symbolic structures and the system memory, the demonstrative

[4] Means-ends analysis was nominally defined only with GPS (General Problem Solver), also created by Newell and Simon in 1957.

process is progressively directed toward the objective. In this sense, while remaining entirely internal to a symbolic domain, the heuristic process preserves a structure functionally akin to cybernetic retroaction: the system generates a configuration, evaluates its distance from the goal-state and selects the operator capable of reducing that gap.

In summary, heuristics constitute the strategies selected by the control unit, in the constant dialogue between generator and test, to minimize the syntactic distance between the symbolic structures that follow one another starting from the axioms in order to arrive at the final thesis.

The entire mechanism of symbolic manipulation is possible thanks to a specific programming language, *Information Processing Language* (IPL), also created by Newell and Simon for the realization of their LT.

The unprecedented novelty of IPL and, more specifically, of the process of list processing, lies in the concept already introduced of physical symbol system. Before 1956 the data contained in memory cells were considered inert, that is, "still" and without any connection with the other memory cells. Newell and Simon realize that, if it is possible to transform every logical entity or inference operator into a symbol, it is possible to do the same with the addresses of memory cells. In doing so, it is possible to allocate in a memory cell a symbol that represents the address of another cell, that is, to create lists that connect different cells and to pass from the concept of fixed memory to the concept of dynamic memory. This capacity of the address-symbol to "point" to another cell is what is called *designation*. This "invention" produced no little disorientation in the scientific community:

> *«That this was a new view was shown to us many times in the early days of list processing, when colleagues asked us where the data were, that is, which list finally contained the collections of bits that constituted the content of the system. They found it strange that there were no such bits: there were only symbols that designated other symbolic structures.»* (Newell, Simon, 1976)

LISP (1958)

If with LT we have the first model of reasoning of symbolic AI, with LISP one passes to the generalization of symbolic programming. LISP (*List Processor*) is not properly an AI model but a programming language and, in fact, it presents itself as an implementation of IPL. Newell and Simon state that LISP completed the act of abstraction, lifting list structures from their embodiment in concrete machines (Newell, Simon, 1976). In other words, the creation

of a memory that is no longer rigid but dynamic makes it possible to operate on all "types of data" without having to worry about how the machine physically stores them. In LISP, dynamic symbolic memory does not depend only on the possibility of constructing list structures through designation, but also on the automatic management of memory space. *Garbage collection* indicates precisely the automatic recovery of portions of memory no longer reachable from the active symbolic structures. In this way, the machine can continue to construct, modify and abandon complex symbolic structures without requiring an explicit manual de-allocation.

One of the most important points that makes LISP decisive in the evolution of AI is the concept of *homoiconicity*. A language is homoiconic if a program written in it can be manipulated as data using the language (Ceravola & Joublin, 2021). Still within the hypothesis of the physical symbol system, it is possible to identify the program itself of a machine (i.e. a list of instructions) with a list of symbols. Since the system is able to manipulate symbols, it then becomes able to manipulate code, that is, the instructions themselves. Code, being symbol, is data for the machine. In other words, LISP implements the possibility that the system can modify itself.

This identity between code and data was fundamental for the development of the first automatic learning systems. Over the decades following its creation, LISP became the dominant programming language in AI.

SHRDLU (1970)

With SHRDLU another important stage in the history of AI is marked, with respect to NLP. Terry Winograd, with his doctoral thesis defended at MIT, presented his demonstrative program in 1970, called precisely SHRDLU, a name taken from the second column of letters of a typewriter. The program answered questions posed in natural language and moved blocks here and there in a virtual world (Kaplan, 2018). More precisely, the program operated in a specific virtual environment called *Blocks World* (BW) in which there were geometric shapes such as pyramids and boxes of different colors. The implementation of SHRDLU rested mainly on the LISP language (see previous section), which constituted its logical-syntactic infrastructure. The program allowed the user to give commands such as moving or positioning the blocks in particular ways in the BW, in Fig. 1. All this occurred in a question-and-answer dynamic, so much so that SHRDLU did not limit itself to executing

instructions, but if these had been ambiguous it could ask for clarifications instead of freezing, or it was able to give explanations about actions carried out previously.

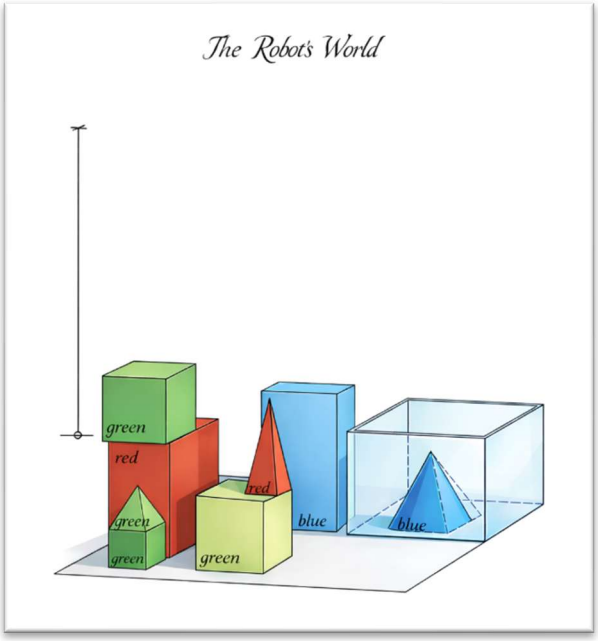


Figure 1: AI-generated reconstruction of SHRDLU's Blocks World.

Why did SHRDLU highlight the limits of the logical-symbolic approach? Before SHRDLU, other symbolic AI programs had been conceived and realized, including ELIZA, recognized as the first *AI chatbot* in history (implemented on pattern matching mechanisms). Winograd expressed reservations about its modelling in these terms:

> *«[this program] would respond to a question like "What time is it?" by asking in return "Why do you want to know what time it is?", or muttering "You want to know what time it is!". This can be done without any understanding. All it has to do is take the words of the question and rearrange them in a simple way to create a new question or statement. In addition it recognizes a few key words, to respond with a fixed phrase whenever the patient uses one of them»* (Winograd, 1971).

Beyond specifically criticizing ELIZA program, Winograd exposes what are the limits of the logical-symbolic approach most used up to that moment. Pure symbolism, that is, pure syntactic manipulation, expressly eliminates all semantic content, that is, it produces responses that are «meaning-less». SHRDLU highlighted the fragility of symbolic systems in the face of the complexity and ambiguity of natural language outside a restricted domain (Singh, Banerjee, & Ghosh, 2025). What Winograd proposes is the necessary next step, thus made explicit:

> *«If we really want computers to understand us, we need to give them the ability to use more knowledge. In addition to a grammar of the language, they need to have all sorts of knowledge about the subject they are discussing, and they have to use reasoning to combine facts in the right way to understand a sentence and respond to it. The process of understanding a sentence has to combine grammar, semantics and reasoning in a very intimate way, calling on each part to help with the others»* (Winograd, 1971).

The addition of semantics is taken on by the programmer, who provides the program's basic knowledge with all the instances necessary to carry out the mechanisms of grammar, semantics and reasoning.

How is this new step practically realized in SHRDLU? Winograd implements a "vertical" architecture in which grammar, semantics and reasoning are expressed, as a minimal list, in the actions of *parsing* (i.e. syntactic analysis), *meaning* and *thinking about the meaning*, respectively. The verticality of the architecture is, according to Winograd, essential so that the actions of reasoning are not sequential (as they would be in the purely syntactic approach) but in a dynamic and continuous relation with one another. Syntactic analysis itself is «designed to deal with semantic questions» (Winograd, 1971). Instead of being an action of analysis of syntactic rules alone in order then to reformulate symbolic structures heuristically functional to the purpose – as it was for LT – the *parsing* of SHRDLU is the way in which language is organized around the choices for conveying meaning (Winograd, 1971). In other words, it is not enough to analyze the structure of the sentence, but it is necessary for the pre-established aims to abstract the linguistic characteristics that are important for interpreting its meaning. To perform this abstraction «the system has a network of semantic FEATURES[5] […] used for an initial phase of semantic analysis. The features subdivide the world of objects and actions into simple categories, and the semantic interpreter uses these categories to make some of its choices between alternative definitions of a word» (Winograd, 1971).

For example,[6] in the two instructions "take the cube" and "take note of the position", the word "take" assumes a different meaning depending on the context, more precisely depending on the type of object to which that "take" refers. Semantics identifies the object of reference and chooses. If the object is [GRASPABLE], as in the first case, the system activates the physical procedure of movement of the mechanical arm. If the object is

[5] We use uppercase after Winograd, and as a link to the semantic features cited in the example below.

[6] The example given here is not from Winograd, but is freely drawn from the analysis of the paper.

[INFORMATION], as in the second case, the system interprets "taking" as an operation of writing into memory. In this example, what is being indicated in square brackets are precisely the semantic features we have spoken about.

Moreover, unlike LT where one even arrives at proving new theorems but always within a logic that we could define as "static" (in the sense that the new proofs were already true in potency from the axioms and the new knowledge remains within the well-delimited field of the *Principia Mathematica*) and, therefore, within the use of an accumulative memory, with SHRDLU we have a model that interacts with a dynamic world, the BW, which is modified every time the program performs an action. SHRDLU must therefore adapt its internal memory because the "reality" in which it operates is changed because of its intervention. The comparison is synthetically simple: static world, static memory; dynamic world, dynamic memory. It is not by chance that this model is implemented on LISP, a language that allows the program to determine the meaning of words through their reference (incorporated within the concept of *designation* cited previously) to objects, properties and specific events within its virtual environment. Conversely, the effectiveness of SHRDLU confirmed that the LISP language was the ideal instrument for managing complex symbolic lists and for treating code as data (see *homoiconicity*), allowing the program to plan actions.

SHRDLU is innovative precisely because, for the first time, language engages in a dynamic relation with a world, even if it is virtual. This characteristic will return as decisive in the development of LLMs (*Large Language Models*). Historians of computer science today use the term "grounding" (coined by Stevan Harnad in 1990)[7] to describe SHRDLU's capacity to connect symbolic systems to physical referents (virtual in the case in question).

What implements this new dynamic? All this architectural framework allows a further comparison between LT and SHRDLU. In the first, the process of reasoning was defined as *heuristic search*. With the second, one passes to *procedural semantics*. Ontologically, it is a change in the very concept of knowledge: in doing so SHRDLU represents knowledge in the form of procedures rather than tables of rules or lists of patterns (Winograd, 1971).

The architectural characteristics described up to this point allowed SHRDLU to sustain a "significant" question-answer exchange with a user, responding at least ideally to Turing's hypothesis of 20 years earlier. If it did not have a mechanism of semantic analysis, it could

[7] To delve into the topic see Harnad (2002)

not have “awareness” of the world in which it is immersed and, therefore, would not be able to answer specific questions about that world.

It is appropriate to underline that there is no intention here to support, without due caution, the hypothesis of self-consciousness on the part of these programs with respect to their own operation, nor that of evaluating what degree of “awareness” they have of it. For example, according to critics such as Searle, SHRDLU remains a system of formal manipulation, since it does not possess real sensory experience or intentionality. The language used to describe the phases of the “reasoning” of a machine necessarily derives both from the mentalistic setting of early cybernetics and from the lack of a subsequent and appropriate terminological differentiation between algorithmic processes and human thought. In any case, the scope and nature of such theoretical implications are complex themes that would deserve a dedicated extemporaneous reflection.

Numerous other symbolic AI models followed one another, including ELIZA (already cited), DENDRAL, MYCIN and PROLOG, but the evolutionary history of AI has shown that, «with all its fascination, the physical symbol system hypothesis is not the only possible one» (Kaplan, 2018).

1.3.2. Machine Learning and Symbolic AI

Up to this point, Turing’s idea of *learning machines* had not yet really been pursued. All symbolic AI models concentrated on the interpretation of language, through purely syntactic procedures or with in-depth treatments at the semantic level. This was made possible by following the physical symbol system hypothesis proposed by Newell and Simon and by aprioristically endowing each model with a knowledge base. Normally, while considering learning as a dimension not necessarily antithetical to the physical symbol system hypothesis, if in an application that follows the symbolic systems approach there is learning, this occurs in advance, in order to develop the symbols and rules that ultimately will be assembled and used for the pre-established application (Kaplan, 2018). Therefore, what has been discovered up to this point about the capacity of machines to learn?

If in symbolic models knowledge is to a large extent supplied a priori in the form of symbols, rules and procedures, with Machine Learning the center of gravity shifts toward *data-driven* forms of learning. In epistemological terms, this passage can be interpreted as a shift from the prevalence of deduction to the centrality of inductive generalization: the system does not

only apply already formulated rules but learns general regularities starting from sets of examples.

Following Peirce's theoretical framework (cf. Peirce, 1934), the tripartition is the following:

- Deduction (Rule + Case → Result) is the analytic application of a Rule to a Case in order to predict a Result. It is the only method that necessarily preserves truth (for example, executing code to verify the output).
- Induction (Case + Result → Rule) is the synthetic derivation of a Rule from the accumulation of Cases and Results. It validates hypotheses through statistical frequency (for example, generating a function to satisfy unit tests).
- Abduction (Rule + Result → Case) is the inference of a Case (or of a new Rule) to explain a surprising Result.

However, in the case of Deep Learning and LLMs (as we will see later), this induction should not be understood as the explicit formulation of rules, but as statistical optimization of parameters that allows the model to approximate functions or probability distributions starting from data. In summary, induction in general and grounding on data specifically is the capacity to extract general patterns from specific data based on statistical frequency. In this conceptual shift, the programmer no longer has to provide in advance the "rules of behavior" to the model but merely designs the architecture and provides the training data.

Recalling what has already been said previously, Turing decomposes the entire problem of learning machines into two points: the realization of a child machine and the implementation of an educational process. Although it is incautious to identify totally the dynamic of human learning with processes of induction, it is possible to associate the educational process of machines with the induction that ML brings to the fore. It can be maintained that the educational process proposed by Turing, composed of positive and negative reinforcements and instruction, is inductive in the general epistemological sense: the machine analyzes a series of particular examples in order to generalize stable patterns.

This is exactly what the mechanisms of supervised learning in machine learning do (we will see later): labeled examples are given, the machine adjusts the parameters, a generalizing function emerges.

SOAR

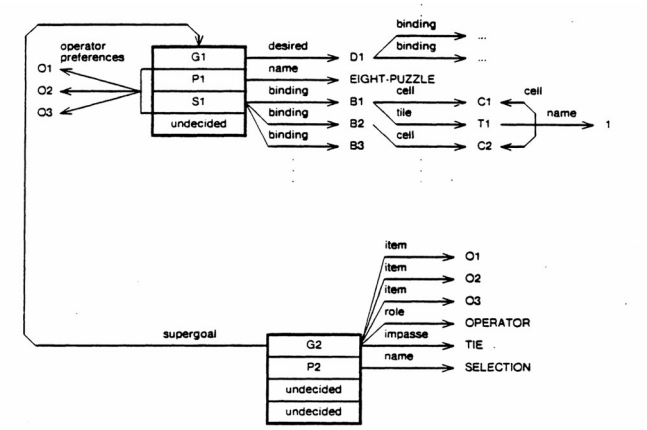
SOAR (*State Operator And Result*), developed by John Laird, Allen Newell and Paul Rosenbloom in 1987, represents the first attempt to overcome the limits and rigidity of symbolic systems. «Soar is an architecture for a system that is to be capable of general intelligence» (Laird, Newell, Rosenbloom, 1987). With these words the descriptive article on SOAR opens. The term *general intelligence* was taken up in 2002 by Ben Goertzel, Cassio Pennachin and Shane Legg, to give the title to the book on which they were working. It was Legg who suggested the term Artificial General Intelligence (AGI, coined previously by Mark Gubrud, 1997). Goertzel, Pennachin and Legg are credited as the main popularizers of the term (cf. Goerztel, Pennachin, 2007). When Goertzel and his colleagues coined the term AGI, this referred to the capacity of AI to be able to generalize, that is, to be able to solve tasks pertaining to disparate domains, without being previously programmed for the specific tasks. In some way, Goertzel wanted to distinguish AGI from narrow AI (ANI, *Artificial Narrow Intelligence*), that is, "restricted AI", a term that also emerged at the beginning of the 2000s to indicate AI designed to carry out a specific task and highly efficient in that specific domain. Over the years the term AGI has undergone a semantic drift. The term today is used to indicate the maximum development that AI modelling will reach in its evolutionary parabola, insofar as by *general* one means an algorithm capable of learning, understanding and carrying out any intellectual task that a human can carry out.

According to the authors, SOAR was designed to be able to: «(1) work on the full range of tasks, from highly routine to extremely difficult and open-ended problems; (2) employ the full range of problem-solving methods and representations required for these tasks; and (3) learn about all aspects of the tasks and its performance on them» (Laird, Newell, Rosenbloom, 1987).

The structure of SOAR's aims revolves around the concept of *problem-solving*. The latter is not an external or juxtaposed function to the model. The architecture itself is, intrinsically, a *problem-solving system*. Its three pillars are: memory, decision cycle and learning.

The structure of memory. SOAR divides memory into two families: working memory and productive memory (or *productions*), that is, procedural memory. The first contains the representation of the current situation (what the system "is looking at"), structured as a symbolic graph facilitating a flexible representation of knowledge (Laird, Newell, Rosenbloom, 1987). The second contains a series of instructions to follow depending on the

conditions of the problem. These instructions are stored as rules of the IF-THEN type ("IF X happens – THEN do Y"), namely links of logical causation or condition-action pairs. Described in this way, procedural memory represents – without differences at the conceptual level – the knowledge system provided as a base in all symbolic AI models.




```
(goal G1 ↑problem-space P1 ↑state S1 ↑operator undecided ↑desired D1)
  (desired D1 ↑binding DB1 ↑binding DB2 ...)
 (problem-space P1 ↑name eight-puzzle)
(state S1 ↑binding B1 B2 B3 ...)
 (binding B1 ↑cell C1 ↑tile T1)
  (cell C1 ↑cell C2 ...)
  (tile T1 ↑name 1)
 (binding B2 ↑cell C2 ...)
  (cell C2 ↑cell C1 ...)
 (binding B3 ...)
 ...
(preference ↑object O1 ↑role operator ↑value acceptable
        ↑problem-space P1 ↑state S1)
(preference ↑object O2 ↑role operator ↑value acceptable
        ↑problem-space P1 ↑state S1)
(preference ↑object O3 ↑role operator ↑value acceptable
        ↑problem-space P1 ↑state S1)
 (operator O1 ...)
 (operator O2 ...)
 (operator O3 ...)

(goal G2 ↑problem-space P2 ↑state undecided ↑operator undecided
        ↑supergoal G1 ↑role operator ↑impasse tie
        ↑item O3 ↑item O2 ↑item O1)
 (problem-space P2 ↑name selection)
```

Figure 2: left, snapshot of fragment of working memory. Right, working memory representation of the left structure (Laird, Newell, Rosenbloom, 1987).

Working memory is mainly composed of:

1. Context stack (or stack of context, the boxed parts in Fig. 2): it is an ordered sequence of contexts, each of which contains four *slots*, that is, a quadruple $(g, p, s, o) = goal$, $problem-space$, $state$ and $operator$. The upper context contains the priority objective; the lower contexts contain sub-objectives of the first. Therefore the context stack is an ordered hierarchy of dependent quadruples. The structural dependence is strong: if an object in a slot is modified or replaced, all the slots below it in that context are instantiated differently.
2. Objects: they are any entities defined by a symbol, *identifier*, and a set of *augmentation*.
3. Preferences: they are symbolic judgments, absolute or comparative, through which SOAR establishes which object must be selected in a slot of the context, for example which operator to apply among several possible operators.

Augmentations are triples of symbolic relation identifier-attribute-value. For example, if we take in Fig. 2 the row

$$(goal\ G1\ \hat{}problem-space\ P1\ \hat{}state\ S1\ \hat{}operator\ undecided\ \hat{}desired\ D1)$$

within it we have four augmentations of the object G1, which are:

- $G1, problem\text{-}space, P1$
- $G1, state, S1$
- $G1, operator, undecided$
- $G1, desired, D1$

Augmentations do not constitute a simple implementative detail, but the fundamental minimal unit of symbolic representation within SOAR. They guarantee access to memory, the management of contexts and the generality of learning. Therefore in SOAR, working memory does not manipulate single or compact objects, but networks of symbolic relations, true symbolic graphs and sub-graphs that establish the structural dependence among all the objects involved. In doing so the system can always determine which objects are effectively accessible starting from the *context stack* by following, backward, chains of *augmentations* and *preferences*. Here one finds again, in a more evolved form, the same intuition that emerged with list processing: a symbolic memory that is not monolithic but configured on the relations among its elements.

This *augmentation* structure does not remain an abstract concept but becomes operative in the representation of the state of the problem. In the *eight-puzzle* problem (which we will comment on better below), the state is not encoded as a monolithic and rigid configuration of the grid. The connections between the tiles and the cells are specified by objects called *bindings*. A given state, such as S1 (Fig. 3), consists of a set of nine *bindings*. Each *binding* points to a tile and a cell; each tile points to its value; and each cell to the adjacent cells» (Laird, Newell, Rosenbloom, 1987). The operators manipulate only the bindings; the representation of the cells and of the tiles does not change (Laird, Newell, Rosenbloom, 1987).

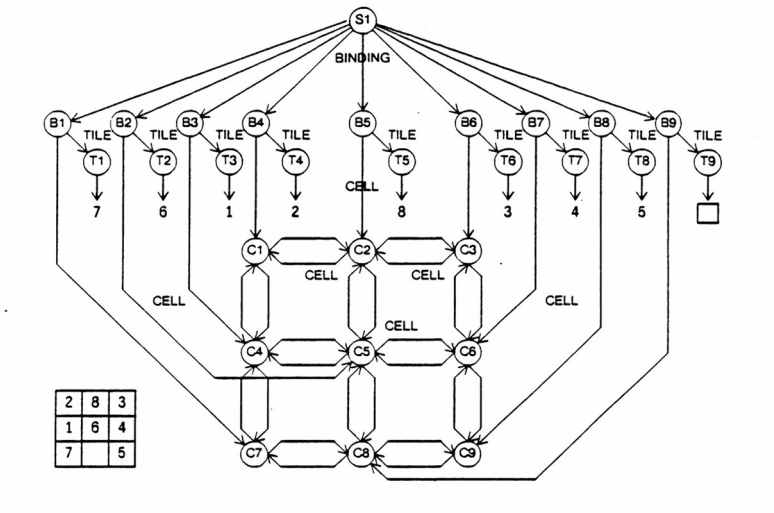


Figure 3: Graphic representation of an eight-puzzle state (Laird, Newell, Rosenbloom, 1987).

This articulation of working memory does not have only a descriptive function. Since states are decomposed into elementary relations and their elements remain accessible through structured chains of *augmentation*, the architecture can identify precisely which aspects of the situation are relevant for the current *problem solving*. This representational granularity is necessary in order to assist the process of *chunking*, described below.

The decision cycle. An element of novelty of SOAR with respect to the past is represented by the decision cycle (see Fig. 4). It consists of two distinct phases alternated continuously: the *elaboration phase* and the *decision procedure*.

During the elaboration phase, new objects, new augmentations of existing objects and preferences are added to working memory, with respect to the context in which one is located (i.e. the initial state and the desired one are instantiated, or the operators that will serve in the decision phase). This is a monotonic process (the elements of working memory are not erased or modified) that continues until quiescence is reached because there are no more elaborations to generate (Laird, Newell, Rosenbloom, 1987). The elaboration phase proceeds thanks to *productions,* that is, the elements of long-term memory or productive memory, namely the IF-THEN rules of which we spoke earlier. For example, when the elaboration phase reaches quiescence, the *decision procedure* begins, which examines the accumulated preferences and the context stack.

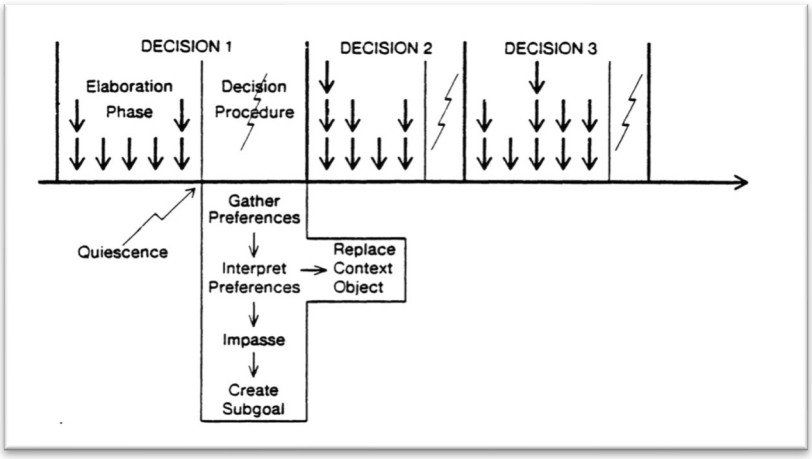


Figure 4: SOAR decision cycle (Laird, Newell, Rosenbloom, 1987)

If the available knowledge is not sufficient to solve a problem or a task, an *impasse* occurs (trans. from Fr. "dead end", Treccani, n.d.). In this case, the system does not continue reasoning within the same problem space, but creates another one, creating a *subgoal* (sub-objective).

Let us take up the example studied by the authors, reported in Fig. 5. The problem to be solved is the well-known *eight puzzle*, a classic problem in which a 3x3 grid of cells is given, eight of which are randomly occupied by the numbers from 1 to 8. The problem consists in transforming an initial state into a desired state, performing only lawful moves, that is, the movement of one of the numbered cells into the only available empty cell. The numbered cells that can be moved are only those adjacent (to the right, to the left, above or below) to the empty cell. In the situation shown in Fig. 5 the system finds itself choosing three different operations but all equally possible: moving the number 6 down, moving 7 to the right or moving 5 to the left. Thus, an *impasse* occurs and the consequent opening of a *sub-goal*. In this *subgoal*, the system creates three instances corresponding to the three movement operators. To evaluate which is the best choice, it is necessary to evaluate what result is produced by the application of each operator. The test of the operators occurs in a further subgoal (*evaluation subgoal*), testifying to the fact that a branching hierarchy of sub-objectives is created. It is straightforward that solving each capillary objective gradually leads to solving the branching, proceeding in reverse and returning to the main objective, saving memory and computational costs. The economy of this process is favored by *bindings* (see above), representations that allow one to "move" tiles without rewriting the whole state of the puzzle grid.

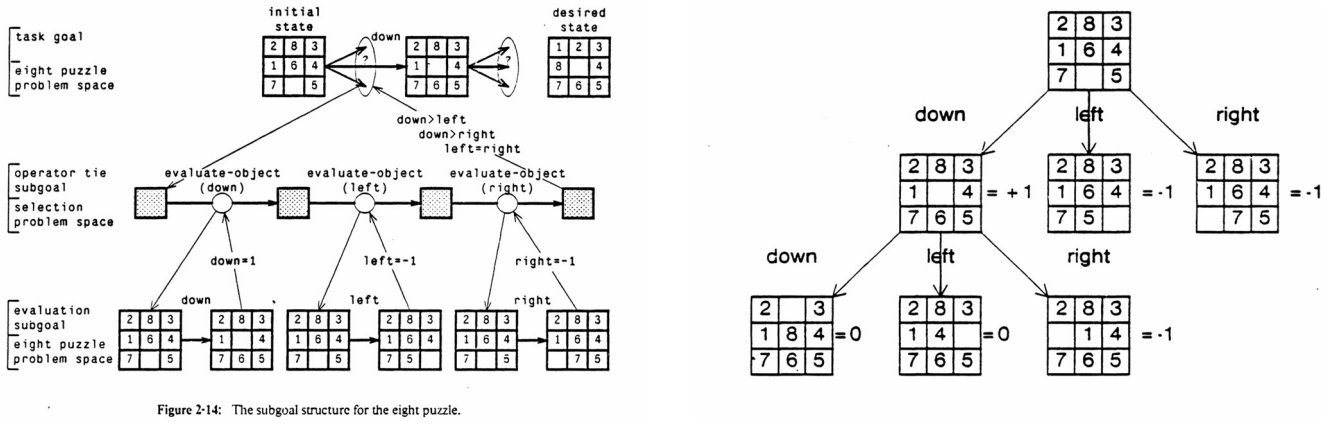


Figure 2-14: The subgoal structure for the eight puzzle.

Figure 5: Hierarchical structure of subgoals and evaluation of preferences in the eight puzzle. On the left, the hierarchy of subgoals automatically generated by the SOAR architecture to resolve the impasse, from the main task to the operator selection subgoals and their evaluation subgoals. On the right, the local evaluation of the three possible moves (down, left, right): down gets a score of +1 by bringing a tile to the correct position, left and right get -1, generating the down > left and down > right preferences that resolve the impasse in the upper level (Laird, Newell, Rosenbloom, 1987).

The evaluation of the applicative outcome of the movement operators is called evaluation of preferences. It consists of an evaluation of the local improvement or worsening of the actions of the operators. Specifically, the move is evaluated with +1 if the moved tile passes from "out of place" to "in correct position", -1 if it passes from correct to out of place, 0 if it passes from out of place to out of place. In the case shown, SOAR obtains

$$down = 1$$

$$left = -1$$

$$right = -1$$

From these results it constructs the preferences $down > left$, $down > right$, $left = right$, which resolve the impasse at the higher level. At that point the selection subgoal ends, SOAR returns to the main task, selects $down$ as the puzzle operator, applies it to the initial state, chooses the new resulting state and thus continues the problem solving.

Figure 5 read in sequence outlines these steps:

1. Main task: I must choose a move in the initial state of the puzzle.
2. Impasse: I do not have sufficient direct knowledge, therefore an impasse arises.
3. First subgoal: choose between movement operators down, left, right.
4. Evaluate-object: evaluation of the operators.
5. Second local impasse: to evaluate a move I must try it; therefore, an evaluation subgoal arises.

6. Test: I try down, then left, then right and obtain down = 1, left = -1, right = -1.
7. Preferences: I generate down > left, down > right, left = right.
8. Resolution: the subgoal ends, in the main task down is selected, and the puzzle continues from there, that is, the elaboration-decision cycle is iterated.

This whole process is defined by the authors as a process of *automatic subgoaling*, because the creation of subgoals occurs naturally as a consequence of the type of architecture thus constructed.

At the end of the decision phase, if the preferences generated in the subgoal are sufficient to determine a final choice, the system proceeds to replace in the context stack the pertinent object of the higher context. In the case discussed, the initially undecided operator is replaced by the operator that resulted preferable (Fig. 6).

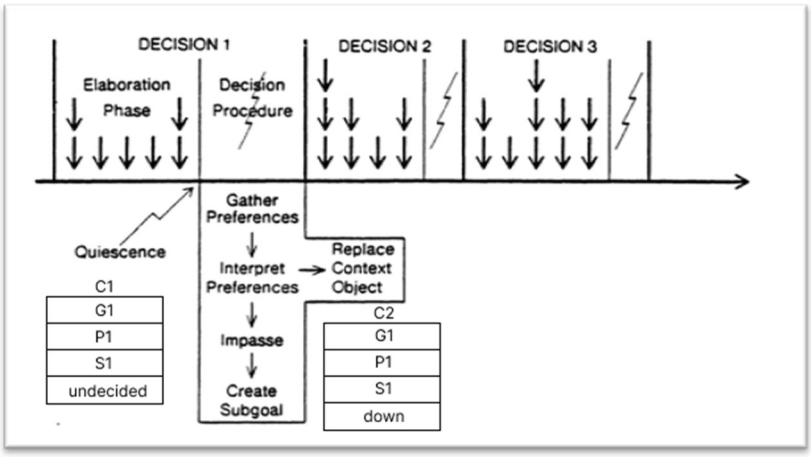


Figure 6: Updating the context stack at the end of the decision cycle. On the left, the state of the context with the operator still undecided; on the right, the updated state after the impasse has been resolved, in which the undecided operator is replaced by the down operator, which is preferable from the preference evaluation. The process-decision cycle is then iterated from the new state (Laird, Newell, Rosenbloom, 1987).

In general, the replaced object can be any one of those in the slots of the context (problem space, state or operator) depending on the type of impasse solved. If instead the preferences do not yet determine a final choice, the system does not proceed further but generates a new impasse. Once the subgoal is finished, working memory eliminates the elements created for the local elaboration and preserves only the results. On this basis *chunking* then intervenes.

Learning. The point of strong change with respect to purely symbolic models consists in the fact that SOAR implements a learning dynamic. In this architecture, learning occurs through a mechanism called *chunking*. The authors first recall the previous tradition and literature, according to which *chunking* is a learning scheme for automatically organizing and

remembering ongoing experiences on a continuous basis (Laird, Newell, Rosenbloom, 1987). SOAR starts from this definition and introduces a significant novelty. When one enters the cycle of analysis of sub-objectives, here learning is required because a subgoal is created if and only if the available knowledge is insufficient for the next step in problem solving (Laird, Newell, Rosenbloom, 1987), which is the condition of impasses. Therefore, SOAR does not learn in a random or continuous way over the entire process, but learns in localized moments, precisely when it encounters an insufficiency of knowledge. The subgoal serves to find the missing information; chunking transforms the results of the subgoal into a new production that the system will use automatically when in the future it encounters a situation sufficiently similar to the starting one, avoiding the reopening of the same subgoal. The representation of objects through augmentation plays a key role in determining the generality of learning in Soar. Generality is maximized when those aspects of a state that are functionally independent are represented independently (Laird, Newell, Rosenbloom, 1987). As we have already seen, in order to move a tile into an empty cell, the system does not have to modify in memory all the instances relating to the entire grid of the eight puzzle. It is sufficient to reorganize the order of the bindings.

SOAR combines the centrality of problem spaces with a uniform elementary representation (objects and their augmentation), and this combination makes it possible (as it was not before) to treat all tasks, even routine ones, as heuristic search.

Before SOAR, symbolic AI models centered their development around the concept of heuristic search, implementing what are called *weak methods*. With this expression one indicates general problem-solving strategies that can be applied to different domains precisely because they do not depend on a very rich specialist knowledge. In previous symbolic systems such methods were often treated as relatively determined procedures; in SOAR, instead, they tend to emerge from the interaction between architecture, available knowledge and decision cycle.

If *weak methods* are not fixed modules but emergent behaviors, then they can also be learned. If SOAR learns continuously by automatically and permanently storing the results of its sub objectives as productions. This single mechanism of *chunking* also produces the acquisition of new heuristics and of implementations for operators.

In conclusion, SOAR embodies a series of eleven basic hypotheses for modelling a cognitive architecture (not only a program) for a general intelligence:

1. **Physical symbolic system hypothesis:** A general intelligence must be realized with a symbolic system
2. **Goal structure assumption**: Control in a general intelligence is maintained by a symbolic system of goals.
3. **Uniform Assumption of Elementary Representation**: There is a single elementary representation for declarative knowledge.
4. **Problem-space hypothesis:** Problem spaces are the fundamental organizational unit of all goal-oriented behavior**.**
5. **Production system hypothesis**: Production systems are the appropriate organization to encode all knowledge in the long run.
6. **Universal sub-goal hypothesis**: Any decision can be the subject of goal-oriented attention.
7. **Automatic sub-goaling hypothesis**: All goals are dynamically born in response to impasse and are automatically generated by the architecture.
8. **Control Knowledge Assumption**: Any decision can be controlled by indefinite amounts of knowledge, whether domain-dependent or independent.
9. **Weak Method Hypothesis**: Weak methods are the basic methods of intelligence.
10. **Weak Method Emergence Hypothesis**: Weak methods arise directly from the system's response based on its knowledge of the task.
11. **Uniform learning assumption**: Goal-based chunking is the general mechanism of learning.

However, SOAR realizes only in part the capacities of general intelligence, with significant aspects still missing (Laird et al., 1987). According to scientific literature, SOAR is assigned to the family of ML but develops an architecture still tied to logical-symbolic approaches and, for this reason, it can be said to belong to the model of *Symbolic Machine Learning*. In summary, it is an ML model because it implements a cognitive architecture endowed with a mechanism of learning (chunking) based on experience, unlike previous models where learning was a priori, and it satisfies Turing's requests because it is a machine that self-improves. However, it is still a symbolic model because its reasoning follows explicit and encoded rules. It can be said that SOAR, while opening the field of development of ML, represents the apex of GOFAI.

1.3.3. Machine Learning and Sub-Symbolic AI

The passage from symbolic Machine Learning to sub-symbolic Machine Learning coincides with a decisive transformation in the way of understanding artificial knowledge. In symbolic systems, knowledge is represented through explicit and manipulable structures, such as logical rules, ontologies, planning procedures or inferential chains; in this sense, the symbolic paradigm remains coherent with the physical symbol system hypothesis, according to which general intelligence can be realized through the manipulation of symbolic structures (Newell & Simon, 1976). In sub-symbolic systems, instead, the relation between input and output is not formulated primarily in terms of interpretable symbols, but through numerical configurations distributed in the parameters of the model. This distinction is central in the *connectionist* tradition, which will interpret cognitive behavior not as application of explicit rules, but as the outcome of interaction in the network. In this sense, algorithms such as artificial neural networks are sub-symbolic because they learn statistical correlations between variables without necessarily translating them into humanly readable rules (Ilkou & Koutraki, 2020; LeCun, Bengio, & Hinton, 2015).

This transformation intertwines with the affirmation of the *Big Data Era*. The increase in data availability, together with the growth of computational capacity, makes it possible to train models capable of extracting regularities from increasingly large datasets; in this sense, the recent progress of Machine Learning depends as much on the development of new algorithms as on the explosion of available data and computation (Halevy, Norvig, & Pereira, 2009; Jordan & Mitchell, 2015). However, precisely this predictive effectiveness produces the epistemological problem of opacity, which does not depend only on the scale of the models, but also on the fact that learned knowledge takes the form of a distributed configuration of parameters, difficult to reduce to a finite set of readable instructions. For this reason, the literature on the opacity and interpretability of Machine Learning insists on the fact that the "black box" is not only a practical limit of access to code or data, but also concerns the difficulty of making intelligible the internal functioning of complex and non-linear models (Burrell, 2016; Lipton, 2018).

Machine Learning (ML) in the proper sense is the study of algorithms that make predictions to solve a specific class of problems. If in ML one studies *machines that learn how to predict*, such machines fundamentally have two algorithms: an algorithm for learning (i.e. algo1: LEARNER) and an algorithm for prediction (i.e. algo2: PREDICTOR). *Machine Learning*, however, needs another important ingredient, that is, data. In the last twenty years the

phenomenon of Big Data has been witnessed, born from the encounter between the ever greater production and collection of data of the digital era and, on the other hand, the evolution of the computational power of computer systems. It is precisely the need to manage and synthesize the vastness of modern datasets that has catalyzed the development of artificial intelligence systems that learn precisely from data.

The learning of these systems is articulated on three main paradigms: supervised learning, unsupervised learning and reinforcement learning.

Supervised learning (SL) concentrates attention on the concept of labeled data, that is, data in which each example is associated with a correct response or with a predefined class of belonging (Buzzi, Priori, 2025). Models trained with this supervision learn mappings between inputs and the corresponding outputs, inferring generalizations in order to be able to predict the classification of unlabeled data. «Applications in which the training data comprises examples of the input vectors along with their corresponding target vectors are known as supervised learning problems. Cases such as the digit recognition example, in which the aim is to assign each input vector to one of a finite number of discrete categories, are called classification problems» (Bishop, 2006). The learners most used in this type of learning are decision trees and support vector machines (SVM).

Unsupervised learning (UL), on the contrary, operates on unlabeled data, in which no information about the correct answer is present. «In this learning approach, an artificial intelligence system gets just the input data and not the associated output data. Unsupervised machine learning, unlike supervised learning, does not need the presence of a person to oversee the model» (Naeem, Samreen & Ali, Aqib & Anam, Sania & Ahmed, Munawar, 2023). The algorithm must infer autonomously, without any supervision – neither external, nor provided by the data themselves – the hidden relations in the data. The objectives of unsupervised learning are multiple: «discover groups of similar examples within the data, where it is called clustering, or to determine the distribution of data within the input space, known as density estimation, or to project the data from a high-dimensional space down to two or three dimensions for the purpose of visualization» (Bishop, 2006).

Finally, reinforcement learning (RL) provides for the maximization of a cumulative reward. In other words, an agent (i.e. the model itself) operates in an environment receiving feedback in the form of positive reinforcements (rewards) or negative ones (punishments). «Here the learning algorithm is not given examples of optimal outputs, in contrast to supervised

learning, but must instead discover them by a process of trial and error» (Bishop, 2006). As can be seen in Sarker (2021), the problems addressed with RL are defined as Markov Decision Process (MDP), namely everything that can refer to «sequentially making decisions» (Sarker, 2021). An RL problem typically includes four elements: Agent, Environment, Rewards, and Policy (or optimal decision strategy), as expressed in Sarker (2021). RL techniques are divided into *model based* and *model-free*. Their fundamental difference does not lie in the structure of the decision strategy, but in the integration of a model of the dynamics of the environment (transitions and rewards). While model-based methods construct and use such a model for planning, model-free methods estimate the policy directly from experience, without explicitly modelling the environment (Sutton & Barto, 2018).

In summary, supervised learning uses labeled data to learn to predict a correct response, unsupervised learning uses unlabeled data to discover hidden structures, and reinforcement learning uses interaction with an environment to learn to maximize a reward (Buzzi, Priori, 2025).

An intermediate level of learning, which will be important when we speak about the learning of Large Language Models, is called Self Supervised Learning (SSL). The underlying, and together revolutionary, intuition here is that we can just use running text as implicitly supervised training data for a model. For example, a word [c] that occurs near the target word [apricot] acts as gold correct answer to the question "Is word [c] likely to show self-supervision up near [apricot]?" This method avoids the need for any sort of hand-labeled supervision signal (Jurafsky & Martin, 2026).

1.3.4. Connectionism and Neural Networks

Within the vast panorama of *Machine Learning*, *neural networks* have developed. The drive toward this new approach was motivated by a partial return to the unity between intelligence and the support of intelligence. In other words, if before (cf. Turing, 1937, 1950) one had tried to develop models of intelligence that explicitly tried to elude their relation to the support, whether biological or artificial, now one returns to that point. The idea that guides this reversal is the one according to which human intelligence may not reside totally in logic, but in the structure of the support that realizes it (the brain): it is a "bottom-up" approach, from the complex network of neuronal connections to intelligence.

The development of neural networks is born with the idea of realizing artificial architectures that can help study human intelligence by attempting to simulate it. As Kaplan argues, the interesting thing is that we know a great deal about the structure of the brain, down to the details, which layers and which regions are usually involved in various activities. Surprisingly, however, very little has been understood about how the brain is wired (metaphorically speaking), and this is precisely the area of interest of AI researchers who try to build neural networks (Kaplan, 2018). It is necessary, however, to distinguish between neural networks as an empirical (or engineering) methodology, aimed at solving practical problems, and neural networks as a theoretical explanatory and simulative apparatus of mental processes, which has taken the name of *connectionism*.

A step back. The development of the connectionist approach (and no longer only symbolic) spreads from the 1980s and 1990s of the twentieth century, but its original trace goes back to the realization of the first formal neuron in 1943 by McCulloch and Pitts, whom we have already cited when speaking of the Macy Conferences. They theorized that the discrete action (i.e. binary, digital logic) of nervous components was the only way in which the nervous system could manage the immense quantity of information that crosses it. To arrive at this formalization, the authors consider some basic assumptions of theoretical neuro-physiology (cf. McCulloch & Pitts, 1943/1990):

> assumption 1. The "all-or-none" law. A neural impulse between synapses occurs only above a certain threshold. Therefore, either the neuron "fires" (produces an impulse) or remains at rest.
>
> assumption 2. The impulse depends only on the time and place in which this threshold is exceeded, and not on any other characteristic.
>
> assumption 3. The threshold for the impulse corresponds to a fixed number of synapses that must be excited simultaneously (within a short period of "latent addition") for the neuron to activate.
>
> assumption 4. The threshold for the impulse also depends on the presence of inhibitory synapses.

Consequently:

consequence 1. Since a neuron either "fires" or remains at rest, every neural event can be represented mathematically as a discrete event, consequently corresponding to a simple logical proposition (cf. Franks, 2024).

consequence 2. It is possible to designate the neurons of a given network N with $c_1, c_2, \ldots$ The "action" of c_i is indicated by the term $N_i(t)$, that is, the activation of a neuron at a given moment. The time t corresponds to the number of synaptic delays from the beginning of the process. To describe this, the *functor S* is introduced, which indicates a backward shift in the discrete time of the network. If $P(t)$ means that a certain property P holds at time t, then $S(P)(t)$ means that that same property held at the immediately preceding time. Generalizing, $S^n(P)(t)$, with n the number of steps.

consequence 3. The fact that a neuron fires only if it receives impulses from several neighbors simultaneously (spatial summation) is equivalent to a logical conjunction (AND) or to a disjunction (OR), depending on the excitation threshold.

consequence 4. The presence of inhibitory synapses, which prevent the neuron from firing, corresponds mathematically to logical negation (NOT).

Consequences 3 and 4 strictly depend on the first, which represents the fundamental epistemological leap in McCulloch and Pitts.

Continuing, the authors encounter a problem with regard to learning. This is understood as a process in which concomitant activities at a previous time have permanently altered the network, so that a stimulus that previously would have been inadequate is now adequate (McCulloch, Pitts, 1943/1990). The question turns out to be that of modelling learning as a historical-biological process. To integrate this plastic dimension of the neural network, the authors implement "nets with circles", that is, with recursive structures (McCulloch, Pitts, 1943/1990).

A cyclic net (or circle) is defined as a chain of neurons that returns to the starting point. The minimal set (or base) of neurons that maintain the cyclic net is called cyclic set, $c_1, \ldots, c_p$ with p defining the cardinality of the set and therefore the order of the network. After some normalization steps, one arrives at the form:

$$N_i(z_1) = Pr_i\left[S^n N_1(z_1), \ldots, S^n N_p(z_1)\right] \qquad (2)$$

where Pr_i is the propositional condition that determines the firing of the cyclic neuron i. It is the logical law that says when neuron i fires, as a function of the other neurons of the cycle.

Finally, the configuration of the system is defined as $X_i(z_1)$, namely the overall state of the cyclic set at a certain instant of time z_1. If the cyclic set is composed of two neurons, the possible configurations are:

$$X_1(z_1) = N_1(z_1) \wedge N_2(z_1)$$

$$X_2(z_1) = N_1(z_1) \wedge \neg N_2(z_1)$$

$$X_3(z_1) = \neg N_1(z_1) \wedge N_2(z_1)$$

$$X_4(z_1) = \neg N_1(z_1) \wedge \neg N_2(z_1) \tag{3}$$

where:

- $N_j(z_1)$ represents neuron j firing at time z_1;
- $\neg N_j(z_1)$ represents neuron j not firing at time z_1.

In general, given p neurons, the total number of possible configurations is 2^p. Therefore, it holds that:

$$X_i(z_1). \equiv . \sum_{j=1}^{2^p} Pr_{ij}(z_1) \, . S^n X_j(z_1) \tag{4}$$

which regulates the evolution of the configurations of the cyclic network and says that the global state X_i at time z_1 holds if at least one of the terms $Pr_{ij}(z_1).S^n X_j(z_1)$ holds.

The expression found, with logic formalism, demonstrates the capacity to evaluate the configuration of the system at a given instant of time, knowing the configuration of the system in the preceding instant of time. This causality has appeared in different forms in different sciences, but never, except in statistics, has it been so irreciprocal [asymmetrical] as in this theory (McCulloch, Pitts, 1943/1990). In other words, this model allows the description of a state starting from the preceding one but prevents the reverse because (4) includes disjunctive relations. For this reason, «we cannot deduce […] [the] past from present activities» (McCulloch, Pitts, 1943/1990).

The authors conclude that formal equivalence (the capacity of the mathematical model to replicate the logical behavior of the neuron) must not be exchanged for a factual explanation (an exhaustive description of biological reality). This means that the theory privileges the logical description (i.e. propositional, binary)[8] of behavior, but not the concrete process through which the network changes. Real learning remains "enduring change" (i.e. "lasting change") of the network, that is, it opens the problem to the modelling of dynamic memory.

In conclusion, the model of McCulloch and Pitts – although presenting evident limits – already contained in 1943 the fundamental idea that allowed the development of connectionism, namely the idea that intelligent properties can emerge from the organization of networks of simple interconnected units, and that such organization can be formalized as a computational system. If the model of McCulloch and Pitts incorporates learning only formally as an enduring change of the network, without yet assuming it as the central object of modelling, in connectionism the network becomes precisely the place of learning.

The perceptron. In fact, from what emerges in *The Master Algorithm* by Pedro Domingos, it can be said that the neuron of McCulloch and Pitts does not know how to learn, but to give it the possibility, we must associate variable weights with the connections between neurons (Domingos, 2020). The result of this operation is Frank Rosenblatt's *perceptron* (1958). In this model, excitatory and inhibitory synapses are identified by positive and negative weights respectively, and it is possible to vary the weights and thresholds to modify the activation function of the perceptron (Domingos, 2020), whose output is 1 or 0 depending on whether the weighted sum of the inputs exceeds the threshold. Unlike the McCulloch-Pitts neuron, where the binary output was the direct result of a logical proposition, in the perceptron it is the result of a threshold function applied to a continuous weighted sum of inputs. It is precisely this continuity of the weighted sum (even if then reduced to a binary output) that makes possible the geometric description of the decision boundary as a hyperplane in the space of inputs, as we will see shortly.

What does the learning of weights by the perceptron consist of? In a network with n inputs (the data), these are represented by an n-dimensional vector. Each input has its weight $\omega_1, \dots, \omega_n$. Geometrically, the decision boundary with which to determine whether a given example is positive or negative is the locus of the points of the hyperplane whose weighted

[8] It is not by chance that McCulloch and Pitts are contemporaries of Turing, that is, they enter into a reflection that revolves entirely around the problem of computability.

sum is equal to the threshold. If we were in two dimensions, the hyperplane would be a line and the learning process of a perceptron would consist in varying the direction of the line (that is, varying the weights) so that it separates the positive examples from the negative ones (Domingos, 2020).

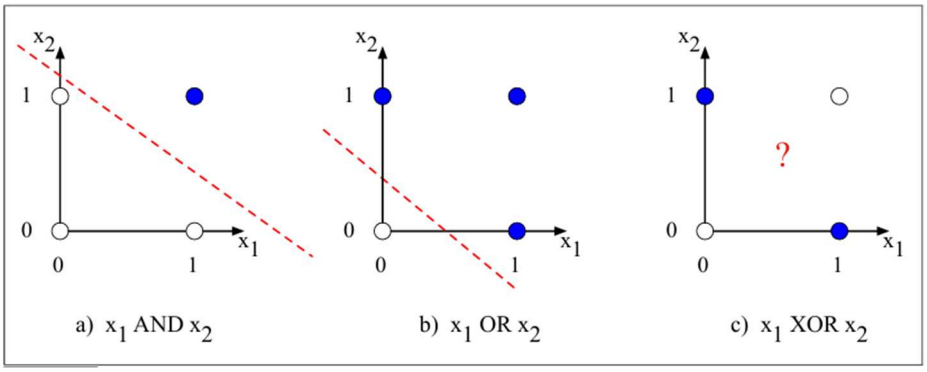


Figure 7: The AND, OR and XOR functions, represented with input x_1 on the x-axis and the input x_2 on the y-axis. Filled circles represent perceptron outputs of 1, and white circles perceptron outputs of 0. There is no way to draw a line that correctly separates the two categories for XOR. After Jurafsky and Martin (2026).

The geometrical consideration made leads to the evidence that the perceptron is able to trace a decision boundary that is linear. For this reason the problems solvable in this model are only those in which the classes of data are linearly separable, and the model fails whenever it is faced with a problem that is not linearly separable (i.e. the problem of the logical function of "exclusive OR" or XOR; Fig. 7), a problem demonstrated by Minsky and Papert in their *Perceptrons: An introduction to computational geometry* (1969). These limits and together the capacity to have rethought learning as variation of the weights of the network proved fundamental for the consolidation of the concept of neural network and then for the passage to *Deep Learning*.

1.3.5. Deep Learning

The limit of the perceptron is solved by the concept of *neural network* properly used, namely the addition of *layers* to the single *neural unit*. «The core of the neural network is the hidden layer h formed of hidden units h_i, each of which is a neural unit […], taking a weighted sum of its inputs and then applying a non-linearity» (Jurafsky & Martin, 2026). Together, the use of *layers* of intermediate neurons (*hidden* in this sense), which mediate between input data and output neurons, and the addition of a non-linear activation function allow the network thus constructed to solve non-linearly separable problems. The non-linear activation function or, directly, *activation function*, replaces the threshold in the architecture of the perceptron.

Following the review by Jurafsky and Martin (2026), neural units, once the weighted (with the weights w_i) and linear sum of the inputs (x_i) has been computed, represented through scalar product with

$$z = \boldsymbol{w} \cdot \boldsymbol{x} + b \tag{5}$$

where b is the bias, apply a non-linear function

$$y = f(z) \tag{6}$$

as can be seen in Fig. 8.

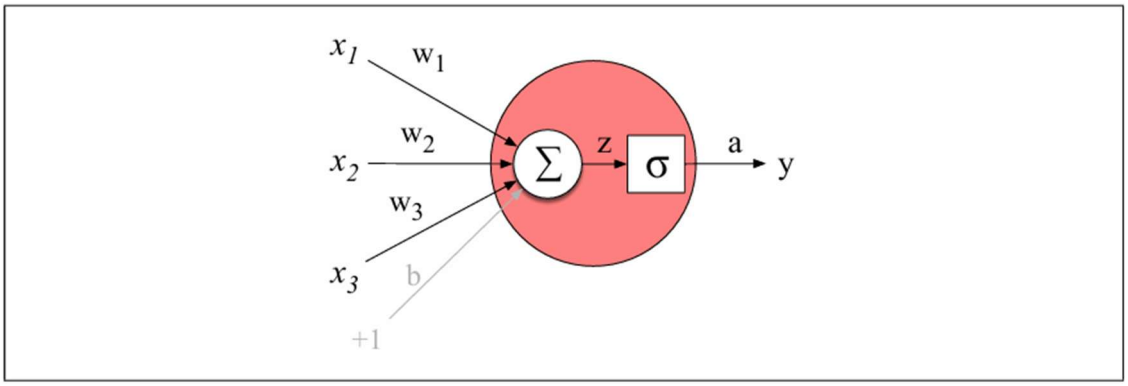


Figure 8: Schematization of the function of the neural unit or *perceptron.* After Jurafsky and Martin (2026).

The activation functions most used in modern neural networks are *sigmoid*, *ReLU* and *tanh*.

This double innovation (*hidden layer* and *activation functions*) is strategic. The former do not limit themselves to mediating and transmitting passively from input to output, but provide an intermediate representation of the input data, in a new space, in which the subsequent layer can work better. In other words, they carry out a remapping of the space of the data. In the XOR case cited before, the two positive points [1;0] and [0;1] are made to collapse into the same hidden representation $h = [1,0]$, for which it becomes easy to trace the decision boundary (Fig. 9).

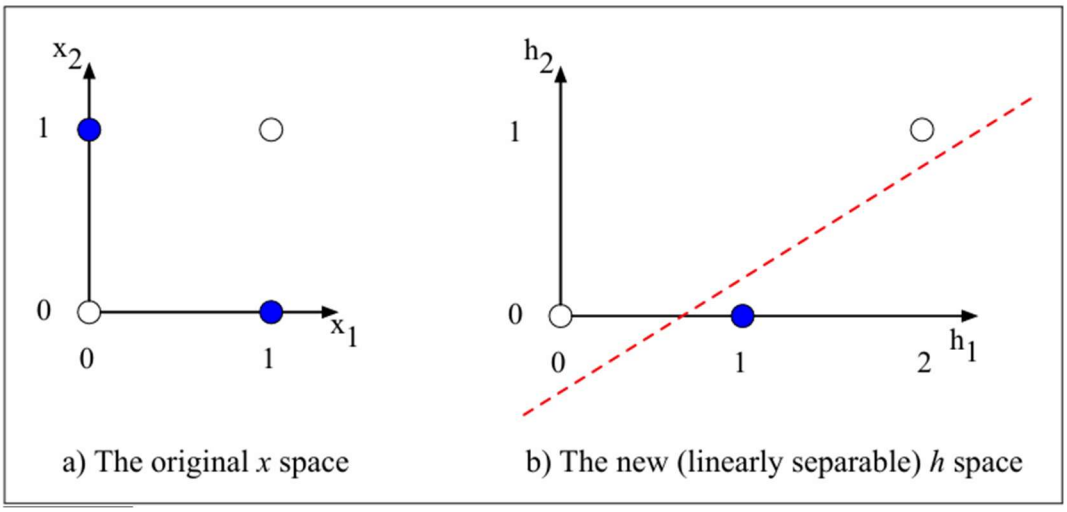


Figure 9: The hidden layer forming a new representation of the input. (b) shows the representation of the hidden layer, **h**, compared to the original input representation **x** in (a). Notice that the input point [0, 1] has been collapsed with the input point [1, 0], making it possible to linearly separate the positive and negative cases of XOR. After Jurafsky and Martin (2026).

In LeCun et al. (2015) the same idea is formulated in general terms: «hidden layers can be seen as distorting the input in a non-linear way so that categories become linearly separable by the last layer» (LeCun et al., 2015). Activation functions, instead, allow the layers to be functional: «Without activation functions, a neural network would simply perform linear transformations, limiting its ability to model complex relationships in the data» (Mienye & Swart, 2024). In summary, the depth of the network does not function without non-linearity, and vice versa.

Deep Learning (DL) represents the implementation of these concepts, that is, it is realized with deep networks in which a vast number of *hidden layers* is employed. DL «allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction» (LeCun et al., 2015). The more layers, the more the intermediate representations will come to abstract increasingly complex characteristics from the input data. However, the increase in the number of layers involved in the "deep network" also entails the increase in the number of parameters (i.e. the typical weights and biases of all neural networks, as we said before) to be fixed during training. In fact, the paradigm of Deep Learning becomes historically relevant as soon as the problem of representation intertwines with the problem of optimization. Jurafsky et al. (2026) recall that the revival of neural networks occurs in the 1980s, after a moment in which they had been set aside, thanks to the diffusion of practical instruments for constructing deeper networks, such as the backpropagation algorithm (follows). In the early 2000s, thanks to better

hardware and optimization techniques (Bengio et al., 2007; Hinton et al., 2006), it becomes possible to train larger and deeper networks, beginning Deep Learning in the modern sense.

The way in which neural networks adjust parameters occurs automatically (Jurafsky & Martin, 2026), and this represents the learning mechanism, something that the model of McCulloch and Pitts could not implement.

In general, the learning of a multi-layer architecture concentrates on three essential elements, *forward pass*, *loss function* and *gradient descent*, which follow one another in an iterative process. The first concept enucleates the fact that learning starts from a computation "forward", that is, from the input data $\boldsymbol{x}$ to the estimate $\hat{y}$,[9] calculated with (6), of the true value of the output y (the target)[10]. The process proceeds by correcting the parameters that generated this estimate. The quantification of the distance between $\hat{y}$ and y, that is, the error of the estimate, occurs through a *loss function* $\mathcal{L}$. Generally, it is common to use *cross-entropy loss* as *loss function* (Jurafsky & Martin, 2026). Now that one has an error associated with the output generated by the network, training the network means finding the parameters that minimize the value obtained from the *loss function*. The optimization algorithm used to accomplish this passage is the *gradient descent algorithm*, in its *stochastic* version (*Stochastic Gradient Descent* or SGD). In the first case the algorithm calculates the gradient of the parameters using the entire training dataset for each single update, an operation that can prove extremely slow and costly on large datasets. The stochastic variant, instead, uses small random subsets of data (from which it takes the name stochastic), called *mini-batch*, to obtain an estimate of the gradient; this approach makes it possible to execute more frequent updates and guarantees that the calculation time for each step does not grow as the total dimension of the dataset increases (Goodfellow et al., 2016). Synthetically, the updating of the parameters occurs according to the rule:

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta \mathcal{L}(\theta) \tag{7}$$

Within θ are contained the parameters (weights and biases) of the model. The gradient $\nabla_\theta \mathcal{L}(\theta)$ is the vector of the partial derivatives of the loss with respect to the parameters. The SGD algorithm updates θ by following the direction opposite to the gradient of the cost

[9] In this case we are specifically in a fully connected, feedforward network: every unit of layer i is connected to every unit of layer i+1, without cycles. The same does not hold, for example, for CNNs.

[10] y is given as target because we are in the specific case of SL. Other families of DL also implement UL without target.

function, that is, by following the direction of "descent". Finally, η is called learning rate, a hyper-parameter that controls the amplitude of the step with which the parameters are updated at every iteration.

Calculating the gradient of the loss is immediate when one has networks with few layers. In the case of DNNs (i.e. *deep neural network*) this is no longer true and, in particular, «it's much harder to see how to compute the partial derivative of some weight in layer 1 when the loss is attached to some much later layer» (Jurafsky & Martin, 2026). The solution to this problem is represented by the already cited *backpropagation algorithm* or *backprop*. Jurafsky traces the invention of backprop back to Rumelhart, Hinton and Williams (1986). In LeCun, Bengio and Hinton (2015) a broader genealogy is given. The idea of training multilayer architectures with gradients and backprop «was discovered independently by several different groups during the 1970s and 1980s», namely Werbos (1974), Parker (1985), LeCun (1985). Concretely, backprop is nothing other than the application of the *chain rule* typical of the mechanisms of derivation of composite functions. In this case, the entire process proves to be «the same as a more general procedure called backward differentiation, which depends on the notion of computation graphs» (Jurafsky & Martin, 2026), where «a computation graph is a representation of the process of computing a mathematical expression, in which the computation is broken down into separate operations, each of which is modeled as a node in a graph» (Jurafsky & Martin, 2026). An example of this is reported in Fig. 10, where $L = f(e)$ and $e = f(d)$.

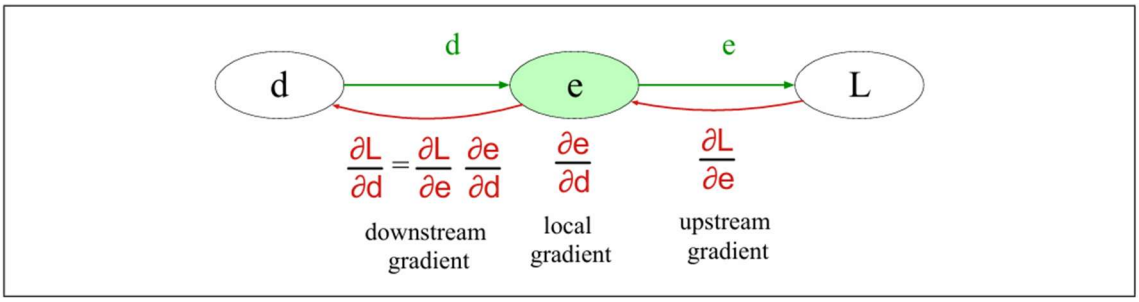


Figure 10: Each node (like *e* here) takes an upstream gradient, multiplies it by the local gradient (the gradient of its output with respect to its input), and uses the chain rule to compute a downstream gradient to be passed on to a prior node. A node may have multiple local gradients if it has multiple inputs. After Jurafsky and Martin (2026).

Backpropagation is applied repeatedly through all the modules of the graph, starting from the output (where the network produces its prediction) to the first layer (where the external input is provided). In doing so, it becomes simple to calculate the gradients of the cost function (loss function) with respect to the weights of each module (LeCun et al., 2015).

The only condition for applying this algorithm, a condition that also clarifies the non-tautology of the approach, is that the modules of the network be «relatively smooth functions of their inputs and of their internal weights» (LeCun et al., 2015), that is, that the modules are effectively differentiable or quasi-differentiable. The local differentiability of the modules together with the existence of the chain rule makes possible the use of backprop.

The examples taken above are particular and specific. Having taken the true value y as target for the network in the calculation of the loss function, since the target is known for each input (Jurafsky & Martin, 2026), shows us that we are in the case in which the training of the network is part of the supervised learning, but the different families of deep neural network also implement the other approaches to learning in ML (unsupervised learning and reinforcement learning). The main families of DNNs, diversified by type of architecture and purpose, are:

1. FNN (*Feedforward Neural Network*): networks in which information flows exclusively in one direction, from the input toward the output, without cycles. Within this family different structures are distinguished on the basis of the density and topology of the connections. We have:
    a. *Multi-Layer Perceptron* (MLP): they are feedforward fully-connected networks. This last notion specifies that each unit of layer i is connected to each unit of layer $i + 1$, without cycles.
    b. CNN (*Convolutional Neural Network* or ConvNets): also, without cycles, networks most used to process data that present themselves in the form of multiple arrays, such as images. The typical architecture of a ConvNet provides a succession of layers that extract characteristics in a hierarchical way (Tian et al., 2025), called *convolutional layer* and *pooling layer*. The former have the task of «detect local con junctions of features from the previous layer» (LeCun et al., 2015), the latter intervene to merge semantically similar characteristics into one, creating a more synthetic and abstract image. This is made possible because the fundamental idea of CNNs is that images present local statistics «invariant with respect to position» (LeCun et al., 2015), which makes it possible to consider the characteristics as shared within the array through the use of shared weights (Tian et al., 2025).

2. RNN (*Recurrent Neural Network*): they are designed for tasks involving sequential inputs, namely data in which each element is intrinsically correlated with those that precede or follow it. In other words, in a sequence the meaning of an element often depends on its temporal or positional context. For this reason this type of network is mostly used for language processing. At the architectural level, these networks maintain «in their hidden units a "state vector" that implicitly contains information about the history of all the past elements of the sequence» (LeCun et al., 2015), an element that allows the network to be given a temporal dimension in its computational structure.

RNNs, in their evolution, have been recognized as useful in predicting the next character in a text or the next word in a sequence, or also in translating texts from one language to another. Indeed, in addition to the use of state vectors, subsequent developments (Sutskever et al., 2014) allowed the integration of the encoder-decoder paradigm within the architecture of RNNs, which will be peculiar for the development of the *transformer* architecture.

For what has been said so far, the developments of the Deep Learning paradigm have shifted attention from simple representation to learning, specifically to learning from data. From the perceptron to deep neural networks one does not pass through a simple succession of models, but through a progressive deepening of the same paradigm. First the possibility of representing non-linear functions through hidden layers, then the possibility of training increasingly deep architectures through backpropagation, then the specialization of these architectures with respect to different types of data.

Chapter 2

In the state of the art panorama, the models and tools developed are much more numerous than those taken here as the main axis of analysis, and belong to different functional families which, although not central to the present discussion, have played a decisive role in the consolidation of the entire ecosystem and in the extension of the areas of application of AI. This line of research now intends to focus on the analysis and study of *Transformer* architecture, necessary for the discussion around the concept of *Large Language Models*.

2.1. LLMs and Transformer architecture

The definition of *Large Language Model* is conceptually simple. It is «a *neural network,* trained to predict the next *token,* in a very large corpus of text» (Lewis, 2025). This type of network is involved in the field of *Natural Language Processing*. Specifically, having a given context as input, it outputs a distribution over possible next token or word (Jurafsky & Martin, 2026). It is straightforward that the definition provided make LLMs the ideal culmination of the evolutionary path we have described so far, that leads from symbolic modelling of language to forms of statistical learning from data. Having already depicted the concept of neural network, the peculiar LLMs' features will be clarified in what follows.

Transformers. The heart of LLMs is the *Transformer* architecture, introduced by Vaswani et al. with the 2017 article "Attention Is All You Need". If recurrent neural networks were considered the state of the art in sequential data processing (i.e. step-by-step method), that was also their limitation, so that transformers imposed themselves as a turning point as they are able to analyze and process entire input sequences in parallel, drastically improving computational efficiency. Transformers, therefore, eliminate recursion in favor of parallelism, relying on one and only mechanism called *attention* (Vaswani et al., 2017). Specifically, these systems are composed of two main blocks consisting of N layers each (normally N=6)[11], called *encoders* and *decoders* respectively, a basic structure inherited from seq2seq encoder-decoder models based on RNN/LSTM[12] (Sutskever et al., 2014), as it is shown in Fig. 11.

[11] We are referring to the original transformer architecture of the 2017 paper by Vaswani et al., i.e. the one designed for sequence transduction tasks such as automatic translation.

[12] LSTM means *Long Short-Term Memory* (cf., Hochreiter & Schmidhuber, 1997)

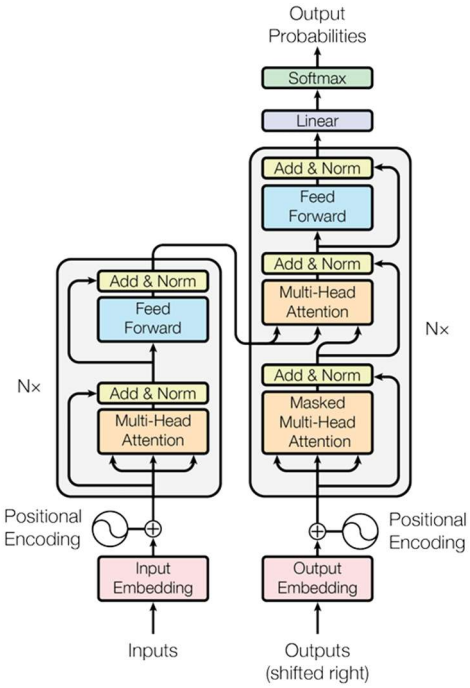


Figure 11: The Transformer – model architecture (Vaswani et al., 2017).

From tokens to vectors

Since the purpose of Transformers is to process language, the first problem that arises is the fact that neural networks operate on vectors, while natural language comes in the form of sequences of discrete symbols. The network, therefore, first of all needs to use the *embedding mechanism*.

Tokenization and Embedding. The system is not able to read the input text, so it proceeds to segment the text into discrete units called tokens. The token may not match with an entire word of the given sentence, but also with a part of it, depending on the characteristics of the *tokenizer* in use. Since the system operates on numbers, each token is replaced with an integer numerical index taken from a fixed vocabulary. At this point the input text has become a sequence of integers, but this process represents a simple symbolic substitution. This sequence of numbers does not give any information about the meaning of the word or associated token. To take into account the semantic dimension (i.e. the meaning) of the tokens, the *embedding matrix* is introduced, that is

$$X_{emb} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^{N \times d_{model}}$$

where N is the number of tokens in the vocabulary and d_{model} the dimension of the model, in which each row represents a vector associated with a token of the previous sequence. In this way, it is possible to link each token with a position in a continuous vector space. It is reasonable to think that semantically similar words or tokens also reveal a geometric proximity in the given vector space, idea canonically attributed to Bengio et al. (2003) and Mikolov et al. (2013). An example could be the words "Norway" and "Sweden", or even the words "Tuesday" or "Wednesday" (LeCun et al., 2015). However, as Lewis (2025) has pointed out, vectorized tokens within the embedding matrix are "static" vectors that represent lexical meanings that are distributed but still lack specific context. In other words, the deep meaning of tokens (and their embeddings) emerges only from the relationships with the other tokens within the sentence, subsequently captured by the model in the *contextualization phase* that takes place in the encoder. It is for this reason that Lewis (2025) contrasts the rigidity of the old symbolic structures with the dynamic nature of the vectors in a continuous space which, despite being semantically static, are functional to the predictive capacity of the Transformer, i.e. they are the mathematical foundations on which the system rests.

Positional encoding. Before entering the encoder, the embedding matrix retrieves the structural value of the position of the tokens within the sentence, originally managed by the recursivity of the previous models. Then, each element x_i of X_{emb} is added with its position p_i, suitably calculated (cf. Vaswani et al., 2017) and the matrix obtained is

$$X_0 = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix} \in \mathbb{R}^{N \times d_{model}}$$

where $z_i = x_i + p_i$.

It is worth noting that the positional encoding described above represents the original formulation proposed for the standard Transformer architecture, by Vaswani et al. (2017), in which positional information is added through a fixed sinusoidal signal to the token embeddings prior to the attention computation. However, subsequent research has shown that this approach presents limitations in terms of generalization to sequence lengths unseen during training and of sensitivity to absolute rather than relative token distances. For this reason, modern decoder-only (see section 2.2) large language models have largely moved towards relative or rotary positional encoding schemes. Two approaches have become particularly prominent in this line of research. *Rotary Position Embedding* (RoPE),

introduced by Su et al. (2021), encodes positional information by applying a rotation to the token representations, so that the resulting proximity and similarity depends exclusively on the relative distance between token positions rather than on their absolute indices. *Attention with Linear Biases* (ALiBi), proposed by Press et al. (2021), follows a different strategy, introducing no positional signal into the embeddings and instead adding a fixed negative linear bias as a function of the distance between tokens, thereby encouraging the model to favour closer contextual dependencies. Both approaches have demonstrated stronger extrapolation capabilities beyond the training context length compared to the original sinusoidal formulation and have been widely adopted in the architectures underlying contemporary large language models.

The matrix X_0, output of the embedding and the positional encoding mechanisms, now become the encoder input. Before entering in its internal functioning, we should describe the general attention mechanism introduced by Vaswani et al. (2017).

Attention Mechanism

As we said, the major outcome of Vaswani et al. work is to have developed «the first transduction model relying entirely on self-attention to compute representations of its input and output without using sequence aligned RNNs or convolution» (Vaswani et al., 2017). At the most general level, attention is defined as a function that maps a set of *queries* and a set of *key-value* pairs to an output. More precisely, the tripartite structure – i.e. *queries, key, values* – is introduced by the fact that a word in a given sentence has different roles based on the context and the way in which it is interrogated. In particular:

1. **Query (Q):** Represents what the token is "looking for" in the rest of the context.
2. **Key (K):** Represents the label or information that the token offers to others.
3. **Value (V):** Contains the actual information content of the token that will be extracted if the Query and Key match

In other words, the key can be read as the profile (or *sign*, to make an analogical-operational comparison with the *theory of meaning*) with which a token makes itself recognizable, the value as the operational content the token provides once selected, and the query as a contextual interpretative instance, i.e. as the local question that a token asks the context to establish which other token is relevant in that precise sequential configuration.

Formally, given an input representation matrix X the model computes

$$Q = XW^Q, \qquad K = XW^K, \qquad V = XW^V$$

where Q, K and V represent the query, key, and value matrices and W^Q, W^K and W^V the matrices of *learned weights*. Thus, input tokens are projected into three different vector spaces through matrices. Since two vectors that are more similar to each other have a higher dot product with each other, the dot product acts as a similarity function (Jurafsky & Martin, 2026). For this reason, attention is computed as

$$Attention\ (Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where the relevance or pertinence of the tokens with respect to the others is calculated by first comparing the *similarity* (i.e. scaled dot product) between query and key (i.e. the *softmax* arguments), while the content is transferred only later, through the weighted combination of values (Lewis, 2025). Comparing only Q and K first, allows us to isolate the structure of the relationships of the context, before applying it to the information content (V).

Furthermore, the softmax function is the mathematical step that transforms similarity scores into probabilistic weights. First, it takes the result of QK^T and scales it by the factor $1/\sqrt{d_k}$, where d_k represent the dimension of query and keys vectors. The reason for performing this scaling is motivated by the fact that, for great values of d_k, the dot product increases, pushing $softmax$ into regions of quasi-null gradients, making it very difficult for the model to update the weights and learn from the data. Second, the function maps scaled dot product results into a suitably normalized probability distribution (Jurafsky & Martin, 2026). This mechanism allows us to determine the weight assigned to each value with respect to the current query.

At this stage, the formulation is still fully general as it does not yet specify whether Q, K, and V are derived from the same sequence or from different sequences of input. This distinction will become crucial when moving from the general definition of attention to its specific realizations in encoder self-attention and decoder cross-attention.

Multi-head attention

While scaled dot-product attention defines the general mathematical form of the attention mechanism, the Transformer implements this mechanism through *multi-head attention,* that enable the model to capture diverse relationships and patterns. Instead of using a single set of Q, K, V matrices, the input embeddings are projected into multiple sets (i.e. heads, represented by h layers in Fig. 12), each with its own Q, K, V.

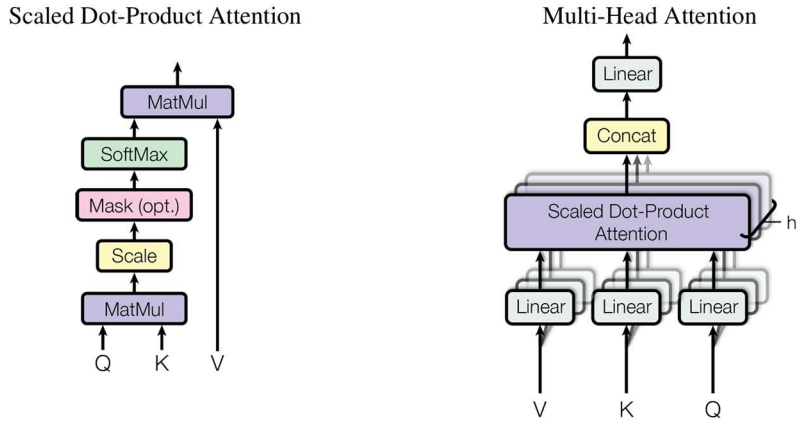


Figure 12: (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of h attention layers running in parallel (Vaswani et al., 2017).

Mathematically, multi-head attention is expressed as:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_h)W^O$$

where:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

where W_i^Q, W_i^K, and W_i^V are learned projection matrices specific to the i-th head. The typical dimension for each head is $d_k = d_v = \frac{d_{model}}{h}$. W^O is a final weight matrix to project the concatenated output back into the model's required dimensions, to produce the final contextualized representation.

As we said before, this architecture has drastically reduced computational costs, also related to training times. This model has set new records in terms of precision with respect to automatic translation tasks and also for constituency parsing. For example, the system excels

at resolving distant relationships in the text (i.e. long-distance dependencies), such as the resolution of anaphoras, i.e. linking a pronoun to its antecedent that appeared much earlier (Vaswani et al., 2017; Lewis, 2025). As a comparison, while in RNN the number of operations needed to relate two distant signals grows linearly with their distance, in Transformers this complexity is reduced to a constant number of operations, regardless of distance. For all these reasons, Transformer has represented an important turning point in AI research field.

To summarize, multi-head attention is the architecture the system uses to simultaneously capture different kinds of dependencies across different representational subspaces and positions in the sequence of tokens. Thus, it is used for all the three types of attention mechanism we'll see: *encoder self-attention*, *decoder self-attention* and, lastly, *encoder-decoder attention* (or *cross attention*).

Encoder

The encoder, the first major block of the Transformer, aims to "understand" the sentence provided as input. As we said before, the focus lies in the fact that tokens (or words) do not have meaning in themselves, but in relation to the other parts of the sentences, following the argument of associationism (Lewis, 2025). To do this, each of the encoder's N layers is composed of two sub-layers: a *multi-head self-attention* layer and a *position-wise feed-forward network* layer (cf. Fig. 11). The first provides the system with the possibility of enriching the semantic content of a word or token by collecting information from all the other words in the sequence, i.e. it provides the contextualization mechanism; the second layer operates – through complex non-linear operations – a refinement of the previously enriched vector.

Encoder self-attention. To operate the contextualization mechanism, i.e. progressively reconstructs tokens as context-dependent units, the encoder needs to use the attention mechanism. Let $X^{(l-1)}$ be the input representation to the l-th encoder layer, with $l \in \{1, 2, \dots, N\}$ and N is the number of the encoder's layers. For $l = 1$ we have $X^{(0)} = X_0$ that is the output of embedding and positional encoding mechanism. Thus, in all subsequent layers, $X^{(l-1)}$ is the contextualized output produced by the previous encoder layer. The attention mechanism computes queries, keys, and values from this same matrix as:

$$Q^{(l)} = X^{(l-1)}W^{Q,(l)}, \qquad K^{(l)} = X^{(l-1)}W^{K,(l)}, \qquad V^{(l)} = X^{(l-1)}W^{V,(l)}$$

Since Q, K, and V are all derived from the same input matrix through distinct linear projections, the mechanism relates each token to the other tokens of the same sequence. This is precisely the condition that defines the process as *self-attention*.

In the multi-head implementation we have this operation repeated in parallel across h different heads. For each head i, the layer computes:

$$\text{head}_i^{(l)} = \text{Attention}\left(X^{(l-1)}W_i^{Q,(l)}, X^{(l-1)}W_i^{K,(l)}, X^{(l-1)}W_i^{V,(l)}\right)$$

and then combines the outputs as

$$\text{MultiHead}^{(l)} = \text{Concat}\left(\text{head}_1^{(l)}, \dots, \text{head}_h^{(l)}\right)W_0^{(l)}.$$

Encoder output. After N layers, the encoder returns a final matrix of contextualized token representations. This final output no longer corresponds to a simple vectorization of the input tokens, as it was for the embedding process, but it represents a distributed and context-sensitive representation of the whole sequence, which will then be used by the decoder in the following stages of the Transformer architecture.

Self-attention and Parallelization

From the discussion made up to now, it is straightforward that the N encoder (and later also the N decoder layers) work in series along the depth of the network, because each process takes the output of the previous layer as input. At first glance, this structure could be in contradiction with the parallelization capacity of the Transformer, a central novelty of this architecture. In reality, parallelization is not about how layers communicate, but about how tokens are treated within each layer. Recalling that, unlike recurrent architectures – in which tokens must be processed one after the other – in Transformers the entire sequence simultaneously enters the calculation of the matrices Q, K and V, and the attention matrix QK^{T} relates all the positions with all the others in a single operation. In this sense, the Transformer retains a layered serial structure in depth, but realizes a parallel processing on the token axis, through the structure of multi-head attention.

Decoder

If the encoder is responsible for transforming the input token sequence into a continuous, contextualized vector representation, the decoder – on the other hand – is responsible for generating the output sequence. We have already mentioned that the purpose of the Transformer is to predict the next token given a sequence; the decoder plays precisely this role. Its structure includes N layers each composed of a *masked multi-head self-attention* (i.e. a modified version of the encoder's *multi-head self-attention*), an *encoder-decoder attention* (or *cross-attention*) and, as for the encoder, a position-wise feed-forward network.

Causality or Autoregression. At each iteration, the decoder provides an output probability distribution of the possible subsequent words in the input sequence. After appropriate sampling (cf. Jurafsky & Martin, 2026), one word is chosen from those possible. In other words, decoders are generative models, meaning that, given input tokens, they generate novel output tokens (Jurafsky & Martin, 2026). The generated word – the new token – is added to the input sequence, and the now updated sequence re-enters as decoder input in a loop process (i.e. feedback loop). Since the system generates output tokens iteratively – one at a time – and each newly updated sequence is fed back as input to the decoder, the information flow in decoders goes left-to-right, meaning that the model predicts the next word only from the prior words (Jurafsky & Martin, 2026). Decoding from a language model in a left-to-right manner, and thus repeatedly choosing the next token conditioned on our previous choices is called causal or autoregressive generation. (Jurafsky & Martin, 2026).

Embedding and positional encoding for decoder. Each time the model generates the next token, since it is chosen within a probability distribution, the output is a numerical index, not a vector. Thus, the generation loop requires re-embedding and, also, positional encoding, as it was for the encoder.

How is causality preserved? Since we want to address this question, first we need to distinguish two different level of discussion: training and inference times (i.e. “real” use). The latter is the one described above, genuinely iterative and left-to-right generation, the fact that new prediction is conditioned only on the tokens that have already been generated. During training, the goal is *to teach* the decoder to predict the next token at every position of the target sequence. To do so, the model is given the entire correct target sequence, so that it can learn many next-token predictions in parallel (cf. *multi-head attention*). This is the so called *teacher forcing* mechanism (Jurafsky & Martin, 2026). However, if the target

sequence were provided without modification, the model would simply receive at each position the very token it is supposed to predict, making the task a trivial copy-and-paste operation. To avoid this, the target sequence is right-shifted as $(\langle SOS \rangle, y_1, y_2, \dots, y_{t-1})$, while the prediction target remains $(y_1, y_2, \dots, y_t)$, with $\langle SOS \rangle$ representing the starting message. Nevertheless, right-shifting alone is not sufficient to preserve causality, since parallelization stands.

Masked Multi-head attention. Thus, additional constraint is added, that is *masking* (Vaswani et al., 2017). The decoder employs a *causal mask M* that operates directly onto the attention matrix, setting the entries corresponding to future [available] positions to minus infinity before the softmax (cf. Vaswani et al., 2017). Formally (Zhao et al., 2026), this is achieved by:

$$MaskedAttention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}} + M)V$$

After this masked self-attention step, the decoder has built a causally valid representation of the partial target sequence. This representation then enters the encoder-decoder attention, or cross-attention.

Cross-attention (or encoder-decoder attention). It is the third sub-level present in the decoder and, again, is implemented on a multi-head structure to allow parallelization. The substantial difference with the other mechanisms of attention is that *cross-attention* is not a process of *self-attention*. In this case, inputs come from different levels and do not belong to the same sequence. Cross-attention receives keys and values from a "side-channel", that is the encoder output – K_{enc}, V_{enc} – and queries from the current generation of the decoder (i.e. from masked multi-head self-attention) – Q_{dec}, so that the cross-attention score (Zhao et al., 2026) is computed as:

$$CrossAttention = softmax\left(\frac{Q_{dec}K_{enc}^T}{\sqrt{d_k}}\right)V_{enc}$$

The purpose of cross-attention is to allow the decoder to generate the output sequence while remaining conditioned by the input sequence "encoded by the encoder". This conditioning is essential as the architecture implemented by Vaswani et al. (2017) is primarily designed for transduction tasks (see "Attention mechanism" section), such as automatic translation. In these cases, the model must transform a source sequence into a target sequence (e.g., from a

sentence in English to a sentence in Italian). We have already seen that the encoder reads the entire source sentence and transforms it into an array of contextualized representations, and the decoder generates the target phrase one token at a time. Cross-attention is where these two processes meet. The distinction of origin of queries, keys and values allows the decoder to focus on the input phrase and establish which parts are most relevant to produce the next token. For example, if the input sentence is $[The, black, cat, is, sleeping, on, the, sofa]$, after the encoder contextualization, the decoder begins to generate the Italian translation. Suppose it has already produced $[Il, gatto, nero, sta, \dots]$, i.e. the translation of the first four tokens. To decide the next token, the decoder pays attention to the content of the English phrase. Through cross-attention the system can interpret $[sleeping]$ as most relevant token, since it is the other part of the predicative verb. The decoder will assign more weight to the corresponding representations and will use the associated values to generate a coherent Italian expression of the next token, for example $[dormendo]$. Thus, the cycle repeats itself.

To complete the architectural description made, it is necessary to mention a transversal component that accompanies all the blocks of the Transformer and that, for expository reasons, has not been discussed separately. In Vaswani et al. (2017), each sub-layer of the architecture is wrapped by a *residual connection* followed by *layer normalization*, according to the general scheme $LayerNorm(x + Sublayer(x))$. This mechanism, commonly referred to as *Add & Norm* (cf. Fig. 11), does not alter the functional logic of the individual modules. Residual connections facilitate gradient propagation and preserve information across successive layers, while normalization helps stabilize internal distributions during training. For this reason, Add & Norm should be regarded as an essential architectural component, rather than as a merely implementational detail. Its transversal position within the Transformer justifies addressing it here in a unified way, as a technical condition that supports the trainability, stability, and computational depth of the architecture as a whole.

Having just depicted the foundational Transformer architecture, further important considerations need to be made.

2.2. LLM and Generative AI

We said that the heart of LLMs is the Transformer architecture. We shall now delve deeply in this statement.

The paper by Vaswani et al. (2017) introduces the Transformer architecture in its original encoder-decoder form, that makes sense especially in transduction tasks (Vaswani et al., 2017) in which there are two distinct sequences, i.e. a source sequence to be encoded and a target sequence to be generated. This architecture then evolved into two main variants, each specialized for specific tasks: *encoder-only* and *decoder-only* structures. Before describing these variants, it is worth noting that the encoder and decoder concepts predated the Transformers. However, the variants we intend to describe here lie on the major outcome of Transformers, with respect to RNN limitations: standing on attention mechanism only. In other words, before 2017 encoder-decoder architectures were already used, but based on RNN or LSTM.

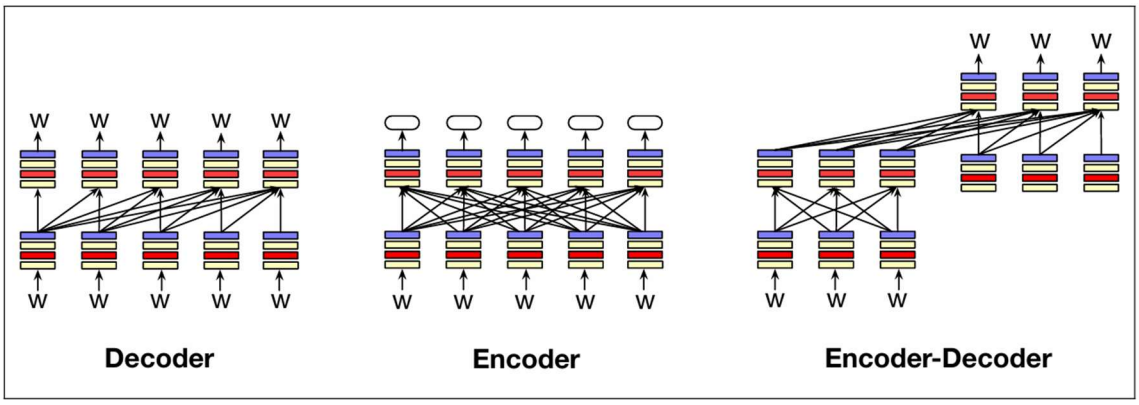


Figure 13: Three architectures for language models: decoders, encoders, and encoder-decoders. The arrows sketch out the information flow in the three architectures. Decoders take tokens as input and generate tokens as output. Encoders take tokens as input and produce an encoding (a vector representation of each token) as output. Encoder-decoders take tokens as input and generate a series of tokens as output (Jurafsky & Martin, 2026).

Encoder-only. This variant considers only the encoder part of the Transformer. It became popular with the release of BERT – *Bidirectional Encoder Representations from Transformers* – released by Google in 2018 (cf. Devlin et al., 2019). Since the encoder is made up of multi-head self-attention mechanism, it is *bidirectional*, that is being able to look for all the tokens in input (i.e. left or right). This bidirectional context power produces not text word-by-word, but deep vector representations. Thus, BERT was considered the state-of-the-art model for NLU-tasks (i.e. *Natural Language Understanding*), such as classification, sentiment analysis (Singh et al., 2025) and other tasks where the input is a text and the output is something different, like labels (Jurafsky & Martin, 2026).

Decoder-only. In a decoder-only setup the architecture is streamlined into a stack of identical layers. Each layer focuses on predicting the next token in a sequence by looking only backward at the preceding context (i.e. autoregressively). Since the architecture is decoder-

only, the cross-attention component from the foundational Transformer structure loses its purpose. What remains is layers of *masked multi-head self-attention* and *feed-forward network*. This type of architecture represents the actual breakthrough of Transformers.

A fundamental intuition underlying language models is that almost anything we want to do with language can be modeled as conditional generation of text. Conditional generation is the task of generating text conditioned on an input piece of text (Jurafsky & Martin, 2026). Conditional generation is the profound intuition behind the applicability of decoders. Since they are able to provide conditional generation, and since the purpose of LLMs is the same, then *decoder-only Transformers*, became the heart of modern LLMs.

There is a common tendency to make the concept of Large Language Model coincide with that of Transformer, even if not equivalent. Having sufficiently described the Transformer architecture, it is useful to point out that an LLM is a neural network trained on *extremely large* textual corpora to learn to predict the next token given a context. Although modern LLMs are not the same as Transformers, they have been made possible by the application and scalability of Transformer architectures to language modeling. So it is reasonable to see LLMs as the result of the convergence of three elements: a specific neural architecture – in most cases based on Transformer decoders, a very large number of parameters learned during training and large textual datasets used for pre-training. This leads us to a refined definition of LLMs: large-scale language models that leverage the Transformer architecture to learn linguistic representations and generate text through autoregressive prediction of the next token.

Having clarified this distinction, It is usually said that the Transformers are the basis of the Generative AI (or GenAI) revolution led by OpenAI with the GPT series (i.e., *Generative Pre-trained Transformer*) – the engine underlying the well-known user interface ChatGPT (or others like Claude by Anthropic and Gemini by Google). What is meant is that, although the origins of GenAI are earlier (cf. Goodfellow et al., 2014), LLMs constitute the largest development achieved so far.

Finally, Decoder-only Transformers and LLMs have revolutionized AI and GenAI in particular. We will then see which aspects characterize this revolution (see. Section 2.4)

2.3. Training Large Language Models

The training process of an LLM generally consists of three phases: *pre-training, fine-tuning* and *preference alignment*.

Pre-training. Jurafsky and Martin define pretraining as «learning knowledge about language and the world from iteratively predicting tokens in vast amounts of text» (Jurafsky & Martin, 2026). During pre-training, the model is subjected to huge textual corpora, that is then stored in memory. The typical learning of bidirectional (encoder) architectures, like BERT, involves masking some words before the text is passed into input. The purpose of the model is to predict missing target words, that are revealed only after each system output attempt. GPT-like models (so decoders-only), on the other hand, follow the auto-regressive approach described above. Also in this case it is necessary to provide a masking mechanism, to avoid copying and pasting. Here, unlike bidirectional models, masking is not applied to the pre-input text, but takes place within the attention process itself. This is allowed by the architecture of the Transformer decoder-only itself (see "Decoder" section), i.e. by *the masked multi-head self-attention* layers.

Recall that LLMs are Large not only because they are trained on large textual datasets, but also for the very large number of parameters the net contains, i.e. *weights* (e.g. attention matrices, feed-forward networks weights, embedding matrices, layer normalization weights) and *biases*, that constitute the group of learned parameters, and also *architectural hyperparameters* and *sampling parameters,* project choices fixed externally by developers.

How does training actually take place? The system produces the first outputs on the next token (again through probability distributions over the entire vocabulary). At first, because the network weights are initialized in an uninformed way, these deployments are inadequate. The correct token, however, is already known, since it is simply the token that actually appears in the corpus in that position (see teacher forcing mechanism). At this point, the network uses cross-entropy to minimize the error between output and target, and then the backprop to update the network internal weights in the direction that reduces the error. By repeating this process over huge amounts of text, the model becomes progressively more capable of assigning high probability to statistically appropriate tokens in a given context.

Self-supervised training. The process described in this way is an example of self-supervised learning, or SSL (see Section 1.3.2), because it does not require external or human-made labels. In fact, It is the natural sequence of words that act like its own supervision (Jurafsky & Martin, 2026). As we have already said before, SSL represents a paradigm that has revolutionized the mechanism of neural network training, because it allows models to be trained on colossal amounts of data without human intervention for labeling.

Fine-tuning. After pre-training, the model can be further trained on particular or uncommon contexts or data. For example, one can specialize the model within the medical context by subjecting it to *instruction tuning*, or *supervised fine-tuning*, which trains the model on instruction-response pairs to perform specific and content-oriented tasks (Jurafsky & Martin, 2026). Finally, it is possible to subject the model to *preference alignment* (often via RLHF, i.e. *Reinforcement Learning from Human Feedback*)[13]. Here, human preferences or comparison data between accepted and rejected responses are used to orient the model towards more useful and less harmful continuations (Jurafsky & Martin, 2026; Lewis, 2025).

2.4. What LLMs have changed?

The revolution brought about by Large Language Models (LLMs) is located not only on a technical and performance level, but also on an epistemological one. On the other hand, some epistemological consequences that we are going to describe have been made possible thanks to the technical innovations of this field of research.

The first epistemological consequence of the advent and large-scale diffusion of LLMs concerns the relationship between knowledge and its representation. LLMs are located at the end of a trajectory that leads from symbolic AI to Machine Learning, then to Deep Learning and finally to Transformers and LLM. As highlighted by Barelli et al. (2025), while the first paradigms – imperative-procedural and logical-declarative – operate through top-down knowledge, in the form of an explicit play algorithm, machine learning and, in particular, deep learning is characterized by a bottom-up flow of knowledge, starting from the examples that are provided (Barelli et al., 2025) and where the machine learns from those examples. This technical-architectural transformation has radical epistemological consequences. In the

[13] RHLF has been formalized by Christiano et al. (2017) and applied on large scale to LLM (instruction tuning + RLHF) by Ouyang et al. (2022).

first case, the machine executes instructions or applies clear and explicit rules provided by the programmer, making the process interpretable and symbolic (Barelli et al., 2025). So too the knowledge that the model produces is interpretable and symbolic, since it is codified in legible propositions, symbols or rules. On the contrary, in Deep Learning the model learns from the data themselves, according to numerous and continuous iterations during which the network progressively updates its matrices of numerical weights. The knowledge produced is, following Barelli et al. (2025), sub-symbolic and not interpretable. Hence the constitutive opacity of neural networks translates the fact that they are sub-symbolic not because they hide within them a rule of operation – which could therefore be made explicit – but because knowledge itself does not take the form of a rule.

This distinction is also reflected in the psychological field. Traditionally, psychology, and consequently symbolic AI, has imagined the mind as populated by highly structured information, such as logical rules, *grammars*, or lists of attributes (Lewis, 2025), which can be summed up, for example, in Newell and Simon's physical symbolic system hypothesis (see section 1.3.1). Superficially, in the classical view, knowledge is considered to be contained a priori in the mind. Completely different is the connectionist approach (see section 1.3.4), i.e. the framework of cognitive sciences that has driven neural networks and deep learning development. Here, knowledge is generated by the neural network (i.e. during the process) and, thus it is not a priori infused in the form of an algorithm. That is more acceptable if we think about how we reason on everyday life.

LLMs radicalize this transformation. It should be recalled that the learning of an LLM is a distributed learning, i.e. knowledge is not memorized literally but encoded in contextualized and sub-symbolic vectors and numerical weights (Barelli et al., 2025). Consequently, it is reasonable to say with Lewis that this knowledge is «a model of the corpus, not a record of the corpus» (Lewis, 2025). This abstract nature is crucial to allow the model to generalize, i.e. to apply learned regularities to unprecedented prompts and contexts never seen before (that is the major outcome of LLMs). Thus, when asked to answer it does not look for a memorized sentence, but produces what is statistically and semantically plausible in that context (Lewis, 2025). Bialek (2025) interprets these innovative performances as an emerging phenomenon in a physical sense. In other words, they did not manifest themselves until the networks had enough degrees of freedom and until the datasets became large enough (i.e. LLM conditions of existence). Cristianini is also of the same idea and specifies that the

most useful skills of language models emerged spontaneously while they were being trained to perform a different task. So, to know what else may emerge in the future, we need to consider the effect of model size on such a phenomenon (Cristianini, 2024). In Transformers, moreover, the attention mechanism allows to control the combinatorial explosion of possible interactions, ensuring that the model focuses only on certain relevant combinations (Bialek, 2025). This is the so-called phenomenon of head specialization (i.e. the components of attention mechanisms). In this sense, scale (i.e., the size of the language model) is not a simple architectural factor; beyond a certain threshold it produces new qualities, that is the definition of *emergency*.

Lewis (2025) insists that the unexpected abilities of LLMs compared to previous forms of AI (e.g., answering very different questions, solving puzzles, producing analogies, mastering natural language, writing code), emerge from the seemingly simple task of predicting the next text. Prediction, therefore, is not a marginal skill but a foundational principle, therefore epistemologically relevant. Recalling Lewis (2025) – «a model of the corpus, not a record of the corpus» – in order to well predict a word, the model must capture grammatical, semantic, pragmatic, cultural and inferential regularities. It must, at least functionally, take into account what is being talked about, the context, the relationships between concepts and discursive conventions.

Mugleston et al. (2025) argue that, since LLMs learn correlations in high-dimensional vector spaces, by modifying network weights to associate words and concepts in statistically significant contexts, their ability is relational. Although they do not necessarily possess an embodied understanding of the world (see Section 2.5), they nevertheless construct patterns of relationship between linguistic elements that allow semantically sophisticated behaviors (Mugleston et al., 2025). This ability cannot be explained only as emerging from the computational scale (from the size of the model and the vastness of the data), but must also be traced back to the nature of the object itself on which the model is trained, i.e. natural language. The latter is not a simple set of strings, but a relational structure in which human knowledge, practices, inferences and ways of organizing experience have settled. LLMs help to bring out some characteristics that are already internal and latent in the language itself. This is Stephen Wolfram's view, according to which LLMs have done nothing but discover that «human language (and the patterns of thinking behind it) are somehow simpler and more "law like" in their structure than we thought» (Wolfram, 2023). This strong claim makes us

consider at least that the emerging capabilities of LLMs emerge not only from the interaction of the Transformer architecture with the scale of the model and the quantity of data; also a fourth factor is added: the relational structure of natural language. That is the very reason to say that the epistemological revolution of LLMs focuses on natural language as a computational medium and that, consequently, linguistic prediction becomes a driver of much broader skills than textual completion.

In conclusion, the epistemological revolution of the LLMs has brought to light several relevant aspects, even in tension with each other. On the one hand, they show that knowledge, representation, and reasoning can emerge without being explicitly encoded in symbolic form. On the other hand, they show that generative, sub-symbolic and distributed knowledge is effective but opaque, probabilistic and semantically plausible. It is knowledge that emerges from data, errors, corrections, weights, attention, scale and language. The stakes are high and, for this reason, criticism is not spared.

2.5. From Performance to Episteme: The Problem of Knowledge

When it comes to the limitations of LLMs, several topics are always discussed. We talk about hallucinations (Huang et al., 2023; Ji et al., 2023), absence of reflective knowledge (Ackerman, 2025; Griot, 2025), strict dependence on training data and therefore transmission and amplification of biases (Bender et al., 2021; Wang et al., 2024), abductive leap impossibility (Zahavy, 2026), and others. Some of these limitations could be addressed by improving architectures, others are structural. The problems that arise with LLMs are not, however, only of a technical or infrastructural nature, but also concern dimensions that intercept human life, in different ways.

In recent decades, the rapid diffusion of Machine Learning models throughout the social fabric – for example in sectors such as education, finance and healthcare (Kim et al., 2024) – has brought with it problematic aspects. The situation is even more delicate if we consider that in some of these areas, characterised as high-stakes contexts, artificial decision-making systems are already required and used (Baron, 2025). The problematic aspects stem from the fact that these systems are commonly considered opaque, that is, they show a lack of transparency regarding the internal mechanisms involved in the mediating process between input and output, with the consequent lack of transparency regarding the computation modality of the instances involved. To describe such systems, the widely used metaphor is

that of the black-box problem, firstly formalized by Ashby (1956). The opacity of machine learning systems stems from the convergence of data complexity, the large number of parameters and their non-linear interactions, in addition to the dynamics of the learning process itself (Chandrasekaran & Jordan, 2013). For this reason, the field of eXplainable AI (XAI) has emerged as an attempt to make AI systems and their outputs more interpretable and explainable (Boge & Mosig, 2025), while recently developed human-centered XAI approaches insist that an explanation is relevant not only when it is formally accurate, but when it is meaningful for the user in the context of use (Kim et al., 2024). These issues are inseparable from broader normative questions concerning trust, trustworthiness, responsibility, privacy, reliability and fairness (Buschmeier et al., 2025; Durán & Pozzi, 2025). In this sense, the social integration of AI requires not only technical explainability, but also an ethical framework capable of orienting the design and use of algorithmic systems toward human values, rights and responsibilities – what Benanti calls the urgency of an "algorethics" (Benanti, 2023).

This brief overview makes us comprehend the complexity and multidimensionality of the situation. By now, we want to focus firstly on the knowledge-related issues, that are the wellspring from which many of the problems discussed above. The purpose is to question which type of knowledge is really produced by LLMs and how it interacts with human beings.

The real output. In the previous section we said that the knowledge produced by an LLM is sub-symbolic and uninterpretable. It does not take the form of a rule but emerges in the functioning of the network and in the continuous updating of the weights of the network. The architecture of LLMs is oriented towards the purpose of generating the next token of the input series. This predictive and, at the same time, generative capacity results in an output that is statistically and semantically plausible (given a specific context). Polverini and Gregorcic (2024) well summarize this point by saying that the LLM, in a simplified form, always answers the question: "According to your model of the statistics of human language, what words are likely to come next?" (Polverini & Gregorcic, 2024). The model can learn ontological, semantic, syntactic, mathematical or factual associations, even biases and spurious correlations through its mechanism of repeated token prediction but, ultimately, the forms of thought involved (e.g., knowledge, reasoning or competence) come from nothing more than the internalization of statistical patterns present in the data (Jurafsky & Martin, 2026; Polverini & Gregorcic, 2024). To support this argument comes the literature

agreement on the fact that LLM's learning does not possess any characteristic of intentional learning, i.e. they do not possess consciousness.

Butlin et al., in *Consciousness in Artificial Intelligence* (2023), propose a rigorous neuroscience-based approach to assessing the degree of AI consciousness. The analysis suggests that no current AI system is conscious. However, Butlin et al.'s approach brackets the biological substrate. In other words, the authors adopt as a research methodology quantitative indicators that refer to the approach of computational functionalism, according to which the execution of calculations of the correct type is a necessary and sufficient hypothesis for consciousness, eliminating a priori any consideration of the biological support (somehow linked to Turing's thought).

The other part of the literature agrees on the opposite. Since '90, Harnad (1990; 2002) has argued that bodies or sensorimotor experience are necessary for true language understanding, thus for having and producing knowledge. Since AI systems lack this, they have no way to "ground" the symbols they are trained on. In this framework, human learning is not only characterized by intentionality, but also by the possibility of experiencing (the world), i.e. the *being-in-the-world* dimension (Gahrn-Andersen, 2025), that is, the incarnation (i.e. better called *embodiment*). In fact, as Dreyfus (2006) famously argues, what separates humans from machines is the fact that we exist *in situations*, not apart from them, and that we do so in an embodied and non-conceptual way, and – as Heidegger said – «handiness is the ontological categorical definition of beings as they are "in themselves"» (Heidegger, 2010). This is useful if we want to avoid reducing the problem to a Cartesian framework.

Subsequently, part of the literature argues that the natural solution to this problem (i.e. unintentionality and non-experience) could be found in *Multimodal Large Language Models* (MLLMs) (Driess et al., 2023; Girdhar et al., 2023; Huang et al., 2023; Radford et al., 2021), which learn to associate linguistic representations with information from other *modalities*, such as vision or sound. Chalmers (2023) believes that there is a greater than 25% chance that extended systems endowed with senses and embodiment could become conscious within a decade (Chalmers, 2023). However, there is still considerable disagreement over whether MLLMs exhibit the necessary interaction between linguistic and sensorimotor inputs that appears to underpin grounding in humans (Mollo & Millière, 2023; Tong et al., 2024;

Shanahan, 2023; Bisk et al., 2020). Since the situation and the judgement are complex, specific problematic questions arise.

***Epistemia* or The Illusion of Knowledge.** We agree on the fact that LLM's fundamental purpose is to generate statistically plausible next tokens. Thus, it is reasonable to record that the generated text is not grounded in communicative intent. This «can seem counter-intuitive given the increasingly fluent qualities of automatically generated text, but we have to account for the fact that our perception of natural language text, regardless of how it was generated, is mediated by our own linguistic competence and our predisposition to interpret communicative acts as conveying coherent meaning and intent, whether or not they do» (Bender et al., 2021). We could say, poetically, that "coherence lies in the eye of the beholder". For example, from Loru et al. (2025) emerges that humans are strongly influenced by stylistic heuristics, such as the fluidity of writing or emotional tone, which can mislead the perception of truth. Walter Quattrociocchi – professor at La Sapienza University in Rome and leader of this research – insists on the concept of *epistemia,* the tendency to confuse linguistic form with epistemic reliability (Loru et al., 2025). Epistemia means the illusion of knowledge that emerges when surface plausibility replaces verification.

Although the research of Loru et al. (2025) focuses primarily on classifying differences in the type of reasoning employed by humans and LLMs in assessing the reliability of information, they warn that if humans begin to uncritically delegate their judgments to LLMs, there is a risk of a substitution of reasoning, from deliberative and critical to approximate reasoning, based on statistical patterns (Loru et al., 2025), which seems to be able to describe man's tendency to homologation with artificial systems. This results in the aforementioned "illusion of knowledge", the dynamic in which plausibility replaces the veracity of facts. Moreover, «as these systems scale, the issue is no longer only whether outputs are correct, but how the very notion of judgment is operationalized once decisions are delegated to statistical models» (Loru et al., 2025).

We should recall that epistemia did not arise only with the widespread use of LLMs. In general, we speak of the illusion of knowledge even within the classical psychological approach, where the illusion lies in the fact that any form of knowledge – be it explanation, judgment, argumentation – is produced without the epistemic conditions of knowledge itself being guaranteed (Rozenblit & Keil, 2002). LLMs and modern AI only made this process more evident.

The need for Agency. For the reasons described so far, Luciano Floridi argues that AI should be understood not by unduly expanding the concept of intelligence, but rather by broadening that of *agency* – traditionally defined as the capacity for autonomous action and deliberate decision-making (Himma, 2009) – including within it artificial entities that, in fact, act without possessing mental states, cognition or intentionality. Floridi, in this sense, advocates a real conceptual and lexical change, proposing to use the concept of *Artificial Agency* (AA) instead of *Artificial Intelligence* (AI), as the very notion of intelligence is misleading. Starting from this, the author – in *AI as Agency without Intelligence* (Floridi, 2025) – identifies a new frontier called *Agentic AI* (or Social Artificial Agency), that represents an emergent form of collective agency, where autonomous, coordinated AI systems interact to achieve complex goals with minimal human oversight. It is in this horizon that AI can be best developed without the risk of considering it intelligent, i.e. avoiding anthropomorphic and biological fallacies (Floridi, 2025).

Why can't we stop at «it's just statistics»? On a purely descriptive level, the argument built up so far is fundamentally correct. LLMs produce essentially statistical knowledge, in the sense that they learn regularities and produce outputs based on context-conditioned probabilities. At the epistemological level, however, arguing that the statistical framework of these models excludes a priori any possibility of relationship with knowledge, learning or reasoning may prove to be a partial judgment, which risks reducing the complexity at stake, or at least the possibility of interesting and fruitful reflections on the problem of knowledge – human and artificial.

The research of Tenenbaum et al. (2011) helps us to understand, at its essential level, that there are forms of probabilistic modeling that are much closer to explanation, inference, generalization from few data, construction of hypotheses, than simple LLM statistics. The starting point of the paper (Tenenbaum et al., 2011) is not a general defense of statistics, nor a direct reflection on Large Language Models, but a more radical question about how learning works: «How do our minds get so much from so little? We build rich causal models, make strong generalizations, and construct powerful abstractions, whereas the input data are sparse, noisy, and ambiguous – in every way far too limited» (Tenenbaum et al., 2011), common problem of cognitive science. In fact, this already marks a non-negligible difference compared to LLMs which, on the contrary, learn from huge amounts of data. This human capacity for generalization – at two different levels of expertise, the child, who can learn how to use a new word from seeing just a few examples (grasping the actual meaning

because he can use the word appropriately in new situations), and the adult, who constructs larger-scale systems of knowledge such as intuitive theories of physics, psychology, or biology or rule systems for social structure or moral judgment – appears difficult to reduce to a simple accumulation of observed frequencies. This is the well-known "problem of induction", on which the greatest philosophers have confronted each other from Plato and Aristotle through Hume, Whewell, and Mill to Carnap, Quine, Goodman, and others in the 20th century (Godfrey-Smith, 2003). The answer to this problem that Tenenbaum et al. (2011) argue is that human learning can be described as probabilistic inference driven by abstract structures (remaining within the constraint of inductive learning).

How? Abstract knowledge is encoded in a *probabilistic generative model* (Tenenbaum et al., 2011). *Probabilistic*, because human learning takes place – in this theoretical framework – in conditions of uncertainty about the states of the latent variables involved in the process. More specifically, the learner never possesses the global and entire structure of the world, at most he experiences one of its partialities, which refers to the already depicted ability to generalize from few data. Again, from these data the subject infers something that is not immediately observable: a category, a link, a cause, a rule, a hidden property, a latent structure. Probabilistic, in this sense, does not refer to the need to count frequencies, but to attempt hypotheses from incomplete data. *Generative*, not in the sense of generating an output (as in AI), but because it is understood as a mental model that describes the causal processes of the world from which the subject's observations derive (i.e. a model that links latent structure or cause to the observed data). In this process, the mind uses Bayes' rule as a tool to update its beliefs (the plausibility of hypotheses) about latent variables in light of the observed data. It should be noted that the process combines the ability of the hypothesis to make those data expected with the prior knowledge (Tenenbaum et al., 2011), a reflection of an already structured hypothesis space, made up of categories, hierarchies, causal models and abstract schemas, a real organized theoretical framework – contrary to what we will see later with *Knowledge in Pieces* (see Chapter 3). Without this space of hypothesis, a meaningful generalization would be impossible.

Thus, the argument of Tenenbaum et al. shows that the statistical property that drives a process of knowledge production is not sufficient to define what that knowledge really is and, in particular, that "statistics" is not synonymous with "non-knowledge". This important emphasis, in reality, does not save LLMs from criticism and, on the contrary, can make it more solid. The use of Tenenbaum’s thesis in this respect is motivated by the need not to

stop at immediate answers, but to deepen the judgment, even to the point of best corroborating the distinctions between human and artificial learning.

Thus, the problem with LLMs is not simply that they use probabilities – because humans also use probability (cf. Goldstein et al., 2025) – but that their probability is not, at least structurally, a probability on explicit hypotheses about the world, but a probability on linguistic configurations (beyond reflections on deductive and inductive learning).

World Models. Tenenbaum et al. allowed us to clarify that "statistical" does not necessarily mean "superficial". Yann LeCun allows us to take a next step, namely that to talk about intelligence, learning and reasoning it is not enough to produce plausible outputs, but it is necessary to have a model, at least implicit, of how the world presents itself.

LeCun, currently head of the startup *AmiLabs*, is involved in the development of new AI models that can overcome current limitations. As we have already said, if on one hand man has the ability to generalize and build theories of organized knowledge starting from a few examples of the world – which LeCun calls world *models*, i.e. internal models of how the world works (LeCun, 2022) – by contrast, to be reliable, current ML (and DL) systems need to be trained with very large numbers of trials so that even the rarest combination of situations will be encountered frequently during training. Still, «our best ML systems are still very far from matching human reliability in real-world tasks such as driving, even after being fed with enormous amounts of supervisory data from human experts, after going through millions of reinforcement learning trials in virtual environments, and after engineers have hardwired hundreds of behaviors into them» (LeCun, 2022). So LLMs, in this topic, are also defined by these same limits. For LeCun, current artificial intelligence has not yet reached the reasoning and planning capacity typical of humans and animals. For this reason, the line of research within which LeCun is located is in close proximity to the positions taken by Floridi, i.e. with the focus on the agency dimension. The aim is to create *autonomous intelligent agents*, which have the opportunity – by interacting with the environment, that is, by experiencing it – to make informed decisions. On the basis of these theoretical limitations, it is even more reasonable to see LLMs limits, in terms of experience of the world and of embodiment. Cristianini in *Machina Sapiens* (2022) is not so critical of LLMs and, on the contrary, argues that what Transformer architecture (the foundation of LLMs) produces, in contact with huge amounts of text, is certainly a model of language, but probably also of the world – even if this understanding extends to very general aspects (Cristianini, 2022).

The notion of *world models* must therefore be formulated with caution. Lewis (2025) argues that the LLM is a model of the corpus, not a record of the corpus; other authors discuss whether and to what extent, from sufficiently rich texts, a model can learn regularities that correspond to states of the world. Mugleston et al. (2025), for example, recall studies in which Transformers trained on textual sequences of moves in chess seem to plot the state of the chessboard to predict legal moves. This does not automatically prove that LLMs possess a world in the human sense of the term, but it does show that, under certain conditions, linguistic prediction may require the construction of internal representations of states not immediately explicit in the text (Mugleston et al., 2025). LeCun's research, on the other hand, has led to the development of LeWorldModel (LeWM), a stable end-to-end method for learning latent world models of environments (Maes et al., 2026), based on a completely different architecture from that of Transformers (precisely because it wants to overcome the limitations of LLMs, i.e. computational problems of generation, lack of grounding of the language model). This paves the way for us to enter even deeper into the debate between intelligence and agency, verifying their intersections and differences.

The purpose of this topic is not just to evaluate the differences or similarities between human and machine (or autonomic) learning. In a broader horizon, we are concerned with evaluating the reasonableness of the construction of an adequate theoretical space in which it is possible to compare different models of knowledge and learning. It is an attempt to build a critical parallelism between frameworks.

Up to this point, we have managed to avoid two opposite extremes of judgment regarding the nature of LLMs: on the one hand, it is reasonable not to reduce LLMs to mere "stochastic parrots" (Bender et al., 2021); on the other hand, it is even more prudent and sensible not to attribute to LLMs intelligence and knowledge beyond their actual possibilities. Thus, as a result, it seems interesting to draw parallels with human learning: on the one hand, it is not reasonable to think that the human subject learns as a simple statistical system, limiting himself to accumulating regularities or reproducing patterns encountered in experience; on the other hand, it would be equally simplistic to imagine human knowledge as a system that is already fully coherent, explicit, logically organized and transparent.

The need for a Knowledge Analysis approach. In this space of hypothesis lies the need to take an approach that is intrinsically able to address the problem of learning. If the previous analysis has allowed us to question the status of the knowledge produced by LLMs, now it

is necessary first of all to deepen the dynamics of human learning, useful for establishing the theoretical space within which to compare them (see Chapter 4). The approach we intend to take is that of Knowledge Analysis (diSessa, Sherin & Levin, 2016) and, in particular, that of *Knowledge in Pieces* (diSessa, 1993; Amin, 2025).

Chapter 3

In this chapter, we intend to deepen the dynamics of human learning, following the argumentative structure of the previous chapter. This will make it possible to develop, in a natural way, the theoretical scaffolding with which the theoretical frameworks under consideration will later be compared. In order to investigate human learning, we have chosen to use the Knowledge in Pieces approach (diSessa, 1993; Amin, 2025) and the methodology of Knowledge Analysis (diSessa, Sherin & Levin, 2016), developed within the learning sciences in the context of research on conceptual change in the learning of physics.

This chapter follows a structure that is parallel to that of the chapter on LLMs, though not identical. In the case of Transformers, it was possible to begin directly from the architecture, because the object in question was already technical and formalized. In the case of Knowledge in Pieces and Knowledge Analysis, it is necessary to begin from their theoretical genealogy, because they arise within a specific context of inquiry into the nature of intuitive knowledge in the learning of physics. Only after carrying out this historical reconstruction will it be possible to describe the internal architecture of KiP, the way in which it interprets learning, and finally the epistemological changes and limits that this perspective introduces.

3.1. Positioning

Knowledge in Pieces (KiP) and, subsequently, Knowledge Analysis (KA) emerged in conjunction with the birth of Physics Education Research (PER) and in its intertwining with cognitive and learning sciences.

The fundamental research question of KiP is: what form does students' knowledge have before formal instruction, and how does this knowledge participate in the process of scientific learning? Briefly, KA is the study of the content and form of knowledge for the purpose of understanding learning (diSessa, Sherin & Levin, 2016) or, again, «what is knowledge, and what is the right language with which to characterize it in order to understand how individuals learn?» (diSessa, Sherin & Levin, 2016).

We now examine, historically, how this direction of attention comes to be delineated.

3.1.1. Cognitive revolution and Physics Education Research

Within the field of pedagogy, the most widespread and accepted model during the first half of the twentieth century was *behaviorism*. Developed by John Watson in the early twentieth

century, it assumes the mind of the individual to be an unknowable entity, a black box that cannot be investigated internally. The only visible and, therefore, scientifically examinable phenomenon is *behavior*. From a behavioral perspective, learning is interpreted as an observable change in behavior, produced by the interaction between the subject and the environment. Since internal mental processes do not constitute the primary object of analysis, attention is focused on the external conditions that may facilitate, orient, or stabilize particular responses. In this sense, learning can be explained through mechanisms of conditioning, *classical* in the case of Pavlov and *operant* in the case of Skinner, in which stimuli, reinforcements, and consequences play a central role in the modification of behavior.

In the educational domain, some forms of traditional teaching, particularly lecture-based instruction (i.e., "frontal lecture") centered on the transmission of content and the reproduction of expected responses, can be regarded as partially consistent with this framework. In such contexts, the student tends to assume a predominantly receptive role, while learning is often interpreted as the acquisition and reproduction of correct content or behaviors. Rather than deriving directly from behaviorism, this approach retains some traits compatible with it, such as the centrality of the teacher's action, the control of the instructional environment, and the assessment of observable performances.

Cognitive psychology, which we have already mentioned in Section 1.1, emerged in the late 1950s also as a critical reaction to behaviorism. Unlike the latter, which privileges the analysis of observable behavior, cognitivism brought internal mental processes back to the center of inquiry, including perception, memory, attention, representation, reasoning, and meaning-making. The mind thus came to be regarded as an active system for processing, organizing, and transforming information, through which the subject interprets and acts in the world. Within this framework, authors such as Bruner played a central role in the cognitive turn, while Piaget, Vygotsky, Ausubel, Novak, and Gowin contributed in different ways to the development of constructivist, socio-constructivist, and meaningful learning perspectives. In its earliest formulations, cognitivism focused primarily on the subject's internal processes; this orientation was later broadened by approaches that placed greater emphasis on the role of social interaction, language, culture, and context, as in the socio-constructivism inspired by Vygotsky and in Bandura's social cognitive theory.

Thus, with the *cognitive revolution*, it became possible to conceive learning not merely as an observable modification of behavior, but as a transformation of mental representations

and internal cognitive processes. More specifically, by shifting attention from behavior to the content of thought, it became crucial to observe not only what answer a student provides, but also to ask which representations, resources, and processes make that answer possible (diSessa, Sherin & Levin, 2016), emergent. Within an entirely different field of research, yet still connected to the content of this work, the research field known as *eXplainable Artificial Intelligence* (XAI), Buschmeier et al. (2025) argue that the transformative process that leads the user to reach *agency* – the user's ability to guide their decisions in an informed and flexible way thanks to explanation – arises from the intertwinement of deep comprehension ("knowing that") and deep enabledness ("knowing how"). Thus, if the aim is to study human learning, the analysis must take into account the internal process, the developmental process, the transformations that occur within the student, and not merely the outcome produced.

One witnesses a *focus displacement* in several directions, not only from behavior to mental processes, which directly raises questions about the nature of knowledge, but also – since it is no longer sufficient to act solely through external stimuli in order to condition the subject – from locating knowledge as entirely external to the subject (e.g., in textbooks or in the teacher), to taking into account the subject's internal dimension.

If the internal dimension of the subject matters when we ask about the nature of knowledge, then it becomes necessary to understand how knowledge is organized – if it is organized at all – in the subject's mind. To pursue this line of inquiry, it is necessary to add one further element, namely the fact that – while the nature of knowledge and learning has been a topic of analysis in several fields – «many of the debates spurred by the cognitive revolution played out in the context of science teaching and learning, especially within the domain of physics» (diSessa, Sherin & Levin, 2016). It is therefore essential to ask why cognitive theories found in physics one of their principal domains of development and inquiry. This focus is not accidental. It makes it possible to understand the specific role of physics in the study of cognitive processes and, at the same time, to delimit more precisely the scope of this thesis, which is situated within the field of *Physics Education Research*.

Physics Education Research (PER). The development of Physics Education Research (PER) traces some of its conceptual and critical roots to the end of the nineteenth century, when Joseph Mayer Rice, already in 1893, criticized an educational system based primarily on mechanical memorization. In its modern form, however, PER emerged in the postwar period with the observations of Arnold Arons (Arons, 1998; Beichner, 2009). Noticing that

his lectures left little significant trace in students' minds, Arons began to investigate their conceptual difficulties systematically (Arons, 1998). However, throughout the 1950s and 1960s, the physics community still regarded "education research" primarily as a way of improving the delivery of content rather than as an inquiry into students' cognitive processes (Arons, 1998; Beichner, 2009).

Among the initiatives of this period, it is worth mentioning the *Physical Science Study Committee* (PSSC), a scientific committee established at the Massachusetts Institute of Technology in Boston in 1956 by a group of university physics professors led by Jerrold Zacharias and Francis Friedman, with the aim of revising physics teaching in secondary schools. Following the success of the Soviet space program with Sputnik in 1957, the United States increasingly came to regard its educational system as ineffective, that is, inadequate for providing students – and potential future researchers – with a solid scientific preparation. In response, substantial efforts were directed toward the PSSC project, which sought to innovate instruction through the development of new teaching materials and pedagogical methods. The epistemic vision underlying the project was *content-oriented*, that is, focused on instructional content and, consequently, on the role of the teacher, who was regarded as the primary agent responsible for student learning. The overarching methodological picture was clear: "teaching better means making students understand better". The PSSC certainly contributed significantly to transforming physics education; however, the outcomes fell short of expectations. The instructional framework that it proposed proved to be too rigid and was suited only to a subset of students whose personal epistemological dispositions perfectly resonated with the PSSC approach.

A significant shift took place in the 1970s through the encounter between physics and cognitive psychology. Robert Karplus introduced Piagetian ideas concerning the nature of intellectual development and showed that many university students had not yet acquired the proportional reasoning and variable-control abilities required for learning physics (Arons, 1998). In parallel, Lillian McDermott at the University of Washington established what is generally regarded as one of the first dedicated university-based PER research groups, shifting attention toward the systematic identification of students' difficulties (Beichner, 2009).

The consolidation of PER underwent a decisive acceleration during the 1990s, eventually leading to its official recognition as a field of physics research in 1999 through a declaration

by the *American Physical Society* (APS). A pivotal event in this process was the 1992 publication of the *Force Concept Inventory* (FCI) by Hestenes, Swackhammer, and Wells. The test, based on the dissertation research of Ibrahim Halloun (Halloun & Hestenes, 1985), empirically showed that even university students capable of solving complex mathematical problems performed surprisingly poorly when confronted with qualitative questions involving basic Newtonian mechanics. For example, Halloun and Hestenes, in a study that employed a multiple-choice test given to hundreds of university physics students, found that 44% of these students exhibited the belief that a force is required to maintain motion (Halloun & Hestenes, 1985). The FCI represented a major empirical challenge within research on learning. It should not be interpreted merely as an instrument for assessing students' knowledge; rather, it marks an important epistemological turning point. Students' setback on the FCI cannot be explained simply in terms of a lack of knowledge or of mathematical and procedural difficulties, especially since the questions were word-based and strongly conceptual. Memory-related explanations also appeared insufficient, given that the errors emerged systematically and proved to be remarkably resistant. The issue therefore remained open: how can we explain this unexpected breakdown experienced by students when confronted with simple physics problems? The answer to this question is closely connected to the one raised previously, namely *why the development of cognitive theories found in physics one of their principal domains of inquiry?*

> «To some extent, this is probably an accident of history. But there are also reasons that the learning of physics highlighted what would turn out to be extremely important issues. On the one hand, formal physics seems to encompass a collection of ideas that lies far beyond the informal understanding of the average person. However, at the same time, it is manifestly true that all humans understand a tremendous amount about the natural world simply from our everyday interactions in this world. This leads to some core questions. For example, does it make sense to say that people know any physics, per se, prior to formal instruction?»
>
> (diSessa, Sherin & Levin, 2016)

Prior knowledge and the role of intuition in dealing with the physical world. In order to answer the questions that have emerged, the first observation is that «given that we, as humans, must function in the physical world, it is manifestly evident that we must know a great deal about the behavior of that world» (Sherin, 2006). Human beings, in their being in the world, acquire and structure their own body of knowledge, also through the capacity for generalization mentioned in Section 2.5. This body of knowledge is defined in the literature as *prior knowledge*. The term *prior* denotes the temporal window during which this

knowledge develops, that is, before the experience of formal schooling. The FCI test empirically showed a deep disconnection between prior knowledge and formal or accredited knowledge, organized around explicit disciplinary contents. Particularly evocative is the case of Eric Mazur, a distinguished physics professor at Harvard University. He spent his early teaching career operating under what he calls a «complete illusion» – believing he was a stellar educator simply because his students gave him high ratings and excelled at solving complex, recipe-based textbook equations. However, this illusion completely shattered after seven years when he decided to test his students using basic, concept-driven, word-based problems, exactly the FCI problems. As Mazur vividly recalls, right at the start of the assessment, one student raised her hand and said, looking at her professor: «how should I answer these questions? According to what you taught me or according to the way I usually think about these things?» (Mazur, 2014). And he went: «what's the difference? Shouldn't you think the way I teach you?» (Mazur, 2014). This is precisely what led to the distinction between prior knowledge, viewed as *intuitive* or *common-sense physics knowledge* (Sherin, 2006), and formal physics knowledge, which is considered counter-intuitive. That is why cognitive-based studies on learning started to focus on the emerging questions from physics education.

This staggering realization that his students were merely memorizing formulas without actually understanding the underlying principles fundamentally transformed his approach. It ultimately led him to abandon traditional lecturing entirely and pioneer *Peer Instruction* (Crouch, Watkins, Fagen, & Mazur, 2007) – a flipped-classroom method where students actively debate and explain concepts to one another, proving that real education requires active sense-making rather than passive listening.

Learning in Physics as Conceptual Change. Therefore, if this *prior knowledge* is present in the subject before instruction and often proves resistant to convergence toward formal school knowledge, it becomes necessary to understand how it can be described and modeled. Only starting from this description, it is possible to ask how the transition occurs from intuitive, local, and often implicit forms of knowledge to forms of knowledge closer to scientific ones. In literature, this transformative process is called *conceptual change*. In the 1980s Posner, Strike, Hewson and Gertzog (1982) laid the groundwork for research in this area. Conceptual change, considering its broader aspects, identifies and clarifies that the change is not about adding new knowledge but, instead, concerns transforming existing knowledge into a deeper understanding of scientific concepts. This step is by no means

without challenges. Sherin uses to claim: «conceptual change is not simple learning, but learning when this is particularly difficult» (Sherin, 2021). Different perspectives have emerged, each bringing its own nuance and interpretation to the issue. For example, we have the already cited Posner et al.'s (1982) conceptual change model, Vosniadou's (1994) framework theory and mental model perspective, Chi et al.'s (1994) ontological categories, and Pintrich et al.'s (1993) motivation perspective, some of which will be discussed later.

In summary, alongside the development of instruments such as the FCI, research in PER focused on the problem of describing what intuitive knowledge is made of and how conceptual change occurs (in a general sense). The line leading to Knowledge in Pieces emerges precisely from this need: to understand not only that students possess resistant intuitive knowledge, but how this knowledge is organized, activated, and transformed across different contexts. In what follows, the debate on conceptual change will be taken up again in order to position KiP. To this end, the approaches of misconceptions and Theory-Theory will be introduced in order to underline their differences from KiP, and not in order to represent the terms of the debate. For this, see Amin (2025).

3.1.2. From Misconceptions to Knowledge Analysis

Misconceptions. The first way of responding to this research question (i.e., *what intuitive knowledge is made of and how conceptual change occurs*) follows the paradigm of *misconceptions* or *alternative conceptions*. These are beliefs that the individual develops in preschool age (by virtue of being in the world, as stated earlier), and which are already nominally marked in a negative way insofar as they are contrary and misleading with respect to accredited scientific knowledge. Some would argue that misconceptions research marked the beginning of modern-day physics education research (Docktor & Mestre, 2014). Early works consisted in identifying and listing students' misconceptions and wrong beliefs (Clement, 1982; McDermott, 1984), some of which also included the development of instructional strategies and curricula to revise students' thinking. Since these frameworks considered students' prior knowledge to be incorrect, the methodological approach to education could only consist of "changing knowledge", aligning it with the "true" one. The paradigm of misconceptions has the merit of showing that student error is not random. However, by describing prior knowledge as predominantly composed of incorrect beliefs, it does not make it possible to understand their possible productivity. In other words, a list of incorrect beliefs does not explain well enough why such beliefs recur, organize themselves,

and resist instruction. From here arises the push toward more unitary models, such as *naïve physics* and *theory-theory*.

Theory-theory. Despite students' prior knowledge being considered wrong, agreement developed over the last decades of the twentieth century that this knowledge was instead fairly similar to formal knowledge. The literature in this context moved toward a more structured categorization of intuitive knowledge. The *naïve* resources of novices are much more similar to the structure of theories. Susan Carey's model, the *Theory-Theory of Concepts* (Carey, 1985), consolidates this line. Each of us forms and revises concepts like those developed by scientists. Concepts, in turn, are not seen as isolated notions but as partial theories that embody explanations of the relations between their constituents, of their origins, and of their relations to other clusters of features (Keil, 1989, as cited in Demkanin et al., 2025). Stella Vosniadou also partly follows this direction when she states that «concepts are embedded in theories» (Vosniadou, 1994). Since pupils' intuitive knowledge has theory-like features, it is straightforwardly structured, coherent, and resistant to change (Carey, 1985). One of the most recognized examples of "theory-theory" is the impetus theory, of Aristotelian recall. Students describe objects as needing, in order to move, to acquire an "impetus," an internal force that drives them forward and gradually dissipates along the trajectory, a persistent but inaccurate account with respect to the correct explanation of physical dynamics.

The major limitation of theory-like accounts emerges, in particular, if one observes students' reasoning in relation to counter-intuitive physics problems, exactly like those of the FCI type. In a research project involving Andrea diSessa, David Hammer, and Bruce Sherin, the case of an interview with a university student, the "J case", was discussed and later became a reference point in this field of research (diSessa & Sherin, 1998). In the interview, "J" shows an evident contradiction in describing the forces acting on a ball thrown upward. If at first she provides a scientifically correct answer, claiming that the only force at play during the entire trajectory is gravity, her reasoning changes drastically as soon as the interviewer pushes her to reflect on *what happens at the peak of the toss*, applying what is defined as *focus displacement* or *noticing*. In this second phase, J abandons physics and relies on a more intuitive interpretation, introducing a second force that would have been initially impressed by the hand and would travel with the ball, opposing gravity until it is completely exhausted at the top of the trajectory (Sherin, 2021). This example seems to fully consolidate the theory-like paradigm. However, J's reasoning in the case of *focus displacement* does not

evidently appear as the application of a single and stable theory. On the contrary, Sherin shows that J's answer is "constructed in the moment", through small schemas cued by the context (Sherin, 2021). This is the decisive empirical-methodological point: the fact that students do not always seem to retrieve a ready-made theory, but, on the contrary, often construct the answer locally and contextually, is what directs the change in the modeling.

Standard conceptual change model. Posner et al.'s (1982) model of conceptual change, structured in analogy with the philosophy of science (with respect to Kuhn and Lakatos), around the concepts of *assimilation* and *accommodation*, reflects the epistemic direction just described. Whenever the learner encounters a new phenomenon, he must rely on his current concepts to organize his investigation. Concepts are the protagonists of conceptual change, as seems obvious, and the learner organizes them into a coherent structure. This structure – which has the flavor of a theory-like approach – is what undergoes transformation precisely in the case of conceptual change. It is seen here as a major reorganization of current concepts, where, however, not all current concepts are always replaced (Posner et al., 1982). Individuals will retain many of their current concepts, some of which will function to guide the process of conceptual change. Borrowing a phrase from Stephen Toulmin (1972), Posner et al. (1982) refer to those concepts which govern a conceptual change as a *conceptual ecology*. The work of Posner et al. moves precisely in the direction of identifying the types of concepts that guide conceptual change. For example, there are analogies and metaphors, explanatory ideas, metaphysical beliefs, all entities that explicitly show a "high-level" cognitive structure, that is, they possess an already well-structured intrinsic coherence. This superficial evidence, together with the variety that these concepts display, has led to more than a few reservations concerning this research approach. diSessa and Sherin precisely bring to light the limits of the standard view of conceptual change, defining them as an obstacle to research progress: «Most notably, research has been strikingly inexplicit about what it means by concepts, thus begging the central question, "What changes in conceptual change?" In addition, we believe that researchers have uncritically applied the term concept to what may be very different entities – dog, force, or number. Underlying both of these difficulties is a theoretical imprecision that is insuperable» (diSessa & Sherin, 1998). The limit of the standard model, therefore, is not that it speaks of "concepts," but that it does so through an excessively vague generality. If one does not clarify what a concept is, it becomes difficult to explain what it means to modify, reorganize, or replace it.

Framework Theory. In the perspective developed by Stella Vosniadou, the typical features of the theory-theory approach can be detected, with some differences. At the heart of her work is what she calls a *framework theory*, also referred to below as FT, namely «a skeletal conceptual structure that grounds our deepest ontological commitments and causal devices in terms of which we understand a domain» (Vosniadou, 2018). This structure reflects the typical coherence of the previous approach, although not with the same rigidity and unity. Indeed, the term "theory" is used here differently from what was previously established. On the one hand, a *framework theory* «is not a well-formed, explicit and socially shared scientific theory» (Vosniadou, 2018), in the sense that it lacks the consistency and systematicity of a scientific theory. On the other hand, it is a "theory" because it is a principle-based system that can give rise to explanation and prediction. The individual structures naïve knowledge – relatively coherent but not explicit to consciousness – in preschool age, through interaction with the world. It is formed by ontological and epistemological assumptions that act as constraints in the dynamics of learning. Coherence, therefore, is not necessarily found at the level of single explicitly formulated theories, as in the previous paradigms, but at the level of the deep assumptions that constrain the construction of mental models. For example, in the case of the shape of the Earth, children initially tend to categorize the Earth as an ordinary physical object, applying to it assumptions such as the need for support, the up/down organization of space, and the idea that unsupported objects fall downward (Vosniadou & Brewer, 1992). Vosniadou's paradigm can be considered a sophisticated and structuralist version of the theory-theory approach.

FT differs from previous models not only with respect to the slight de-classification of prior knowledge from rigid theory-theory to principle-based system, but also in how it interprets conceptual change.

In the conception of framework theory, conceptual change, thus, learning is seen as a slow and gradual process of transformation of the framework theory. Vosniadou distinguishes between *enrichment* and *revision*, in quasi-analogy with Posner's model. *Enrichment* consists in the addition of new information to existing conceptual structures, whereas *revision* implies the modification of beliefs, assumptions, or internal relations within the theory. The substantial difference from before is that the object to be changed is not misconceptions as such (also because they are elevated to the rank of *synthetic* or *hybrid conceptions* during learning, that is, marked by a productive nuance even if still scientifically inadequate). Rather, the object to be changed is the ontological and epistemological

assumptions that make them plausible, that is, precisely the framework theory, the supporting skeleton. It is for this reason that Vosniadou states that conceptual change is situated "in" the framework theory, as modifications that occur in terms of categorization, representation and epistemology and the creation of new concepts and new reasoning processes (Vosniadou, 2018). Again, conceptual change requires time: it «is not a sudden replacement of intuitive conceptions with scientific ones, but a slow and gradual process during which fragmented and synthetic conceptions are formed» (Vosniadou, 2018). Sherin (2021) observes that Vosniadou occupies an intermediate position between theory-theory and what will become the paradigm proposed by diSessa: she recognizes that conceptual change requires many coordinated transformations, but attributes the difficulty of change to the fact that initial knowledge is organized in a structure with its own internal coherence and resistance to change. Ultimately, the idea remains that in order to reach a real understanding of physics, parts of the basic framework theory must be changed. This idea will be revolutionized by *Knowledge in Pieces*.

In conclusion, the theory-like approach, also considering its development in FT, has the limitation of attributing excessive coherence to prior knowledge and, consequently, also the limitation of seeking to explain conceptual change according to cognitive structures that are too molar – that is, larger than individual acts of reasoning, such as theories, frameworks, concepts, categories, guiding metaphors, ontological assumptions – and poorly defined. In this sense, diSessa and Sherin (1998) argue that the "concept of concept" is theoretically lacking. On this basis, the next step is not guided only by empirical evidence, although such evidence is present, but is theoretically grounded.

Methodologically, it is appropriate to clarify how we can study the form, content, and transformation of knowledge (diSessa et al., 2016), even before defining a specific model for prior knowledge. This is the level of *Knowledge Analysis* (diSessa et al., 2016), which represents the culmination of this historical-evolutionary trajectory.

Epistemologically, it is necessary to move to a finer grain in the modeling of prior knowledge – to focus attention on *pieces of knowledge smaller than concepts* – and to find the local mechanism (the language) through which these pieces of knowledge are activated, maintained, deactivated, coordinated, and transformed in reasoning. This is the level of *Knowledge in Pieces* (diSessa, 1988, 1993; diSessa et al., 2016).

3.2. Knowledge in Pieces

The previous section brought to light a crucial point in the debate on intuitive knowledge and conceptual change: before explaining how knowledge changes, it is necessary to clarify what it is made of. The difficulties encountered by theory-like interpretations suggest, in fact, the need to describe students' reasoning at a finer grain, one capable of grasping its fragmentation, its sensitivity to context, and its local forms of coordination.

Knowledge in Pieces (diSessa, 1988, 1993) responds to this need not simply by introducing a new class of elements, but by elaborating a theoretical architecture of intuitive knowledge. Within this framework, diSessa pays particular attention to the *sense of mechanism*, defining it as that sense «that accounts for commonsense predictions, expectations, explanations, and judgments of plausibility concerning mechanically causal situations» (diSessa, 1993). It is therefore not intuitive knowledge in general, but a specific form of it, linked to mechanically causal situations. The distinction between *sense of mechanism* and intuitive knowledge will be taken up later. For now, it is sufficient to underline that diSessa conceives it as a *knowledge system* (diSessa, 1993).

Every knowledge system, according to diSessa (1993), must address a set of correlated questions relevant to theory construction. It is necessary to clarify the nature and content of the *elements*; the *cognitive mechanism*, that is, the local process through which such elements are activated and used; *systematicity*, namely the level and type of correlation of the elements in the system; and, finally, *development*, that is, the way in which the overall system evolves over time. This approach will also lead to the definition of the methodological principles of *Knowledge Analysis*, which guide the study of knowledge not only as content possessed by the subject, but as a complex system of representations, forms, contextual activations, and transformations during learning (diSessa, Sherin, & Levin, 2016). KiP, in this sense, is not only a theory that explains learning on the basis of its constitutive elements, but a genuine *epistemological approach to the development of a theory of knowledge*, or, in other words, «a class of theoretical models of knowledge» (diSessa, Levin, 2021), peculiarly connoted. Following the four dimensions of KiP, the research questions become: «What are the elements of knowledge; how do they arise; what level and kind of systematicity exists; how does the system as a whole evolve; and what can be said about the underlying cognitive mechanisms that are responsible for the normal operation of the system and its evolution?»

(diSessa, 1993). In what follows, we will address the description of the elements of KiP, how they interact, and how they develop in learning.

3.2.1. Elements

The first dimension through which diSessa constructs Knowledge in Pieces concerns the elements, that is, the units of knowledge that compose the intuitive sense of physics. KiP does not start from the assumption that the student possesses a naïve theory of motion, nor that the student simply has a list of misconceptions. It starts, instead, from the hypothesis that intuitive reasoning is produced by a complex system of local, often tacit elements that are activated in specific situations (diSessa, 1993; diSessa et al., 2016). These units are called *phenomenological primitives*, or *p-prims*. These small knowledge structures do not have the appearance of explicit propositions or mini-theories. On the contrary, they are cognitive elements that allow the subject to "see" within the situation, for example a physical problem, and to connect certain features of the problem to their own sense of the world.

In fact, they are *phenomenological* because they originate from the subject's interpretation and experience in the world and, once absorbed, constitute a vocabulary through which the subject remembers and interprets experience (diSessa, 1993). In a certain sense, p-prims express the experiential and partly embodied character of intuitive knowledge, since they derive from recurring configurations of acting and perceiving, for example pushing, balancing, resisting. They are also *primitives* because they can function as self-evident explanations. In certain cases, something happens simply «because that's the way things are» (diSessa, 1993). But also, they are primitive since they should be considered the smallest or minimal memory elements, not further decomposable into more elementary parts, «atomic and isolated» (diSessa, 1993). Here, *isolated* does not mean that they are completely apart from everything else, but that they can exist and activate without having to be built by chains of other cognitive elements.

Therefore, in interpretive terms, p-prims represent the basis for explaining the primary and original modality (i.e., recalling the *origin*, the first impact with reality) through which the subject accesses and cognitively reads the physical world.

Examples. diSessa (1993) presents a series of recurring p-prims in the *sense of mechanism*, which he organizes into clusters or phenomenological families. There is the cluster of *force and agency* elements, *constraint phenomena*, the domain of *balance and equilibrium*, and, later, *figural primitives*.

Important p-prims linked to the phenomenological family of *force and agency* are categorized through the triadic structure *agent-resistance-result*. A central example is *Ohm's p-prim*. When one pushes a heavy object and it moves only a little, or when an obstacle limits the effect of an action, the subject can read the situation through the intuitive idea that an action produces an effect proportional to the force of the agent and inversely proportional to the resistance encountered. To the same family belong p-prims with different structures, such as *force as mover* and *continuous push*, for which motion is intuitively associated with the presence of a force that produces or maintains it, close to the idea of impetus. In other cases, instead, the subject relies on p-prims linked to the decay or exhaustion of a certain quantity, such as *dying away*, activated when a motion slows down, a sound fades, or a push seems gradually to be consumed.

A second group concerns *constraint phenomena*, that is, situations in which material constraints guide, block, or support action. P-prims such as *blocking*, *supporting*, and *guiding* make it possible to interpret, for example, a wall that prevents passage, a table that supports a book, or a rail that guides the motion of a cart.

A third set concerns *balance and equilibrium*. Here we find resources such as *dynamic balance*, *abstract balancing*, *overcoming*, and *equilibration*, activated when the subject reads a situation as an equilibrium between opposing tendencies, as in the case of two weights on a scale, one force compensating another, or a system that spontaneously returns to a stable state.

Finally, *figural primitives* depend more directly on the perceptual and geometrical form of the situation, as when a spatial configuration immediately suggests continuity, symmetry, or a privileged direction.

P-prims are not formal definitions, but elementary intuitive schemas through which the subject makes the physical world readable and plausible at first impact (diSessa, 1993).

The Big Picture of KiP. The enormous list of p-prims discussed by diSessa (cf. Appendix B in diSessa, 1993) should not be understood as a closed taxonomy of the elements that constitute intuitive knowledge. Its function is to show types of basic elements that indicate how the subject possesses multiple intuitive schemas, often tacit and non-propositional, through which ordinary physical situations are interpreted. Some p-prims are strongly linked to agency and bodily effort, such as *Ohm's p-prim*, *force as mover*, and *continuous push*; others concern material constraints and configurations, such as *blocking*, *supporting*,

guiding, and *clamping*; still others concern equilibrium, compensation, prevalence, and return to a stable state, such as *dynamic balance*, *abstract balancing*, *overcoming*, and *equilibration*; others, finally, depend on the visual and spatial form of the situation, such as *figural primitives*. This variety is theoretically crucial because it shows that KiP does not seek a single explanatory principle, a law of intuitive knowledge, or of naïve physics, but describes a plurality of local phenomenological resources, each with its own experiential origin, its own context of activation, and its own possible role in the development of scientific knowledge (diSessa, 1993; diSessa et al., 2016).

3.2.2. Cognitive Mechanism

After describing the nature of the elements of intuitive knowledge, it is necessary to understand their mechanism of functioning and interaction, that is, the way in which these resources become operative in a given context. This reflection will allow us to examine more deeply the nature of p-prims themselves.

If in theory-like approaches the problem was to reconstruct the global structure that guides the entire reasoning process, here instead we do not have a unitary guiding theory (a set of rules, or a *production system*), but contextual activation. We therefore move from "possessed" knowledge to "activated" knowledge.

The first form of the cognitive mechanism that describes the activation of p-prims is *recognition*. P-prims act largely by being recognized. Recognition does not literally mean being seen. Rather, it means being cued to an active state on the basis of perceived configurations (diSessa, 1993). Empirical results, for example from the analysis of interview protocols, indicate complex, conflicting, and unreliable strands of reasoning in which students may be guided by aspects of the circumstances that they cannot articulate or check, given the difficulty of reinstantiating a particular mental state that accounted for a judgment. In other words, p-prims are not used through deliberate recall or logical inference. They "kick in" when a learner recognizes, often unconsciously, a situation as fitting one of these familiar explanatory sketches. Learning should provide that p-prims are activated in appropriate circumstances and, in turn, that they should help activate other elements according to the contexts they specify (diSessa, 1993). Thus, we can already see the first picture of the cognitive mechanism as a network of contextual activation of p-prims. Refining this first model requires a description of the local topology of this added-new recognition network. diSessa uses two key concepts: *cuing priority*, which is the "what

comes to mind first" and "what is activated from," and *reliability priority*, what sticks around and gets trusted.

Cuing priority depends on the «local topology of the recognition network» (diSessa, 1993), because an already active element, or a certain set of perceptual and interpretive features of the situation, can make the activation of another element more likely (diSessa, 1993). For example, a situation in which an object is pushed against a resistor can cue *Ohm's p-prim*; a situation in which a movement subsides can cue *dying away*; a configuration of opposing forces can cue *dynamic balance*. The activation of a certain p-prim is, in this sense, priority, because it becomes the main interpretive lens of that specific context. Within theory-like paradigms, that is, in rule-based causal models (i.e., *if something happens, then this should be the consequence*), or in symbolic models (keeping the comparison with AI), the context is "read before" the application of reasoning. Returning to the case of J discussed earlier, according to a model of naïve physics the subject might apply to the highest point of the trajectory the erroneous rule "IF velocity = 0, THEN acceleration = 0". This rule presupposes a description of the situation already structured according to the concepts of velocity and acceleration. It is not, however, automatic that the subject is able to describe the situation in these terms. diSessa emphasizes that interview protocols reveal a type of reasoning in which the subject attempts to explain the problem through categories more fragmentary than the very concept of "zero velocity," such as the idea that the motion seems to have been consumed, that the initial cause no longer seems active, and that the momentary absence of movement can be interpreted as the absence of dynamics. In this sense, in KiP the context is not read before reasoning, but is read directly through p-prims, that is, interpreted through the elements that make some features salient and not others. Thus, p-prims are not rules applied to an already structured input. They are part of the very way in which the input is structured. This is the strong theoretical presupposition for the mechanism of *cuing priority*.

Reliability priority, instead, concerns what, once activated, remains active, is maintained, and ends up guiding reasoning. A p-prim can be cued by a situation and then decay rapidly, because other resources weaken it, make it less plausible, or show its inadequacy with respect to the context. In this case, the p-prim in question would have high cuing priority but low reliability priority. Conversely, a p-prim can be activated and then stabilize, becoming the resource on which the subject continues to build the explanation. The distinction is crucial because it allows KiP to describe reasoning as a dynamic process, not as a simple association between situation and response, between input and output.

In addition, diSessa uses the notion of **structured priority**, identified as the conceptual dimension that brings together the previous two: «structured priorities refers to the pair, cuing and reliability» (diSessa, 1993). The term *structured* is understood as local and not global. Explicitly: «structured means priorities are not global; they do not provide a general ranking. Instead, priorities are structured according to context, the state of "neighboring" knowledge elements» (diSessa, 1993). Moreover, if on the one hand KiP does not describe intuitive knowledge as a global and coherent theory, on the other hand it does not describe it as a set of random fragments either. A *structured priority system* is a weakly organized system in which certain resources are activated and stabilized more easily in certain contexts, but without there being a universal hierarchy. Therefore, if naïve reasoning depends on local priorities, then learning does not simply mean adding correct rules, but reorganizing (priority shifting) the *structured priority system* (diSessa, 1993).

The problem of data-less prediction and bootstrapping. diSessa clarifies that p-prims are not constructs or entities directly observable in interview protocols, nor are they specified in advance – in a deductive way – by the theory (diSessa, 1993). From here arises the problem concerning the particular theoretical status of p-prims. If such elements depend strongly not only on context, but also on the phenomenological experience of the individual, in a sort of idiosyncrasy, then two important considerations follow. First, this serves once again to highlight the limit of theory-like approaches. Second, it tells us that KiP cannot produce many strong «data-less predictions» (diSessa, 1993) for identifying p-prims before empirical analysis.

This does not mean, however, that KiP is a theory without empirical grounding.

Methodologically, the core of KiP lies in s*ynthetic analysis*, that is, in the capacity to organize data, protocols, and observations within a theoretical framework capable of making otherwise opaque structures visible (diSessa, 1993). *Bootstrapping* is the methodologically grounded choice to assume the existence of p-prims and to verify their variety and stability through context, within the constructive process of the theory itself. It is a kind of non-vicious circularity between theoretical construction and empirical analysis. The researcher hypothesizes a p-prim (assuming its existence) in order to make an episode of reasoning intelligible and, at the same time, the p-prim gains theoretical stability insofar as it makes it possible to interpret other episodes, distinguish similar cases, explain hesitations, ambivalences, changes in response, and judgments of plausibility. Moreover, it is precisely

these hesitations, ambivalences, changes in response, judgments of plausibility, spontaneous explanations, satisfaction or confidence toward a description, that represent those which diSessa calls *heuristic principles for identifying p-prims*. These heuristic principles are the applied microanalytic method of bootstrapping.

It is necessary to clarify one further point. KiP is an operational theory, but not a merely pragmatic one. On the one hand, it provides criteria for recognizing elements in the data. On the other hand, these criteria are coherent with a precise epistemology of intuitive knowledge (an issue that will be discussed more fully in Section 3.4), understood as a local, fragile, contextual, and only partially systematic system. KiP does not position itself above the data. Rather, it places itself at the same level as the observed phenomenon and progressively constructs its own analytical objects from within the data themselves.

3.2.3. Systematicity and Developments

We have already said that KiP does not describe intuitive knowledge as a mere collection of random resources, but as organized in a *structured priority system*. This system describes the organization and, in a certain sense, the "probability" of activation and maintenance of certain cognitive resources at the local level, during a specific reasoning process. The level that diSessa calls *Systematicity* describes the systemic organization that exists among all the elements of the system. To specify more clearly, the structured priority system governs the flow of the specific resources activated during a given reasoning context, whereas *systematicities* describe the relational structure of the entire system independently of which resources are called upon. In a certain sense, systematicities represent the latent topology of the system, external to the reasoning context and ultimately emergent.

Among the examples of systematicities, there is *mutual use*, which explains why certain p-prims tend to be used together; *common attributes*, according to which different p-prims share the same attributes, for example agency, effort, balance; or, again, *completeness*, whereby certain configurations of resources give the subject the feeling that the explanation "closes" the problem.

Sense of Mechanism. We can use the topic of systematicities as a useful tool for clarifying the central point of KiP: the *sense of mechanism*. From what has been said, it is evident that it can be reframed as the *systemic* level of intuitive physics. The sense of mechanism does not coincide with individual p-prims, but with the way in which they are organized into a system. One could almost say that the sense of mechanism is an emergent property of the

system of intuitive knowledge, deriving from the regularities and forms of coordination among its resources. The emergence presented here is phenomenological-cognitive, not purely computational. This distinction is not only structural for the type of "subject" considered, human or artificial, but also relative to the type of learning that such a subject undergoes. The dimension of learning in KiP, described by diSessa in the section *Developments* (diSessa, 1993), will be addressed directly in the next section.

3.3. Learning as Tuning Toward Expertise

Within KiP it is possible to reformulate the notion of conceptual change introduced earlier. We have already intuited that within this new paradigm the idea that prior knowledge must be entirely replaced or modified is rejected. However, the trajectory point toward which naïve knowledge must tend remains fixed, namely scientifically accredited and structured knowledge. Within the KiP paradigm, learning is defined as a transformative process of *refinement* of intuitive knowledge (diSessa, 1993; Amin, 2025), understood as the reorganization of resources. A p-prim can generate an error of reading or interpretation in one context but can prove productive in another.

This reorganization of resources is what makes learning possible, understood as *tuning toward expertise* (diSessa, 1993). Under the umbrella of *tuning toward expertise*, diSessa describes learning through three specific mechanisms: *shifting priorities*, *change of function*, and *context migration*, graphically shown in Fig. 14. The first concerns the change in the priorities (of cuing and reliability) of p-prims. For example, some resources may initially be activated as priorities in a given context; in learning, their activation may be dampened. The second represents the change of function of p-prims. A p-prim has, as we have said, the characteristic of appearing self-evident and therefore strongly guides reasoning. In the transition toward expertise, some p-prims may gradually lose this autonomous explanatory force and become resources subordinated to denser configurations of knowledge. Finally, *context migration* concerns the change in the domain of application of a resource. A p-prim such as *dynamic balance*, initially activatable in many situations of apparent stability, is progressively restricted to contexts in which there is genuinely physical equilibrium, that is, zero resultant force or appropriate static conditions. The structured output of *tuning* is what diSessa calls *distributed encoding*, a system of resources organized by priority, function, and domain of application, integrating itself with the more complex dimensions of the concepts, laws, and models of accredited physical knowledge (cf. *distributed encoding* in Fig. 14).

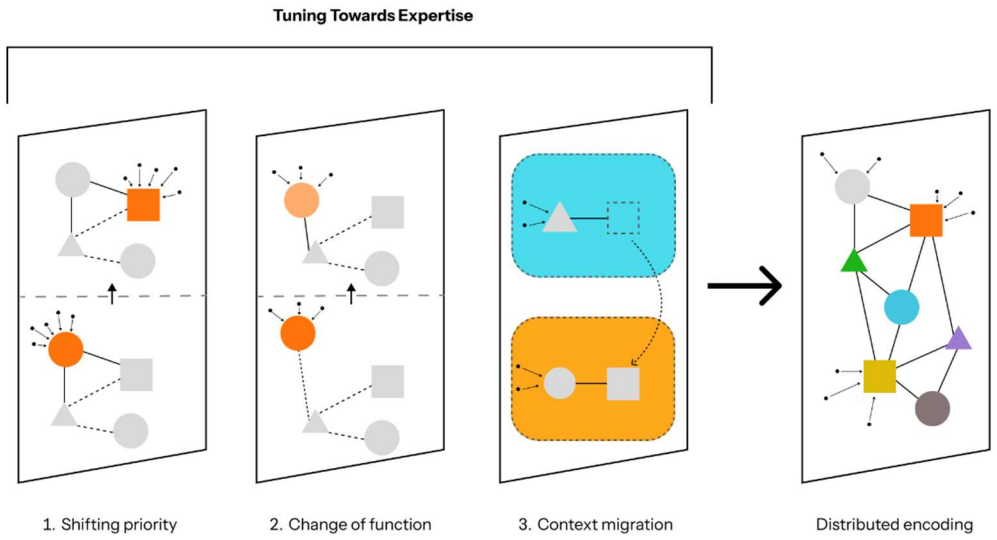


Figure 14: graphical sketch of *tuning towards expertise*. Four stages by which isolated p-prims - intuitive knowledge fragments, shown as large shapes (circles, squares, triangles, hexagons) - are reorganized into expert knowledge. Small dots mark contextual cues, arrows indicate cuing/local activation of p-prims, solid line indicate coordination between p-prims, dashed lines indicates weak coordination. Through shifting priority, change of function, and context migration, these elements culminate in a distributed, interconnected encoding.

3.3.1. Coordination Classes: From Tuning to Systematic Understanding

What does the **systematic understanding** of a scientific concept look like? (Amin, 2025). With this question we move one step further. So far, the model of p-prims – within the infrastructure of KiP – has allowed us to "explain how we explain" the occurrence of things in the physical world, that is, to explain how we make sense of the physical world, the *sense of mechanism*. P-prims are the elements of this *sense-making*, units of the system of intuitive knowledge which, when subjected to tuning, can reach expert knowledge (*distributed encoding*). Now, the question is how this expert knowledge becomes not only local, but *context-stable*; that is, how a subject comes to read the physical world and its concepts in a stable, reliable, and disciplinarily controlled way (and no longer only intuitively) across different contexts. It is for this purpose that the model of *coordination classes*, first introduced by diSessa in 1991, becomes relevant. This new theoretical construct must be read on two different but inseparable levels.

Coordination Classes as a class of concepts. The first level is that of the "concept of concept". This was the argument that diSessa and Sherin (1998) used to bring out the limit of theory-like approaches. In this case, *coordination classes* represent a model of what an accredited scientific concept is, that is, when a concept belongs to expert knowledge. Not all

concepts, in fact, perform the same function. For example, *category-like concepts* are distinguished from scientific ones on the basis of the task generally assigned to them: the prototypical task for category-like concepts is to determine whether something is or is not a member of the category, whereas the prototypical task for *coordination classes* is getting information from the world (diSessa, Sherin, 1998). As diSessa and Sherin (1998) observe, "seeing" things in the world means gaining information about them. In this perspective, a coordination class is not only an internal representation, for example p-prims, but a functional structure that allows the subject to identify, extract, and stabilize conceptual information across different circumstances. This inevitably refers to the second level, that of understanding the strategy to gain information and read the world.

Coordination Classes as a class of strategies. The second level – which better consolidates the definition of *coordination classes* – concerns the strategies for reading the world that expert knowledge develops as its own characteristic. diSessa and Sherin (1998) distinguish the concepts of *readout strategies* and *inferential net*. *Readout strategies* include the set of processes that allow an individual to focus attention and "read" relevant information from a specific situation (Amin, 2025; Levrini, & diSessa, 2008). In the more recent terminological revision, the literature distinguishes between *extraction*, understood as the initial perceptual phase, and *readout*, that refers to the whole chain of processing, from extraction to final determining of the concept-characteristic information (diSessa et al., 2016). Supporting this perceptual activity is the *inferential net*, originally defined as *causal net*, which represents the structural analogue of the entire readout process. It is the whole network of elements, p-prims, physical laws, reasoning strategies, that the individual possesses and that makes the chain of processing possible. To exemplify, whereas in novices the inferential net is dominated by p-prims, in experts it is centered on physical laws which, therefore, act as constraints for inference. The correct functioning of a coordination class is measured through two parameters: *span*, that is, the capacity to apply these strategies to a wide range of different contexts, and *alignment*, the requirement of coherence whereby different inferential pathways or observational strategies must lead to the same synthesis. These two components are not categorically separate, but co-evolve during learning.

In a certain sense, *coordination classes* do not sharply distance themselves from *tuning*, but the two models answer different questions. KiP and *tuning* explain which micro-elements compose knowledge and how they change in a learning process; *coordination classes* explain what specific kind of organization a knowledge that we call a "physical concept" must have

in order for the subject to possess that concept stably. *Coordination class theory*, therefore, does not provide a simple label for naming complex concepts, but an analytical criterion for verifying whether and how a concept functions as a coordinated system of access to information.

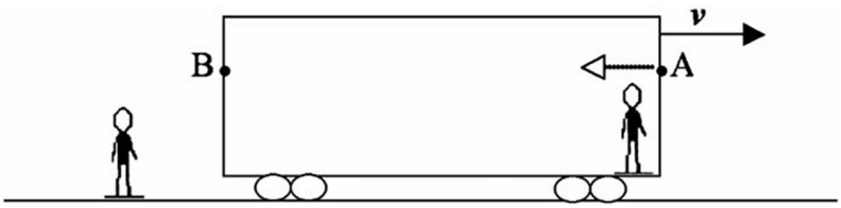


Figure 15: Sketch of the problem by Halliday et al. (1997) and used in the Levrini and diSessa case study (2008).

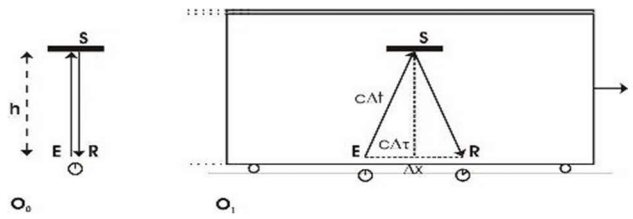


Figure 16: Light clock sketch for proper time classification.

Application scenario. An applicative example of the concept of *coordination classes* is provided by the case of *Proper time as a coordination class* (Levrini, diSessa, 2008). The study in question targets a class of twelve 18–19-year-old high school students who struggle with the difficult idea of "proper time" in special relativity. The specific focus is their difficulty in coordinating the different definitions of proper time in various contexts of use. The path grounding Levrini and diSessa's research can be read as the attempt to coordinate five elements: three contexts of use and two definitions of proper time.

Specifically:

- Context 1 (C1): the thought experiment of the light clock (Fig. 16), a context that the students already knew before Levrini and diSessa's experiment and with which they associate the classical definition of proper time (definition 1).
- Context 2 (C2): the problem from Halliday et al. (1997, see Fig. 15), proposed at the beginning of the reflective work with the target class.

- Context 3 (C3): the same problem from Halliday et al. but replacing the light ray with a small ball; this context is proposed during the reflective work with the students, after analyzing context 2.
- Definition 1 (D1): proper time as the time measured in the reference frame comoving with the object.
- Definition 2 (D2): proper time as the interval measured in the frame in which the two events considered occur at the same spatial position.

The problem initially proposed (C2) is taken from Halliday et al. (1997) and is formulated as follows:

«You are beside the tracks of a railway and a train passes in front of you at constant speed; inside the train, an experimenter sends a laser light pulse from the front toward the rear of a carriage, from point A in the direction of point B.

a. How does your measurement of the speed of the laser ray relate to the measurement made by the experimenter who sends the pulse?
b. For the experimenter, is the measurement of the travel time a proper time?

Justify your answers» (Halliday et al., 1997).

In addressing the problem from Halliday et al., the students draw on the resources they already possess, C1 and D1, which somehow represent a first form of coordination of knowledge. C1 and D1 are linked in an apparently natural way: the relevant object is the clock, the comoving frame is easily identifiable, and the interval to be measured coincides with the round-trip time of the light ray between emission and reception. Precisely this apparent naturalness, however, makes understanding fragile. The light clock allows students to obtain a correct answer without forcing them to make fully explicit what the general criterion is for identifying the relevant reference frame. In other words, the initial context produces a first *readout strategy*: finding the frame comoving with "something" and measuring time in that frame. But what remains ambiguous is precisely that "something."

C2, which constitutes the starting point of Levrini and diSessa's empirical analysis, puts this strategy into crisis. In the problem, the experimenter on the train sends a laser pulse from the front to the rear of the carriage (Fig. 15). In response to the first question, students must recognize that both the observer on the ground and the experimenter on the train measure the same speed of light, in accordance with the relativistic postulate. The second question, however, is more delicate: *is the travel time measured by the experimenter on the train a*

proper time? Most students initially answer yes, because the experimenter is in the same reference frame as the observed phenomenon. They transfer to the new context the strategy constructed in the light clock; that is, they look for the frame comoving with the object or with the phenomenon and assume that the train frame is the correct one, even though it is evidently not clear what this reference frame is, the carriage, the experimenter, the laser ray itself, the "gun" from which the laser ray is fired, and so on. The *readout strategy* is fragile because D1 (a description through objects, the light clock) implies the persistence of classical ontological inferences that the relativistic context requires to be *displaced*. In terms of *coordination class*, this first difficulty can be described as a problem of *span*: D1, and therefore also the first *readout strategy*, works in C1, but does not extend reliably to C2; it does not have sufficient *span*.

To address this difficulty, the second definition (D2) is introduced: proper time is the interval measured between two events that occur at the same spatial position (more compatible with relativistic inferences). This definition modifies the *readout strategy* required of students. It is no longer, first of all, a matter of looking for an object with respect to which being integral, but of identifying the two relevant events and asking whether there exists a reference frame in which they occur at the same point. In Halliday's problem, the two events are the emission of the laser ray from the front of the carriage and its reception at the rear. In the train frame, these two events do not occur at the same point and, therefore, the time measured by the experimenter on the train is not a proper time (for the interval considered). However, even this new definition is not immediately coordinated by the students, because they continue to look for a frame in which proper time can exist. D2 increases explanatory precision but does not automatically produce a stable *coordination class*: it too has a problem of *span* with respect to C2, because in C2 it is not possible to identify the required reference frame.

At this point the third context (C3) is introduced: instead of the light ray, one considers a particle or a ball that moves from the front to the rear of the carriage. In this case, the new definition becomes more operational. The students are able to imagine "riding" the particle and, therefore, to verify that in the frame integral with it the initial event and the final event occur at the same point in spacetime. The readout strategy provided by C3 works because it allows D2 to be translated in a comprehensible way. However, returning to the light ray produces a new difficulty, because it is not clear what it means to "ride" a light ray. Here the problem is no longer only one of *span*, but becomes a problem of *alignment*, since the different projections of the concept are not yet fully coherent.

The discussion is resolved through the interpretation of the light ray as a limiting case. From an Einsteinian perspective, there is no physical observer integral with a light ray; that is, it is not possible to "ride" light because no inertial frame can move at the speed of light. From a Minkowskian perspective, instead, the light path can be interpreted geometrically as a light-like interval, for which proper time is zero. The students thus come to distinguish between two statements that initially seemed incompatible: on the one hand, there is no physical reference frame in which to directly measure the proper time of the light ray; on the other, the geometric formalism allows one to attribute to the light path a zero proper interval as a limiting case. This distinction does not eliminate the difficulty but makes visible the work of conceptual coordination required.

The outcome of the research. The theoretical value of the episode lies in showing how the construction of the concept of proper time requires a progressive articulation of *readout strategies*, inferences, and contexts. Definition 1, linked to the frame comoving with the object, works in the context of the light clock but risks maintaining a reading of the phenomenon that is still too classical-Newtonian. Definition 2, grounded in the co-localization of events, shifts attention toward more properly relativistic reading, but must be made operational through different cases. The context of the particle makes it possible to build a bridge between the two definitions, while the limiting case of the light ray forces clarification of the difference between operational measurability and geometric determination. In this sense, the case of proper time shows that a *coordination class* is not constructed simply by adding a more precise definition, but by coordinating (through *span* and *alignment*) multiple projections of the same concept. Levrini and diSessa interpret this process as a form of *articulate alignment* (Levrini, diSessa, 2008). Students become progressively capable of seeing and identifying the relevant events, choosing the appropriate representation, distinguishing ordinary cases from limiting cases, and recognizing which strategy to use.

The case analyzed shows not only how the concept of proper time can be a coordination class to the point that students are able to appropriate it and use it in an informed and stable way, but also that the coordination of different elements of knowledge necessarily implies, as a first step, a displacement of attentional focus (i.e. *noticing*, *focus displacement*). This can be considered the key to the integral experience of scientific learning.

3.3.2. All you need is Attention (or Noticing)

With *coordination class theory* – and more generally with Knowledge in Pieces – we have reconstructed a theoretical framework in which it is possible to describe the transition from naïve knowledge to expert knowledge no longer only as a reorganization of resources, but also as a class of strategies that define physical concepts as stable concepts within the mind of an expert. In this perspective, expert knowledge is able – as a first step – to read the relevant aspects from a context and apply readout strategies, that is, ways of focusing attention and reading the context. This step, which lies at the beginning of the process of analyzing a problem, helps us understand that learning is not only the evolution of knowledge with respect to disciplinary content, but also the transition from naïve salience to disciplinary salience.

Attention, the capacity to "see," capacity of *noticing*, becomes the central point of learning. It is reasonable to affirm that the core problem of conceptual change is shifting the means of seeing (diSessa, Sherin, 1998). This shift of attention also involves a shift in epistemological posture, that is, in the way of understanding scientific knowledge, in this case relativity. In the case study by Levrini and diSessa (2008), the research concludes with the proposal of an ontological restructuring that involves moving from a view of the cosmos as a place populated by self-moving objects and phenomena to an interpretation of the universe as a set of spacetime events. In terms of *coordination classes*, this involves adopting a preferential focus of attention for all relevant determinations, elevating coordinate-defined event-points to the privileged site for *readout*. This approach makes it possible systematically to problematize the naïve notion of "phenomenon" – understood as an occurrence with absolute position and duration – replacing it with a legitimized phenomenology that considers objects as families of events deposited in spacetime. In summary, this new posture guarantees the *alignment* of inferential pathways and facilitates the *displacement* of persistent classical intuitions, allowing the knowledge system to operate with the coherence and stability required by relativistic physics.

In Chapter 4, we will see how the theme of attention can serve as a criterion of comparison between human and artificial learning.

3.4. The revolution of KiP

The revolution introduced by *Knowledge in Pieces* in the field of the learning sciences does not consist only in having proposed a different description of intuitive knowledge, or in

having proposed a further cognitive model, but in having modified the very way in which it is possible to think about the relation between knowledge, learning, and theoretical modeling.

To summarize, the first contribution of KiP concerns the status of intuitive knowledge. In this perspective, students' ideas are not interpreted primarily as errors to be eliminated, but as cognitive resources that can prove productive or unproductive depending on the contexts of activation. The role of error is also reinterpreted, since a non-normative answer can be the effect of a reasonable resource activated in an inappropriate context, or not yet coordinated with other resources necessary for deep scientific understanding. Common-sense knowledge therefore does not appear as a monolithic obstacle to be replaced, but as a set of *pieces of knowledge*, p-prims, endowed with its own intelligibility. This shift is decisive because it makes it possible to overcome the deficit model of scientific learning, understood again as tuning toward expertise. The term pieces does not indicate a random dispersion of resources, but a different scale of analysis to which theoretical modeling must necessarily refer.

The scope of KiP does not end here. Its revolution also concerns the very way of understanding the construction of a theory of knowledge and learning. Indeed, KiP can be read as an *epistemological approach to the construction of a learning theory*, not so much because it is a unitary theory of scientific knowledge or a comprehensive philosophy of science. Rather, it shows how a theory can be constructed when the phenomenon under study is complex, situated, and processual – in the concrete context of research on scientific learning. Three aspects help us understand KiP as an epistemology of theory.

The first aspect concerns the ontological dimension of KiP, that is, the status of p-prims. They are not immediately observable objects, nor categories deduced a priori. On the contrary, they are theoretical constructs hypothesized in order to make local regularities in students' reasoning intelligible. In this sense, their genesis remains partly assumed through the process of bootstrapping (see Section 3.2.2) – the researcher does not directly observe a p-prim – while their characterization is progressively inferred from the way in which a certain situation, i.e. a physical problem, is recognized, explained, and made intuitively plausible by the student. Moreover, p-prims represent an example of what diSessa and Cobb (2004) call *ontological innovation*. With this expression, the authors indicate the construction of new theoretical categories that make it possible to identify regularities and structures in phenomena which, without such categories, would appear too heterogeneous or

fragmented to be described systematically. P-prims, in fact, make visible a scale of organization of intuitive knowledge that does not spontaneously emerge from the observation of data except through this interpretive process (diSessa, 1993; diSessa & Cobb, 2004).

The second aspect concerns the methodological consequence that KiP assumes for orienting research. It does not function as an aprioristic taxonomy to be applied mechanically to data, but as an *orienting framework*. This notion is introduced by diSessa and Cobb (2004) to describe theoretical frameworks that, by their structure, indicate general perspectives for conceptualizing learning, teaching, and theory design. KiP is more than this. One could call it a *generative* orienting framework, since from it different theoretical models of knowledge can be developed, such as *p-prims*, *coordination classes*, and *strategy systems* (diSessa & Levin, 2021). KiP, therefore, does not predetermine the focus of research in a closed way, but provides a clear analytical and methodological direction. About *methods*, theoretical work proceeds through a controlled loop between constructs and data. The theory suggests what to look at and at what scale to analyze the phenomenon (the microstructure of reasoning, the micro-temporality of change); the data, in turn, force the constructs to be specified, corrected, or refined (diSessa, 1993; diSessa & Levin, 2021). The case study by Levrini and diSessa (2008) is a clear example of this. *Coordination class theory* is applied, and the result is that the concept of proper time can be understood as a coordination class insofar as it requires readout strategies, an inferential network, span, and alignment. However, the analysis of the case also produces a natural extension of the theory itself, since it makes visible the role of definitions, contexts, and articulate alignment in the construction of a more stable understanding. The theory, therefore, does not remain identical to itself after encountering the data, but is refined.

In summary – the third aspect – KiP descends from the "Olympus of the a priori" and moves toward the data (i.e. *humble theory*; diSessa & Cobb, 2004; diSessa, Sherin & Levin, 2016; Levrini & diSessa, 2008), close enough to grasp their fine grain. It is an epistemology of theory because it constructs itself from within the phenomenon it seeks to study.

3.5. Limitations and further work

If KiP defines itself as humble theory, this designation also takes into account its limits, discussed synthetically below.

Pieces or coherence. Critics such as Stella Vosniadou argue that KiP, by focusing excessively on the granular nature of knowledge, the p-prims, does not adequately explain the presence of well-defined and coherent synthetic mental models observed in children, such as those concerning the shape of the Earth. Whereas KiP sees systematicity as an emergent and often local pattern, supporters of *theory-like* approaches maintain that the search for coherence is an initial condition of the cognitive system and an empirically observable need of the subject. Vosniadou suggests that the fragmentation reported by KiP studies may, in some cases, be a methodological artifact due to questions that fracture the internal consistency of students' answers (Vosniadou, 1994; Vosniadou, 2018; Vosniadou & Brewer, 1992). Amin (2025) recognizes the validity of Vosniadou's critique, but argues that this difficulty does not necessarily imply a return to a theory-like conception of intuitive knowledge. KiP in fact possesses internal theoretical resources, such as the model of coordination classes and the forms of systematicities present in p-prims, to explain how local elements can produce relatively stable patterns of inference.

Prediction. An intrinsic limitation acknowledged by diSessa himself concerns the predictive capacity of the model. KiP allows one to describe in very fine detail the role of individual resources in students' reasoning, but it does not always allow one to predict in advance which p-prims will be activated in a given context or what trajectory the learning process will take. This reconnects to *bootstrapping*, that is, p-prims are not directly observable, nor are they specified once and for all before empirical analysis. At first sight, this might suggest that the identification of p-prims depends on an arbitrary process or on the researcher's simple intuition. However, as discussed in Section 3.4, KiP is grounded in a precise analytical methodology. Research proceeds through synthetic analysis, which continuously compares theoretical constructs and empirical data. It is therefore difficult to formulate *data-less* predictions, but there is a method of validation consistent with the theory. Moreover, the question of the ultimate origins of p-prims remains open; there is currently no direct observation of the moment in which these atoms of knowledge first form in childhood (diSessa, 1993).

Timescales. KiP excels in the fine-grained analysis of episodes of reasoning lasting seconds or minutes (i.e. microanalytic and microgenetic scales), but struggles to provide exhaustive accounts of long-term changes spanning years of development, the temporal dimension typical of the Framework Theory approach (Vosniadou, 1994). However, computational techniques that can capture statistical changes in patterns of activation of knowledge

elements in corpora of student reasoning are a promising methodological innovation that could help link analyses at different timescales while adopting the KiP perspective (Sherin, 2013).

In "fraught" with language. The relation between KiP and language is tense. On the one hand, KiP research focuses on tacit, intuitive, and not necessarily propositional knowledge, often difficult to verbalize. On the other hand, the main data on which it is based are almost exclusively verbal protocols. This creates a methodological tension: researchers must infer mental structures that do not have a direct linguistic representation precisely from the nuances of the language used by students. This tension does not cancel the value of the model, but it requires caution. Language must be treated as a partial, situated, and interpretively mediated trace of the activation of resources (Amin, 2025).

Pedagogical challenges. The complexity and scale dimension of KiP represent a non-negligible limitation with regard to curricular and pedagogical implications, which remain predominantly local and not long-term (Amin, 2025).

The choice to take the KiP perspective into consideration as the theoretical framework of reference does not depend only on its innovative components with respect to previous models, but on the fact that it presents a structure that is, in several respects, comparable with that of contemporary AI. The comparison developed in the following chapter, which constitutes the original contribution of this work, arises precisely from this convergence. It is not a matter of immediately and uncritically superimposing human knowledge and artificial knowledge, but of understanding how – through this comparison – it is possible, on the one hand, to interrogate the possible extensions of the KiP paradigm where it itself leaves open questions and, on the other, to explore what KiP allows us to see in contemporary AI and, at the same time, to clarify the differences between the models of knowledge under consideration.

Chapter 4

The objective of this chapter is to compare the models considered so far, LLMs and KiP. Specifically, there is no intention to argue that human learning and artificial training can be traced back to the same mechanism, nor to propose a direct analogy between the human mind and the Transformer. Such an approach would be theoretically fallacious, because it would result in assimilating processes that have different ontological, material, and phenomenological statuses. KiP describes forms of human knowledge, situated, experiential, and oriented toward the construction of scientific understanding; LLMs are computational systems trained on enormous textual corpora through statistical optimization procedures. This distinction represents an inevitable starting condition, but it does not prevent comparison.

The specific research question with respect to which such a comparison becomes viable and theoretically interesting is to understand whether and how the study of LLMs makes it possible to render more visible certain elements and dimensions of KiP, and vice versa. From this perspective, LLMs can be understood as an epistemological laboratory, that is, reconceptualized as a theoretical and cultural "object" that allows us to interrogate, in a new way, questions concerning the nature of knowledge and learning. Consequently, the specific methodology to be applied in order to answer this research question is outlined by three main criteria. First, a criterion of granularity, according to which the comparison is constructed not between "mind" and "machine" in general, but between specific theoretical components of the two reference frameworks. Second, a functional criterion, according to which the analogy concerns the role played by an element within the system, not its ontological nature. Third, an epistemological criterion, according to which every similarity is evaluated in light of the different constraints that define knowledge, learning, and understanding in the two cases. In this way, on the one hand the level of similarity makes the comparison intelligible, while on the other hand the level of differences avoids possible reductionist approaches.

4.1. A bidirectional comparison between LLMs and KiP

In this section we seek to answer the first of the two research questions (RQ1): *Is it possible to construct a synthetic framework in which Large Language Models and Knowledge in Pieces illuminate one another as models of complex and non-symbolic knowledge systems?*

As we have seen so far, this question is articulated into two symmetrical and complementary sub-questions:

> **Sub-RQ1a)** *how the vocabulary of Knowledge Analysis may allow us to map and make intelligible the trajectories of generation and the sub-symbolic learning of Transformers, dimensions that remain opaque to date?*
>
> **Sub-RQ1b)** *to what extent LLMs as an object may serve as a new access route for analyzing and better understanding the theoretical framework of KiP itself?*

The possibility of constructing this comparison and answering RQ1 depends on the path outlined in the previous chapters, starting from the general comparison concerning the form of knowledge that the two frameworks considered model or "produce."

4.1.1. The form of knowledge

Chapter 1 showed how the birth of AI was made possible by a broad interdisciplinary development in which cybernetics, information theory, and computability theory made it legitimate to describe "intelligent behavior" in terms of information, control, manipulation of symbols, and procedures (Turing, 1950; McCarthy et al., 2006). This approach found its most explicit formulation in Newell and Simon's *physical symbol system hypothesis* (1976), according to which general intelligent action requires a system capable of manipulating symbols, producing symbolic structures, and orienting search through heuristics. Models such as Logic Theorist, LISP, SHRDLU, and SOAR progressively articulated this intuition, showing both its effectiveness and its limits; the case of SHRDLU (Winograd, 1971) is paradigmatic in this sense, because its success depended on the restriction of the domain and on the possibility of providing the system with a procedural semantics internal to a circumscribed artificial world, a success that made visible, by contrast, the problem of the semantic scalability of explicit symbolic knowledge.

The transition to Machine Learning and then to Deep Learning radically modified this framework. Knowledge is no longer inserted a priori in the form of rules or axioms, but emerges from training on data, from the updating of the weights of deep networks (multi-layer), and from the progressive optimization of the system. LLMs represent the current point of this trajectory, a passage defined as from the symbolic to the sub-symbolic. LLMs are not programs that apply explicit linguistic rules, but neural networks trained to predict the next token in a large textual corpus (Jurafsky & Martin, 2026; Lewis, 2025), and their

capacity does not derive from a dictionary of encoded grammatical rules, but from a distribution of weights that allows them to generate statistically and semantically plausible linguistic continuations.

Knowledge in Pieces, on a different level, calls into question an analogous conception of knowledge – the one that sees intuitive knowledge as organized from the beginning as a coherent and propositional theory, constituted by a set of formal rules that the mind uses to process information. KiP, within the methodological perspective of Knowledge Analysis, arises from the need to describe intuitive knowledge at a finer grain than theory-like models, hypothesizing a mind constituted by elements or pieces of knowledge, intuitive resources that can be activated, coordinated, transformed, and stabilized in different ways according to context (diSessa, 1993; diSessa, Sherin, & Levin, 2016).

KiP and LLMs, through independent paths, therefore force us to rethink knowledge beyond the paradigm of the explicit rule. On the one hand, KiP shows that human knowledge can be fragmentary without being random; on the other hand, LLMs show that linguistic performance can be complex without being grounded in directly readable symbolic structures. Since both frameworks model granular forms of knowledge, although of profoundly different nature, the comparison that follows must constantly maintain this fine-grained scale of analysis, avoiding a premature return to more global categories that neither KiP nor the epistemology of LLMs authorizes.

4.1.2. Elements

The first level of comparison between KiP and LLMs concerns the nature of the elements that compose knowledge in the two frameworks. In KiP, p-prims are elements of intuitive knowledge derived from the schematization of physical and phenomenological experience, generally perceived by the subject as self-evident (diSessa, 1993). They are neither scientific propositions nor misconceptions in the classical sense of the term, but theoretical constructs inferred by the researcher in order to make recurring patterns in the subject's reasoning intelligible. Their identification requires an interpretive methodology grounded in bootstrapping and synthetic analysis, rather than a procedure of direct detection (diSessa, 1993).

The case of the interview with "J," discussed by diSessa and Sherin (1998), illustrates well the nature of these elements. Asked about the forces acting on a ball thrown upward, J initially provides a correct explanation, recognizing gravity as the only force at play along

the entire trajectory; as soon as the interviewer shifts attention to the highest point of the throw – performing what is defined as focus displacement, or noticing – her reasoning changes drastically, and J introduces a second force, initially impressed by the hand, which would oppose gravity until it is completely exhausted at the top of the trajectory (Sherin, 2021). This oscillation should not be interpreted as simple incoherence, but as the effect of the contextual activation of different resources (diSessa, 1993), solicited by different aspects of the situation, not by the possession of two complete and alternative theories. The term pieces therefore does not indicate random dispersion, but a determinate scale of description; the elements of intuitive knowledge are local but not isolated in an absolute sense, they can be activated together, reinforce one another, enter into tension, change function, and migrate into new contexts, properties on which their productivity depends, varying from context to context.

In LLMs the framework is radically different in nature, but it raises a formally related problem. The initial unit of the system is not the concept but the token. Text is segmented into discrete units, transformed into numerical indices, and projected into a vector space through embeddings. The initial embedding, however, is not sufficient to determine the contextual meaning of the token, because this meaning emerges from the relations that the token maintains with the other elements of the sequence: the Transformer architecture, through self-attention and multi-head attention, produces representations that are progressively reconstructed as a function of the overall configuration (Vaswani et al., 2017; Jurafsky & Martin, 2026; Lewis, 2025).

P-prims are interpretive constructs rooted in physical and phenomenological experience; tokens and embeddings are computational entities internal to an artificial system, trained by error minimization on a textual corpus, without any sensorimotor correlate. The similarity between the two cases therefore does not concern the ontological status of the elements, which is evidently different, but the fact that in both cases the final output emerges from local and distributed configurations, not from an isolated element endowed with its own meaning. In this sense, LLMs help to make less counterintuitive an aspect of KiP that, when applied only to human cognition, sometimes risks appearing fragile: a structure can produce coherence without being organized in advance in a coherent way. At the same time, KiP makes it possible to specify that obtaining coherent performances is not sufficient, in itself, to establish that the artificial system possesses knowledge in the same sense in which a human subject possesses it.

4.1.3. Attention and noticing

The constitutive elements of knowledge, in both frameworks, are not sufficient by themselves to explain knowledge and learning. It is necessary to specify how and why certain elements become operative in one context and not in another. In KiP, this problem is addressed through the notions of *cuing priority* (what is activated first) and *reliability priority* (what, once activated, persists and ends up guiding reasoning). A p-prim can be cued by a situation and then decay rapidly, because other resources weaken it or show its inadequacy with respect to the context. In this case, such a p-prim would have high cuing priority but low reliability priority (diSessa, 1993). The two notions together constitute structured priority, a weakly organized, local and non-global system, in which certain resources activate and stabilize more easily in certain contexts, without there being a universal hierarchy (diSessa, 1993).

Returning to the case of J, the focus displacement performed by the interviewer does not add new physical information but places the subject in the position of modifying her "gaze." This occurs through p-prims, that is, through the very elements that make certain aspects of the situation salient and not others. This is the technical sense – distinct from, and to be preferred here over, the more generic term "attention" – that in KiP is defined as *noticing*, the situated capacity to select which aspects of a situation are epistemically relevant for the reasoning underway.

The comparison with the *attention* of Transformers is at this point direct. Through query, key, and value, the model calculates which tokens are more relevant than others in a given sequential configuration; multi-head attention allows it to capture simultaneously different types of dependencies in different representational subspaces (Vaswani et al., 2017; Jurafsky & Martin, 2026). Attention is, in this sense, a device for selecting relevance internal to the model, a mechanism that structurally performs a function in some respects analogous to noticing. The comparison, however, requires caution, and here the reference left suspended at the end of Section 3.3.2 is explicitly resolved. The attention of Transformers is not phenomenological and embodied attention as noticing is, but a computational mechanism devoid of intentionality and consciousness. This distinction, however, should not lead to a devaluation of LLMs, a scientifically inadequate polarization for the path constructed here, but rather to recognizing that it is precisely this point of tension that makes the comparison epistemologically productive. Both frameworks, in fact, while working in completely

different domains, have had to introduce a mechanism dedicated to “what counts in context” as an indispensable ingredient of knowledge and learning.

4.1.4. Language

The revolution introduced by LLMs, with their emergent capacities, arises from the apparently elementary task of predicting the next word (Jurafsky & Martin, 2026; Lewis, 2025). Precisely this disproportion between the simplicity of the task and the complexity of the performances obtained makes language epistemologically relevant, with a scientific relevance perhaps deeper than it had previously. But how is it possible that LLMs can generate coherent texts, plausible explanations, and linguistic generalizations if they do not directly encounter the world, do not experience phenomena, and do not construct knowledge through an embodied relation with reality? This is possible precisely through language. In it, traces of human experience are encoded in the form of semantic regularities, descriptive interpretations, and recurring intuitive schemas. The attention of LLMs operates precisely on these traces. The model arrives at what the literature discussed in Chapter 2 calls a “world model” (LeCun, 2022) through the traces of meaning present in language itself – a performance that derives neither from an embodied encounter with reality, nor from a purely superficial manipulation of signs devoid of structure (Bialek, 2025; Cristianini, 2024).

This fact produces a twofold insight into KiP. The first does not consist simply in observing that p-prims find expression in language – a circumstance already implicit in diSessa’s methodology, which reconstructs such resources starting from verbal, gestural, and interactional protocols. The new point instead concerns the scale of inquiry. In KiP, language is first of all a local trace of the activation of resources in a specific episode of reasoning; LLMs, by contrast, show that language can also be treated as a distributed field of large-scale regularities. Their capacity to produce fluent and coherent texts indicates that, in collective linguistic practices, there exist sufficiently stable patterns that can be modeled and reused generatively. It is not a matter here of claiming that LLMs activate p-prims, nor that p-prims are simply linguistic structures; on the contrary, it is a matter of making visible a methodological possibility, namely whether LLMs can contribute to an evolution of Knowledge Analysis toward the large-scale identification and validation of linguistic patterns compatible with p-prims, a possibility that Section 4.2 will directly take up.

Second, precisely because LLMs work exclusively on language, they make more visible by contrast what language does not exhaust, namely experience itself, the "doing" of experience. This opens directly onto the next section.

4.1.5. Learning: training and tuning

A further level of comparison is the dynamic of learning. In LLMs, learning occurs mainly during pre-training, when the model is trained on large quantities of text to predict the next token; the target is already present in the corpus, and the system updates its internal weights by minimizing error through cross-entropy and backpropagation, in a self-supervised process (Jurafsky & Martin, 2026). To this are added subsequent phases – fine-tuning, instruction tuning, preference alignment – oriented toward making the model more useful, safe, and aligned with human preferences (Jurafsky & Martin, 2026; Lewis, 2025). This type of learning produces a model of the corpus, not a recording of the corpus (Lewis, 2025); that is, the system does not merely retrieve memorized sentences but constructs a predictive structure applicable to prompts never seen before. LLMs learn, in this sense, to produce plausible continuations of text, but they do not learn to experience the world in the sense in which a human subject does.

In KiP, learning has a different status. Tuning toward expertise does not consist primarily in adding information to a cognitive system, but in reorganizing a system of resources already present, through shifting priorities, change of function, and context migration, up to the progressive construction of more stable forms of coordination (diSessa, 1993; diSessa, Sherin, & Levin, 2016). Learning is not the replacement of error, but the transformation of the functioning of the system. This transformation has a precise experiential dimension, the central thesis of the chapter: *to learn*, in the strong sense, *means to undergo embodied experience*. This is not meant in the reductive sense whereby every form of knowledge would derive immediately from elementary sensory perception, but in the sense that the subject's resources are – in the first analysis – rooted in forms of bodily, perceptual, practical, and social experience (Harnad, 1990, 2002; Dreyfus, 2006; Gahrn-Andersen, 2025). P-prims themselves arise from the schematization of physical experience; even when the subject works on abstract contents, he does so starting from a history of interactions with the world, with his own body, with tools, and with other subjects.

The case of proper time discussed by Levrini and diSessa (2008) decisively clarifies this point. The difficulty of relativity does not consist in memorizing a correct definition of

proper time, but requires a restructuring of the way in which the student reads the phenomenon, that is, the passage from a conception of the world as a set of objects placed in an almost absolute spacetime to an interpretation of the universe as a set of spacetime events implies a change in readout strategy, in which the student must learn to see events (not objects) as the privileged site of physical determination. In KiP terms, this means modifying salience, reorganizing priorities, and constructing new forms of alignment.

An LLM can describe this passage, correctly distinguish coordinated time and proper time, and construct effective didactic examples; but its performance remains a linguistic production constrained by regularities learned in the data, without there being, in the model, a transformation of the way of inhabiting the phenomenon. The comparison therefore does not serve to deny the epistemological value of LLMs, but to distinguish two forms of change. In the case of LLMs, training produces statistical dispositions toward generation; in the case of human learning described by KiP, tuning produces a transformation of the readout, of salience, and of the coordination through which the subject enters into an embodied relation with the phenomenon.

4.1.6. Systematicity

The fourth level of comparison concerns systematicity. One of the most relevant criticisms addressed to KiP is that a theory founded on local elements risks not adequately explaining the coherence and stability of knowledge. Vosniadou's framework theory, while recognizing the presence of fragmented and synthetic conceptions, attributes to initial knowledge a more global structure, endowed with internal coherence and resistant to change (Vosniadou, 1994, 2018; Vosniadou & Brewer, 1992; Sherin, 2021). KiP responds to this criticism not by denying systematicity, but by redefining its status. Systematicity is not assumed from the start as the global form of intuitive knowledge, but is an emergent pattern, to be empirically reconstructed through recurrences, priorities, local relations, and coordinations. diSessa himself (1993) recognizes that no initial assumption about systematicity is incorporated into the theory, and that the available empirical methods are more suited to identifying individual elements than systematic relations among them.

Coordination classes represent the core of this response. A coordination class is not simply a complex concept, but a system of strategies that makes it possible to extract relevant information from a variety of contexts and to coordinate it through an inferential network (diSessa & Sherin, 1998). Span and alignment are its two constitutive parameters. Span

concerns the breadth of contexts in which a coordination class functions effectively; alignment concerns the coherence of the determinations produced in different contexts. The case of proper time, already discussed in 4.1.5, is its most consolidated example. The passage from D1 (the definition through the light clock, of limited span) to D2 (the definition through coincident events in spacetime, with greater span and alignment) shows that expert understanding does not amount to an accumulation of information, but to a gain in stability in the disciplinary reading of the world (Levrini & diSessa, 2008).

In LLMs, systematicity has a different but, in some respects, comparable meaning. Textual coherence emerges from distributed representations, from statistical regularities learned in corpora, from contextual configurations, and from generative constraints, without any of these components, taken in isolation, "containing" the coherence of the final text, and without the system possessing an explicit theoretical structure that guarantees it a priori.

KiP here makes it possible to interpret the coherence of LLMs without reducing it to statistical randomness and without mistaking it for proof of understanding. LLMs, in turn, make more plausible (because empirically observable in a system different from the human one), as a convergence of indications, the KiP idea according to which non-symbolic systems can produce forms of coherence without a global theoretical structure assumed from the start.

This clarification on systematicity allows us to specify the role of error in the two frameworks.

4.1.7. The role of error

In both models, error is not an anomaly to be eliminated, but a route of access to the functioning of the system.

In KiP, error is neither absence of knowledge nor a misconception, in the classical sense in which it is interpreted (an erroneous belief). It is the effect of a reasonable resource activated in an inappropriate context or not yet coordinated with the other necessary resources (diSessa, 1993). The case of J illustrates precisely this mechanism. Error reveals which elements were plausibly activated, which priorities governed the situation, which readouts were selected, and which forms of coordination are still unstable (diSessa, 1993; Sherin, 2006).

In LLMs, *hallucinations* have a radically different status. They are linguistically plausible but factually false or unfounded productions, which do not reveal the activation of a poorly

coordinated experiential resource, since in the model there are no experiential resources in this sense. Hallucinations represent a structural and intrinsic feature of the nature of linguistic generation in Transformer systems. As highlighted by Bender et al. (2021), these models are trained for string prediction tasks, operating a manipulation of statistical patterns. The system generates continuations based on the probability distribution learned during training without stable anchoring to the world (*grounding*), necessary to reflect the causal relations of physical reality (i.e., LLMs lack *world models*, see LeCun, 2022). In this sense, hallucinations derive from the intrinsic incompleteness of the "prior" (Hila, 2025). This refers to the fact that the training corpus, however large, remains a finite and partial sample of reality, without the possibility of equaling its complexity. The error of the LLM marks the breaking point at which the distance emerges between the generative criterion of sequential plausibility and the epistemic criterion of knowledge reliability (reliabilism; Hila, 2025), the latter being necessary to address the risk of *the illusion of knowledge*, or *epistemia* (see Section 2.5).

To summarize, in KiP, error is a window onto the coordination of resources; in LLMs, error is a window onto the distance between linguistic plausibility and epistemic reliability. This difference prevents us from assimilating generative performance to conceptual understanding, but it also prevents us from assigning LLMs to "empty" systems devoid of scientific relevance. As shown in 4.1.6, plausibility is not randomness and the fact that a model can hallucinate does not imply that its coherence lacks structure, just as the fact that a student produces non-normative explanations does not imply that his knowledge is absent or completely erroneous.

4.1.8. Experience, grounding, and understanding

Throughout this chapter, we have repeatedly introduced the necessary distinction between artificial and natural intelligence around the problem of experience. It is necessary to bring this discourse to completion in this dedicated section.

In KiP, intuitive resources are rooted in physical and bodily experience, and the self-evident character of p-prims reveals this anchoring (diSessa, 1993; diSessa, 1983). Learning, within this framework, does not mean acquiring new information but transforming the way in which the subject reads the phenomenon – a transformation that presupposes a subject who acts in the world, manipulates it, receives responses, and is surprised. In LLMs, the relation with the world is entirely mediated by data and representations. There is no manipulation of

objects, there is no perception of resistance, there is no surprise in the human sense of the term. Even multimodal models that receive images, sounds, or sensory data as input, entertain with these representations a computational, not experiential, relation. The distinction, therefore, is not between sensory and linguistic modes of experience, but between the processing of representations and the embodied encounter with the world (Bisk et al., 2020; Gahrn-Andersen, 2025; Mollo & Millière, 2023).

Thus, the learning of LLMs does not coincide with the experiential transformation that KiP describes for human cognition, and this limit is not technical but structural. It is precisely the comparison with KiP that allows us to specify the distinction between performance and understanding. Understanding, in the sense that concerns the *learning sciences*, is not the capacity to produce correct outputs on a domain, but the construction of stable and disciplinarily reliable ways of reading the world (diSessa & Sherin, 1998; Amin, 2025). In this sense, KiP – through the concepts of p-prims, tuning, coordination classes, readout strategies, span, and alignment – offers a theoretical vocabulary for explaining how to reread the framework of LLMs, rethinking knowledge beyond the model of the explicit rule and also making clearer the difference between generative performance and understanding.

4.1.9. Comprehensive overview

The eight categories traversed in the previous sections now allow us to answer directly the two sub-questions of RQ1, gathering the work carried out. With respect to **Sub-RQ1a** – *what the vocabulary of Knowledge Analysis makes intelligible about the sub-symbolic trajectory of Transformers* – the balance is as follows. KiP offers a conceptual apparatus for describing LLMs as granular systems in which the minimal unit is not the concept but the token, devoid of isolated meaning (4.1.2); activated by contextual configuration rather than by rule execution, through the mechanism of attention that structurally performs a function analogous to cuing, although without intentionality (4.1.3); generative in an autoregressive way, that is, producers of progressive trajectories rather than executors of predefined plans (4.1.4); coherent by statistical emergence, not by explicit theoretical structure, in a way analogous and not identical to the emergent systematicity of coordination classes (4.1.6); and whose output, including error, must be treated as interpretive data on the structure of the system, not as transparent access to it (4.1.7). With respect to **Sub-RQ1b** – *what LLMs, as a theoretical object, reveal about KiP* – the balance is equally precise. LLMs make empirically observable, in a non-human system, the plausibility of the KiP thesis that local elements lacking explicit theoretical structure can produce coherent performances (4.1.2;

4.1.6). They make more urgent, by contrast, the specificity of notions that were already present in KiP but sometimes in the background; noticing as an embodied act and not reducible to computational selection of relevance (4.1.3); language as the local trace of experiential activation, not as a field of self-sufficient regularities (4.1.4); grounding as a structural condition of understanding, not as an optional requirement added to performance (4.1.8). Finally, they make visible, by contrast, the difference between plausible linguistic production and stable conceptual understanding, a difference that KiP allows us to articulate through a specific theoretical and epistemological vocabulary and not as a simple summary description (4.1.8).

This analysis closes the first part of the chapter and opens onto the second research question (RQ2).

4.2. LLMs as tools for developing Knowledge Analysis

We now ask whether LLMs, in addition to functioning as a theoretical object useful for understanding KiP, can also become concrete tools for developing its research program. The thesis, as already stated in the introduction, does not empirically realize this application, but evaluates its theoretical possibility and identifies some specific directions for future research.

In order to answer RQ2, the aim is not merely to indicate generic directions of innovation, but to anchor the applications of the proposed comparison to concrete circumstances, specifically to the limits of the KiP paradigm on which LLMs can act. In Chapter 3, some general theoretical limits of KiP were discussed. In what follows, those limits are taken up again and integrated with other methodological nodes and tensions explicitly present in diSessa's work (1993), insofar as they become relevant for evaluating whether and how LLMs can contribute to the development of Knowledge Analysis (KA).

It is also necessary to distinguish between dialectical limits and limits intrinsic to the KiP paradigm.

4.2.1. Characterization of the limits

By dialectical limits, we mean those problems that emerge above all in the comparison between KiP and other models of learning and cognition, such as *Framework Theory*. They are important for situating KiP within the debate on conceptual change, but they constitute neither an exhaustive critique of the KiP paradigm nor the most adequate starting point for RQ2. If, in fact, LLMs were employed to respond to a limit formulated primarily from

outside the theory, as a comparison with other models, the risk would be to develop not so much KiP itself as a partiality of it produced within a specific theoretical comparison.

By intrinsic limits, instead, we mean those problems that emerge from the very structure of the KiP paradigm or that are recognized, at least in preliminary form, within diSessa's own reflection. In Chapter 3 (see Section 3.5), we considered as limits of KiP the categories called *pieces or coherence*, *prediction*, *timescales*, *in fraught with language* and *pedagogical challenges*.

If we consider, for example, the problem *pieces or coherence*, it can be ascribed to the category of dialectical limits. KiP describes intuitive knowledge as a system of local resources, sensitive to context and not organized from the beginning into a coherent global theory. Models such as Framework Theory, instead, attribute greater relevance to the stability of deep conceptual structures and to the presence of ontological presuppositions that constrain students' reasoning (Vosniadou, 2018). From this point of view, the granularity of KiP may appear as a difficulty in explaining the coherence and stability of intuitive knowledge.

However, this limit is dialectical because it depends to a large extent on the comparison with a different image of cognition. If it is formulated as an objection according to which KiP does not assume a global structure of knowledge analogous to that postulated by other models, then the problem does not simply concern what KiP is unable to do, but what it chooses not to assume as a starting point. KiP does not deny the possibility of forms of systematicity, but reformulates their status. Systematicity is not presupposed as an initial global architecture; rather, it is reconstructed through local recurrences, priorities, relations among resources, systematicities and, at a more stable level, coordination classes (diSessa, 1993; diSessa & Sherin, 1998). For this reason, the problem *pieces or coherence* could also be reformulated as an internal question to KiP, insofar as it concerns the passage from pieces to more stable forms of coordination. In this section, however, it will not be developed as an autonomous axis, because it would risk shifting RQ2 toward a discussion among paradigms rather than toward the internal development of KA.

An analogous consideration concerns *pedagogical challenges*. In this case too, the limit can be read in two ways. On the one hand, it can be considered a real limit of KiP, since a theory grounded in microgenetic analyses, local resources, and specific contexts does not immediately translate into extended curricular sequences, linear learning progressions, or

generalizable pedagogical indications. On the other hand, this difficulty becomes dialectical when KiP is asked to perform the function proper to models more oriented toward the construction of curricular trajectories or the description of conceptual structures relatively stable over time. The strength of KiP consists precisely in its capacity to analyze the fine grain of reasoning and to show how apparently naïve resources can be reorganized toward more expert forms of understanding (diSessa, 1993; diSessa & Sherin, 1998). If this same strength is evaluated starting from the demand to produce broad, linear, and generalizable pedagogical pathways, the limit that emerges is at least in part produced by the change of scale and function required of the theory.

For this reason, *pedagogical challenges* too will be kept in the background. They are not denied, nor considered irrelevant. They indicate an important question, namely the relation between fine-grained analysis of knowledge and pedagogical design. However, if they were placed at the center of the present section, the discussion would move from the possibility of developing KA as a research methodology to the broader question of curricular design. Such a direction could constitute a future development of the work, but it does not correspond to the minimal core of RQ2 as formulated in this thesis.

The other three – *prediction*, *timescales* and *in fraught with language* (or simply *the tension with language*) – are instead considered intrinsic limits, because they do not depend primarily on an external comparison with other models, but derive from the very way in which KiP constructs its theoretical object and practices KA; these will be the main object of attention in this chapter.

It should be noted that selecting the intrinsic limits of KiP as the main axis of analysis does not imply a reduction of complexity; on the contrary, it allows the discussion to remain on a methodologically controlled level. LLMs will not be employed, not even hypothetically, to decide between KiP, FT, or other models of cognition, nor to resolve from the outside the controversies of conceptual change. They will instead be considered as potential tools for expanding KA starting from its internal problems. In this sense, the research question is circumscribed not to whether LLMs can correct KiP from the outside, but to whether they can contribute to making some of its central processes more explorable, scalable, and controllable without replacing the theoretical interpretation of the researcher.

In what follows, we will address each of the intrinsic limits identified, evaluating the mode of insertion of LLMs.

4.2.2. Prediction and bootstrapping: LLMs as support for interpretive validation

The first intrinsic limit concerns the very structure of the method through which KA and KiP construct their hypotheses. diSessa (1993) explicitly recognizes that the theory is not able to specify a priori which p-prims a subject has developed; it is individual experiences, in their irreducible specificity, that determine the repertoire of resources that can be activated and are available. It follows that KA and KiP do not produce strong predictions independent (extrinsic) from the data (i.e., data-less prediction; diSessa, 1993). Identification and validation always occur a posteriori, through what diSessa calls bootstrapping and synthetic analysis, that is, a process in which the genesis of resources is assumed and these resources are then tested and refined in a circular and progressive dynamic on episodes and interview protocols (diSessa, 1993; diSessa, Sherin, & Levin, 2016).

The nuance and conceptual depth of synthetic analysis does not immediately emerge within diSessa's work (1993), but within the methodological framework of KA. diSessa, Sherin, and Levin (2016) connect it to the operations of *pattern finding* and *data reduction*. These expressions do not indicate, in KA, a standardized coding procedure analogous to those employed in other forms of qualitative or quantitative analysis. The authors in fact emphasize that the relation between pattern finding and data reduction in KA follows neither the model of verification of hypotheses fixed before data collection, nor that of correlational analysis grounded in quantitative parameters. Rather, *reduction* consists in the selection of relevant episodes and in the construction of theoretically oriented narratives, while pattern finding concerns the process through which the researcher identifies significant configurations in the data and tests them against the theory under construction (diSessa, Sherin, & Levin, 2016).

Since the identification of a resource does not proceed through the application of an already established code, but through an iterative movement among episodes, theoretical hypotheses, reduction of the data, and search for patterns, every analysis requires the construction of a specific argumentative chain. Such a chain must show why a certain limited set of clues, verbal, gestural, temporal, and contextual, can be interpreted as plausible evidence of the activation of a given resource. The result is that synthetic analysis (and, more generally, KA and KiP) produces very rich theoretical descriptions but descriptions that are relatively comparable with one another on a large scale, and that the work of interpretation

remains bound to the competence and intuition of the researcher, contextual to the process of analysis.

It is within this framework that LLMs can be considered methodologically relevant tools. They do not resolve the limit of prediction and bootstrapping at its conceptual root. They cannot produce a priori predictions that the theory does not authorize, and they cannot replace the researcher's theoretical judgment in the construction of the interpretive argument. However, they can intervene as support for synthetic analysis, with the aim of making more controllable the process through which a specific hypothesis of activation is supported starting from the data.

On the side of pattern finding, an LLM could help to explore verbal protocols and interview transcripts by identifying recurrences, structural similarities among episodes, variations in formulations, and possible recurring configurations of explanation. This would not amount to automatically identifying a p-prim, but it would allow the researcher to observe with greater breadth the empirical material on which to construct the interpretive argument. On the side of data reduction, an LLM could contribute to organizing large quantities of data into more manageable sets, signaling relevant passages, grouping similar episodes, and making more explicit the path that leads from the raw protocol to the formulation of the theoretical hypothesis.

This function becomes particularly important because synthetic analysis is, by its nature, a practice of high interpretive intensity. Its strength consists in keeping the analysis close to the complexity of the data; its cost consists in the fact that every reconstruction requires an articulated argumentative chain, which is difficult to transfer automatically to other subjects, contexts, or corpora. LLMs could reduce this cost without eliminating the interpretive nature of the method. They could help compare an episode with cases already analyzed in the literature, make explicit similarities and differences with respect to previous interpretations, identify potential counter-examples, and document which clues support a given hypothesis and which remain ambiguous.

In this sense, a possible development would not consist in simply employing generic LLMs, but in constructing models instructed or specialized on the theoretical vocabulary of KiP, on already validated studies, and on examples of KA conducted in the literature. Such specialization could constrain the exploration of protocols to theoretically relevant

categories, reducing the risk that the model produces linguistic associations lacking grounding in the paradigm.

4.2.3. The tension with language and large-scale validation

The second intrinsic limit we address concerns the relationship between KiP and language. This node touches the main empirical access point to cognitive resources, the elements of KiP.

We already know that p-prims are resources activated by experience and, furthermore, as diSessa argues, they do not coincide with the words or explicit concepts of ordinary lexicon. They are not strongly tied to a vocabulary (diSessa, 1993) and, for this reason, are difficult to translate into explicit verbal concepts. Yet, the KA approach almost always infers them starting from verbal, gestural, and interactional protocols. Language, in this sense, is the main trace, sometimes the only one available, of cognitive processes that by definition exceed it (diSessa, 1993; diSessa, Sherin, & Levin, 2016). This situation outlines an evident and structural methodological tension within KiP: on the one hand, p-prims do not coincide with language; on the other, language represents the main instrument for identifying them. Synthetic analysis, which we have already discussed, in fact uses as its guiding criterion the convergence of different clues (e.g., the form of verbal expression, gesture, hesitation), in order to reduce the risk of confusing a superficial linguistic regularity with evidence of deep cognitive activation. This procedure is practicable only on a very small scale, on a limited sample of data.

It is in this space that LLMs offer a new methodological perspective, already anticipated in 4.1.4 within the framework of the theoretical comparison. As we have seen, language contains semantic regularities sufficiently stable to be modeled (Jurafsky & Martin, 2026; Lewis, 2025; Bialek, 2025). Thus, students' linguistic practices (e.g., recurring metaphors, causal structures, spontaneous analogies, typical formulations of systematic errors) could contain statistically detectable traces of shared intuitive resources. Tools capable of exploring large corpora of these data could identify families of recurring formulations, group linguistically similar episodes, and signal patterns of co-occurrence between certain types of causal expression and certain types of conceptual error. It would not be a matter of identifying p-prims automatically, but of constructing provisional linguistic maps in a more systematized way, such as to orient the subsequent interpretive work of the researcher – that is, the problem of large-scale identification and validation.

The risk of this application, however, is symmetrical to its opportunity. LLMs are linguistic tools that operate on distributed representations of text, and they are structurally inclined to identify lexical and syntactic similarities. Two students who use apparently analogous lexical expressions (e.g., “the force runs out,” “the energy dissipates”) could activate different p-prims, while two students who activate the same p-prim could express it in lexically distant ways. The risk, therefore, is that the artificial system produces linguistically coherent but cognitively heterogeneous groupings. The expertise of the theoretically sufficiently equipped researcher remains an unavoidable prerogative of supervision.

4.2.4. Timescales: LLMs as support for connecting microgenesis and long-term learning

The third intrinsic limit concerns the temporal scale of the model. The analysis of microgenetics is the ground on which KiP has produced its most precise and best documented contributions (diSessa, 1993; diSessa & Sherin, 1998). The difficulty emerges when one attempts to connect this micro level with conceptual transformations that unfold over much longer times, that is, on temporal scales that the model itself is not structured to manage (diSessa, 1993; diSessa, Sherin, & Levin, 2016). The key concepts of *tuning toward expertise* (e.g., *shifting priority* and *context migration*), or also those – with the necessary conceptual distinctions – of *coordination class theory* (e.g., *span and alignment*, *readout strategy*), are not constructed within long-term temporal dynamics. In other words, they do not automatically explain how the reorganization of resources (tuning) and the stabilization of concepts (coordination class) persist over extended periods. Nevertheless, this set of concepts already constitutes a useful vocabulary for describing the outcomes toward which a learning trajectory may tend.

LLMs can offer a concrete contribution in relation to this open challenge. It is necessary to specify that this contribution does not represent a repetition of what was said in the previous paragraphs. Here, in fact, the issue is not the depth of analysis on a situated episode or on a synchronic corpus of data, but its extension over time. To this end, providing LLMs with longitudinal data corpora relating to students’ written productions, such as answers to open problems and explanations produced at successive moments of the same educational pathway, contributions to educational forums or learning platforms, offers material to which KiP has not traditionally had systematic access. Tools capable of analyzing these corpora could contribute to identifying whether certain argumentative patterns stabilize over time;

which strategies are used; how concepts align with one another; whether the same formulations recur in disciplinary contexts progressively farther from the original one (context migration); whether the appearance of new explanatory structures is accompanied by the disappearance or reduction in frequency of previously dominant formulations (shifting priority). It would not be a matter of directly measuring conceptual development (or learning), but of constructing a longitudinal framework of patterns compatible with certain resources, to be subsequently submitted to theoretical interpretation.

In other words, precisely because LLMs are particularly effective in language processing, their role does not consist in directly identifying cognitive resources, nor in automatically validating the existence of p-prims or coordination classes. They can, by using the vocabulary of KiP and KA, help explore linguistic regularities and, consequently, orient the irreplaceable attention of the researcher toward possible patterns of meaning that language (partially) encodes.

4.2.5. Runnability, Boundary Crossing and Future Perspectives

The three directions discussed so far do not exhaust the possible contribution of LLMs in expanding the framework of KiP and KA. In fact, merely sketching these three directions of application allows the general underlying question to emerge. This concerns the essence of RQ2, namely the possibility of making some aspects of the KiP model more explicit, systematic, and operational. One may refer to this question with the concept of *runnability*, generally defined as *computational executability* (diSessa, Sherin, & Levin, 2016).

The research articulated in this thesis intends, for this part, to lay the foundations for addressing the runnability of the theoretical framework of KiP, to be explored in future works.

Runnability does not constitute an external addition to KA but is contained within its very methodological approach. In Chapter 3 we saw KA situated within the development of *cognitive modeling*, interested in the construction of increasingly explicit descriptions of knowledge, its functioning, and its transformations, while not coinciding with the immediate construction of executable computational models. KiP has developed as a theory capable of describing the fine grain and complexity of knowledge (scientific intuitive knowledge). Modeling such complexity (through a *humble* approach) makes KiP a framework that is not directly runnable, but this perspective is not foreign to its research program and its extensions. The extension of the epistemology of physics proposed by diSessa (recalling the

title of KiP's foundational paper, *Towards an Epistemology of Physics*; diSessa, 1993) could be called a *computational epistemology*.

In this context, the comparison with LLMs offers unprecedented perspectives by providing an epistemological laboratory in which systemic properties such as emergence and distributed encoding become empirically treatable. The contribution of LLMs lies in their capacity to act as an artificial phenomenology that makes tangible the dynamics of sub-symbolic knowledge. These algorithms are based on a bottom-up learning paradigm that reflects the non-propositional and emergent nature of knowledge postulated by KiP. Again, the sensitivity to context of the resources of natural knowledge can be supported by the model of artificial knowledge of LLMs, whose probabilistic and non-deterministic approach represents the point of connection in this comparison.

This perspective is also connected to the theme of extending the KiP paradigm to other disciplinary domains, or *boundary crossing*, borrowing the term from Akkerman and Bakker (2011). KiP arises and develops with strength in the context of intuitive physics, especially in relation to the sense of mechanism and physical causality. However, the KA literature itself has progressively shown that its tools can be adapted to different domains, including mathematics, statistics, biology, chemistry, computer science, and social sciences (diSessa, Sherin, & Levin, 2016). LLMs could strengthen this possibility by allowing the exploration of large quantities of linguistic and argumentative productions in domains in which KiP has been less systematically applied.

Caution remains important. The computational systematization of KA must not turn into automation. Runnability should be understood as a methodological provocation, that is, as an invitation to ask which parts of KiP can be made more explicit and controllable, while nevertheless holding firm the principle according to which the cognitive meaning of the patterns identified always remains to be theoretically constructed through the primacy of the researcher's judgment.

In conclusion, LLMs can be assumed as tools capable of opening a new phase of reflection on KiP. They do not resolve the intrinsic limits of the paradigm, nor do they replace its interpretive work. They can, however, contribute to transforming such limits into research directions, showing how prediction and bootstrapping, tension with language, and timescales can become habitable places of methodological development. From here opens a broader perspective, oriented toward understanding which aspects of Knowledge Analysis can be

systematized, which can be made partially runnable, and which can sustain the extension of KiP beyond the original domain of physics.

Conclusions

We have seen how Large Language Models and Knowledge in Pieces represent distinct theoretical frameworks whose bidirectional comparison (Section 4.1) proves fruitful as a theoretical-epistemological deepening of each model. We have also described how LLMs can function as an epistemologically relevant tool for studying the problem of knowledge and, in particular, for expanding and developing the theoretical framework of KiP. This perspective has also made it possible to identify concrete and specific applicative horizons for the development of KiP.

Moreover, we have consciously left in the background a possible symmetrical direction. Just as in Section 4.1 the comparison between KiP and LLMs was constructed in bidirectional form, one could in fact hypothesize that RQ2, too, be developed in both directions, asking not only how LLMs can contribute to expanding Knowledge Analysis, but also how KiP can become an operational tool for developing the LLM paradigm. This possibility is not excluded in principle, but it does not constitute the main object of this section.

The reason is twofold. On the one hand, one of the original contributions of this thesis consists precisely in reconceptualizing LLMs not only as technical objects to be explained or evaluated, but as epistemological tools through which to interrogate human knowledge in a new way. In this sense, it was methodologically more coherent to concentrate the applicative part on the possible use of LLMs to extend a framework of natural cognition, rather than to develop symmetrically also the use of KiP for the design or improvement of artificial models. On the other hand, on the concrete level, the second direction appears less immediately viable within the limits of this work. KiP certainly offers a theoretical vocabulary useful for reading some aspects of LLMs, such as the distribution of representations, sensitivity to context, the locality of processes, and the distance between performance and understanding. However, transforming this vocabulary into a practical tool for developing the LLM paradigm would require a different research design, closer to computational modeling, system architecture, and engineering evaluation of performance. For this reason, this perspective is indicated as a possible future development.

This approach also makes it possible to specify the overall position of the thesis with respect to the contemporary debate on artificial intelligence. Merely affirming that LLMs are statistical systems, or, conversely, assuming them as proof of a new or superior form of knowledge, would not add much to the polarizations already present in the current discussion

and public debate. LLMs are certainly computational systems grounded in statistical learning, devoid of embodied experience and of direct access to the world. However, precisely this nature does not prevent them from being considered relevant cognitive tools. The point is not to attribute to them an understanding analogous to human understanding, but to recognize that they can become instruments of inquiry at our disposal for making certain aspects of human understanding more visible, for testing already existing frameworks, and for developing new questions about the nature of knowledge. From this perspective, the epistemological value of LLMs does not consist in replacing the theory of mind, learning, or understanding, but in forcing it to specify its own concepts more precisely. KiP, reread through the comparison with LLMs, is not overcome nor automated; rather, it is placed before the possibility of explicating, systematizing, and expanding some of its most fruitful intuitions, with a view to further positive developments in research.

References

Ackerman, C. (2025). Evidence for limited metacognition in LLMs. *arXiv preprint arXiv:2509.21545*. https://arxiv.org/abs/2509.21545

Akkerman, S. F., & Bakker, A. (2011). Boundary crossing and boundary objects. *Review of Educational Research*, *81*(2), 132–169. https://doi.org/10.3102/0034654311404435

American Society for Cybernetics. (n.d.). *Summary: The Macy Conferences*. https://asc-cybernetics.org/foundations/history/MacySummary.htm

Amin, T. G. (2025). Learning Scientific Concepts: A Unified Theory of Conceptual Change (1st ed.). Routledge. https://doi.org/10.4324/9780429505126

Amin, T. G., & Levrini, O. (Eds.). (2018). Converging perspectives on conceptual change: Mapping an emerging paradigm in the learning sciences. Routledge. https://doi.org/10.4324/9781315467139

Arons, A. (1998). Research in Physics Education: The Early Years. Paper presented at Physics Education Research Conference 1998, Lincoln, Nebraska. Retrieved May 27, 2026, from https://www.compadre.org/Repository/document/ServeFile.cfm?ID=11986&DocID=2780

Ashby, W. R. (1956). *An introduction to cybernetics*. New York: J. Wiley.

Bateson, G. (1977). *Verso un'ecologia della mente*. Adelphi. (Original work published in 1972).

Benanti, P. (2023). The urgency of an algorethics. Discover Artificial Intelligence, 3, Article 11. https://doi.org/10.1007/s44163-023-00056-6

Baron, S. (2025). Trust, explainability and AI. Philosophy & Technology, 38 (4). https://doi.org/10.1007/s13347-024-00837-6

Beichner, R. (2009). An Introduction to Physics Education Research. In Henderson, C., & Harper, K. (Eds.), Getting Started in PER (2). College Park: American Association of Physics Teachers. Retrieved May 26, 2026, from https://www.compadre.org/Repository/document/ServeFile.cfm?ID=8806&DocID=1147

Bender, E. M., Gebru, T., McMillan-Major, A., & 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) (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, *3*(Feb), 1137-1155. https://www.jmlr.org/papers/v3/bengio03a.html

Bengio, Y., Lamblin, P., Popovici, D., & Larochelle, H. (2007). Greedy layer-wise training of deep networks. In B. Schölkopf, J. Platt, & T. Hoffman (Eds.), Advances in Neural Information Processing Systems 19 (pp. 153–160). MIT Press.

Bialek, W. (2025). Emergence of brains. PRX Life, 3(3), Article 037002 . https://doi.org/10.1103/62p9-wn32

Bianchini, F. (2007). *L'IA e il linguaggio fra storia ed epistemologia*. In F. Bianchini, A. Gliozzo, M. Matteuzzi (Eds.), Instrumentum Vocale: Intelligenza Artificiale e Linguaggio.

Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.

Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., Lapata, M., Lazaridou, A., May, J., Nisnevich, A., et al. (2020). Experience grounds language. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 8718–8735. https://doi.org/10.18653/v1/2020.emnlp-main.703

Bocchi, G., & Ceruti, M. (Eds.). (1985). *La sfida della complessità*. Feltrinelli.

Boge, F. J., & Mosig, A. (2025). Put it to the test: Getting serious about explanation in explainable artificial intelligence. Minds and Machines , 35 (26). https://doi.org/10.1007/s11023-025-09724-1

Buchanan, B. (2006). *What Do We Know about Knowledge?* AI Magazine. 27. 35-46.

Burrell, J. (2016). How the machine “thinks”: Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1), 1–12.

Buschmeier, H., Buhl, H. M., Kern, F., Grimminger, A., Beierling, H., Fisher, J., Groß, A., Horwath, I., Klowait, N., Lazarov, S., Lenke, M., Lohmer, V., Rohlfing, K., Scharlau, I., Singh, A., Terfloth, L., Vollmer, A.-L., Wang, Y., Wilmes, A., & Wrede, B. (2023). Forms of Understanding of XAI-Explanations. In arXiv:2311.08760.

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K, Schwitzgebel, E., Simon, J., & VanRullen, R. (2023). Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv preprint arXiv:2308.08708.

Buzzi, E. & Priori, P. (2025). *Educazione e nuove tecnologie. La questione dell'umano.* 9788852608056. https://books.google.it/books?id=T86N0QEACAAJ. Itaca.

Carey, S. (1985). Conceptual change in childhood. MIT Press.

Ceravola, A., & Joublin, F. (2021). From JSON to JSEN through virtual languages. Open Journal of Web Technologies, 8(1), 1–15.

Ceruti, M. (1986). *Il vincolo e la possibilità*. Feltrinelli.

Chalmers, D. J. (2023). Could a Large Language Model be Conscious? Boston Review. (Based on a lecture presented at NeurIPS, 2022).

Chandrasekaran, V., & Jordan, M. I. (2013). Computational and statistical tradeoffs via convex relaxation. Proceedings of the National Academy of Sciences of the United States of America, 110(13), E1181–E1190.

Chi, M. T. H., Slotta, J. D., & De Leeuw, N. (1994). From things to processes: A theory of conceptual change for learning science concepts. Learning and Instruction, 4(1), 27–43. https://doi.org/10.1016/0959-4752(94)90017-5

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. arXiv preprint arXiv:1706.03741. https://doi.org/10.48550/arXiv.1706.03741

Clement, J. J. (1982). Students' preconceptions in introductory mechanics. American Journal of Physics, 50(1), 66–71. https://doi.org/10.1119/1.12989

Cristianini, N. (2024). *Machina sapiens*. Bologna: Il Mulino.

Crouch, C., Watkins, J., Fagen, A., & Mazur, E. (2007). Peer instruction: Engaging students one-on-one, all at once. Research-Based Reform of University Physics, 1, 1-55. https://doi.org/10.1119/RevPERv1.1.3

De Caro, M., & Giovanola, B. (2025). *Intelligenze. Etica e politica dell'IA*. il Mulino.

Demkanin, P., Cervenová, D., & Sands, D. (2025). Application of Selected Models of Mind in Theories of Physics Education. Journal of Baltic Science Education, 24(6), 1117-1131.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

diSessa, A. A. (1983). Phenomenology and the evolution of intuition. In D. Gentner & A. Stevens (Eds.), Mental models (pp. 15–33). Hillsdale, NJ: Lawrence Erlbaum.

diSessa, A. A. (1988). Knowledge in pieces. In G. Forman & P. Pufall (Eds.), Constructivism in the computer age (pp. 49–70). Hillsdale, NJ: Lawrence Erlbaum.

diSessa, A. A. (1991). Epistemological microworlds: The case of coordination and quantities.

diSessa, A. A. (1993). Toward an Epistemology of Physics. Cognition and Instruction, 10(2/3), 105–225. http://www.jstor.org/stable/3233725

diSessa, A. A. (2018). Knowledge in pieces: An evolving framework for understanding knowing and learning. In T. G. Amin & O. Levrini (Eds.), Converging Perspectives on Conceptual Change: Mapping an Emerging Paradigm in the Learning Sciences (pp. 9–16). Routledge.

diSessa, A. A., & Cobb, P. (2004). Ontological innovation and the role of theory in design experiments. Journal of the Learning Sciences, 13(1), 77–103.

diSessa, A. A., & Levin, M. (2021). Processes of Building Theories of Learning: Three Contrasting Cases. In: Levrini, O., Tasquier, G., Amin, T. G., Branchetti, L., & Levin, M. (Eds) Engaging with Contemporary Challenges through Science Education Research. Contributions from Science Education Research, vol 9. Springer, Cham. https://doi.org/10.1007/978-3-030-74490-8_18

diSessa, A. A., & Sherin, B. L. (1998). What changes in conceptual change? *International Journal of Science Education, 20*(10), 1155–1191. https://doi.org/10.1080/0950069980201002

diSessa, A. A., Sherin, B., & Levin, M. (2016). Knowledge Analysis: An Introduction.

Docktor, J. L., & Mestre, J. P. (2014). Synthesis of discipline-based education research in physics. Physical Review Special Topics - Physics Education Research, 10(2), Article 020119. https://doi.org/10.1103/PhysRevSTPER.10.020119

Domingos, P. (2020). *L'algoritmo definitivo: La macchina che impara da sola e il futuro del nostro mondo* (A. Migliori, Trans.). Bollati Boringhieri. (Original work published 2015).

Dreyfus, H. L. (2006). Overcoming the myth of the mental. Topoi 25:43-49. https://doi.org/10.1007/s11245-006-0006-1

Driess, D., Xia, F., Sajjadi, M. S. M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., et al. (2023). Palm-e: An embodied multimodal language model. arXiv preprint arXiv:2303.03378.

Durán, J. M., & Pozzi, G. (2025). Trust and trustworthiness in AI. *Philosophy & Technology*, *38*(1), Article 16. https://doi.org/10.1007/s13347-025-00843-2

Floridi, L. (2025). AI as Agency without Intelligence: On Artificial Intelligence as a New Form of Artificial Agency and the Multiple Realisability of Agency Thesis. Philosophy & Technology, 38(30). https://doi.org/10.1007/s13347-025-00858-9

Franks, C. (2024). Propositional logic. In E. N. Zalta & U. Nodelman (Eds), *The Stanford Encyclopedia of Philosophy* (Winter 2024 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2024/entries/logic-propositional/

Gahrn-Andersen, R. (2025). Beyond symbol processing: the embodied limits of LLMs and the gap between AI and human cognition. AI & SOCIETY, 40, 3105–3107.

Girdhar, Rohit, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. (2023). ImageBind: One embedding space to bind them all. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15180–15190. https://doi.org/10.1109/CVPR52729.2023.01457

Godfrey-Smith, P. (2003) Theory and Reality: An Introduction to the Philosophy of Science. The University of Chicago Press, Chicago. https://doi.org/10.7208/chicago/9780226300610.001.0001

Goertzel, B., & Pennachin, C. (Eds.). (2007). *Artificial general intelligence*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-68677-4

Goldstein, A., Ham, E., Schain, M., Zada, D., Meshulam, M., Gvirts, H. Z., Dikker, S., Melloni, L., Quian Quiroga, R., Fedorenko, E., Knight, R. T., & Hasson, U. (2025). Temporal structure of natural language processing in the human brain corresponds to layered hierarchy of large language models. Nature Communications, 16(1), Articolo 10529. https://doi.org/10.1038/s41467-025-65518-0

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes,

N. Lawrence, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 27* (pp. 2672–2680). Curran Associates, Inc.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press. http://www.deeplearningbook.org

Griot, C., Hemmer, P., Pammer-Schindler, V., & Baier, L. (2025). Large language models lack essential metacognition for reliable medical reasoning. Nature Communications, 16, Article 283. https://doi.org/10.1038/s41467-024-55628-6

Gubrud, M. A. (1997, November). Nanotechnology and international security [Paper presentation]. Fifth Foresight Conference on Molecular Nanotechnology, Palo Alto, CA, United States. http://www.foresight.org/Conferences/MNT05/Papers/Gubrud/

Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems, 24*(2), 8–12.

Halliday, D., Resnick, R., & Walker, J. (1997). *Fundamentals of physics* (5th ed.). John Wiley & Sons.

Halloun, I. A., & Hestenes, D. (1985). The initial knowledge state of college physics students. American Journal of Physics, 53(11), 1043–1055. https://doi.org/10.1119/1.14030

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

Harnad, S. (2002). Symbol grounding and the origin of language. In M. Scheutz (Eds.), *Computationalism: New directions* (pp. 143–158). MIT Press.

Heidegger, M (2010). Being and time. SUNY Press, Albany.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. The Physics Teacher, 30(3), 141–158. https://doi.org/10.1119/1.2343497

Hila, A. (2025). The Epistemological Consequences of Large Language Models: Rethinking Collective Intelligence and Institutional Knowledge. University of Texas at Austin.

Himma, K. E. (2009). Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? Ethics and Information Technology, 11, 19–29.

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. Neural Computation, 18(7), 1527–1554. https://doi.org/10.1162/neco.2006.18.7.1527

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. arXiv preprint arXiv:2311.05232. https://arxiv.org/abs/2311.05232

Huang, Shaohan, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. (2023). Language is not all you need: Aligning perception with language models. arXiv preprint arXiv:2302.14045.

Ilkou, E., & Koutraki, M. (2020). Symbolic vs. sub-symbolic AI methods: Friends or enemies? *Proceedings of the CIKM 2020 Workshops*.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Chen, D., Dai, W., Chan, H. S., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), Article 248. https://doi.org/10.1145/3571730

Jones, M. L. (2018). How we became instrumentalists (again): Data positivism since World War II. *Historical Studies in the Natural Sciences*, *48*(5), 673–684. https://doi.org/10.1525/hsns.2018.48.5.673

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255–260.

Jurafsky, D., & Martin, J. H. (2026). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models* (3rd ed.). https://web.stanford.edu/~jurafsky/slp3/

Kaplan, J. (2018). *Intelligenza artificiale: guida al futuro prossimo* (2. ed.). Luiss University Press.

Keil, F. C. (1989). Concepts, kinds, and cognitive development. MIT Press.

Kim, J., Maathuis, H., & Sent, D. (2024). Human-centered evaluation of explainable AI applications: A systematic review. Frontiers in Artificial Intelligence, 7, Article 1456486. https://doi.org/10.3389/frai.2024.1456486

Krakowski, S. (2025). *Human-AI agency in the age of generative AI.* Information and Organization, 35, 100560, 10.1016/j.infoandorg,2025, 100560.

Laird, J. E., Newell, A., & Rosenbloom, P. S. (1987). Soar: An architecture for general intelligence. *Artificial Intelligence, 33*(1), 1–64. https://doi.org/10.1016/0004-3702(87)90050-6

LeCun, Y. (1985). Une procédure d'apprentissage pour réseau à seuil assymétrique [A learning procedure for asymmetric threshold networks]. Cognitiva 85: À la frontière de l'intelligence artificielle, des sciences de la connaissance et des neurosciences, 599–604.

LeCun, Y. (2022). A path towards autonomous machine intelligence (Version 0.9.2). Open Review, 62(1), 1–62.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539

Levrini, O., & diSessa, A. A. (2008). How students learn from multiple contexts and definitions: Proper time as a coordination class. Physical Review Special Topics—Physics Education Research, 4(1), Article 010107. https://doi.org/10.1103/PhysRevSTPER.4.010107

Levrini, O., Tasquier, G., Amin, T. G., Branchetti, L., & Levin, M. (Eds.). (2021). Engaging with contemporary challenges through science education research: Selected papers from the ESERA 2019 conference. Springer. https://doi.org/10.1007/978-3-030-74490-8

Lewis, C. (2025). *Artificial psychology: Learning from the unexpected capabilities of large language models*. Springer. https://doi.org/10.1007/978-3-031-76646-6

Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM, 61*(10), 36–43.

Loru, E., Nudo, J., Di Marco, N., Santirocchi, A., Atzeni, R., Cinelli, M., Cestari, V., Rossi-Arnaud, C., & Quattrociocchi, W. (2025). The simulation of judgment in LLMs. Proceedings of the National Academy of Sciences (PNAS). https://doi.org/10.1073/pnas.2518443122

Maes, L., Le Lidec, Q., Scieur, D., LeCun, Y., & Balestriero, R. (2026). LeWorldModel: Stable end-to-end joint-embedding predictive architecture from pixels. Preprint.

Mazur, E. [Serious Science]. (2014). Peer instruction for active learning [Video]. YouTube. https://www.youtube.com/watch?v=Z9orbxoRofI

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. AI Magazine, 27(4), 12–14. https://doi.org/10.1609/aimag.v27i4.1904

McCulloch, W. S., & Pitts, W. (1990). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biology*, *52*(1-2), 99–115. https://doi.org/10.1016/S0092-8240(05)80006-0 (Original work published 1943).

McDermott, L. C. (1984). Research on conceptual understanding in mechanics. Physics Today, 37(7), 24–32. https://doi.org/10.1063/1.2916318

Menabrea, L. F. (1843). Sketch of the Analytical Engine invented by Charles Babbage. (A. A. Lovelace, Trans. & Notes). Scientific Memoirs, 3.

Mienye, I. D., & Swart, T. G. (2024). A comprehensive review of deep learning: Architectures, recent advances, and applications. *Information, 15*(12), 755. https://doi.org/10.3390/info15120755

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. https://arxiv.org/abs/1301.3781

Minsky, M., & Papert, S. A. (1969). Perceptrons: An introduction to computational geometry. MIT Press.

Mollo, D. C., & Millière, R. (2023). The vector grounding problem. arXiv preprint arXiv:2304.01481.

Morin, E. (1990). *Introduction à la pensée complexe*. ESF.

Mugleston, J., Truong, V. H., Kuang, C., Sibiya, L., & Myung, J. (2025). Epistemology in the age of large language models. Knowledge, 5(1), 3. https://doi.org/10.3390/knowledge5010003

Naeem, S., Ali, A., Anam, S., & Ahmed, M. M. (2023). An unsupervised machine learning algorithms: Comprehensive review. International Journal of Computing and Digital Systems, 13(1), 911-921.

Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, *19*(3), 113–126. https://doi.org/10.1145/360018.360022

Nicolis, G., & Prigogine, I. (1977). Self-organization in nonequilibrium systems. Wiley.

Ouyang, L., Lowe, J., Williams, M., Lukaszewicz, B., Chordia, G., Gylfe, E., Denison, J., Reynolds, J., Kleinman, B., Shen, L., Carter, C., Khan, B., Dhariwal, P., Tobin, J., Yin, Q., Hemady, C., McMillan, P., Mishkin, P., Zhao, L., ... & Ziegler, D. M. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730–27744.

https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be36d956074ca93ef3169462-Abstract-Conference.html

Parker, D. B. (1985). Learning logic (Technical Report TR-47). MIT Press.

Peirce, C. S. (1934). *Collected papers of Charles Sanders Peirce*. Harvard University Press.

Pias, C. (Ed.). (2003). *Cybernetics | Kybernetik: The Macy-Conferences 1946–1953* (Vol. 1). diaphanes.

Pias, C. (ed.) (2016). Cybernetics: The Macy Conferences 1946-1953. The Complete Transactions. Diaphanes Publishing House.

Pintrich, P. R., & Boyle, R. A. (1993). Beyond cold conceptual change: The role of motivational beliefs and classroom contextual factors in the process of conceptual change. Review of Educational Research, 63(2), 167–199. https://doi.org/10.3102/00346543063002167

Polverini, G., & Gregorcic, B. (2024). How understanding large language models can inform the use of ChatGPT in physics education. European Journal of Physics, 45(2), Article 025701. https://doi.org/10.1088/1361-6404/ad1420

Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. Science Education, 66, 211-227.

Press, O., Smith, N. A., & Lewis, M. (2021). *Train short, test long: Attention with linear biases enables input length extrapolation*. arXiv. https://arxiv.org/abs/2108.12409

Prigogine, I., & Stengers, I. (1984). *Order out of chaos*. Bantam Books.

Radford, A., Kim, J., W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I., (2021). Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, pages 8748–8763.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review, 65*(6), 386–408.

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. Cognitive Science, 26(5), 521–562. https://doi.org/10.1207/s15516709cog2605_1

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323(6088), 533–536.

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. SN Computer Science, 2(160).

Shanahan, M. (2023). Talking about large language models. Communications of the ACM, 67(2):68–79. https://doi.org/10.1145/3624724

Shannon, C. E. (1948). *A mathematical theory of communication*. Bell Systems Technical Journal, 27, 379-423, 623-656.

Sherin, B. L. (2006). Common sense clarified: The role of intuitive knowledge in physics problem solving. Journal of Research in Science Teaching, 43(6), 535–555. https://doi.org/10.1002/tea.20136

Sherin, B. (2013). A computational study of commonsense science: An exploration in automated analysis of clinical interview data. Journal of the Learning Sciences, 22(4), 600-638. https://doi.org/10.1080/10508406.2013.836654

Sherin, B. (2021). Where are we? Syntheses and synergies in science education research and practice. In O. Levrini, G. Tasquier, T. G. Amin, L. Branchetti, & M. Levin (Eds.), Engaging with Contemporary Challenges through Science Education Research: Selected Papers from the ESERA 2019 Conference (pp. 211–223). Springer. https://doi.org/10.1007/978-3-030-74490-8_17

Singh, G., Banerjee, T., & Ghosh, N. (2025). *Tracing the evolution of artificial intelligence: A review of tools, frameworks, and technologies (1950–2025)*. Preprints.org. https://doi.org/10.20944/preprints202511.0637.v17

Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B., & Liu, Y. (2021). RoFormer: Enhanced transformer with rotary position embedding. arXiv. https://arxiv.org/abs/2104.09864

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. arXiv. https://doi.org/10.48550/arXiv.1409.3215

Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction (2nd ed.). MIT Press.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to Grow a Mind: Statistics, Structure, and Abstraction. Science, 331(6022), 1279–1285. https://doi.org/10.1126/science.1192788

Tian, C., Cheng, T., Peng, Z., Zuo, W., Tian, Y., Zhang, Q., Wang, F.-Y., & Zhang, D. (2025). A survey on deep learning fundamentals. *Artificial Intelligence Review*, *58*, 381. https://doi.org/10.1007/s10462-025-11368-7

Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., & Xie, S. (2024). Eyes wide shut? Exploring the visual shortcomings of multimodal LLMs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9568–9578.

Toulmin, S. (1972). Human understanding. Princeton University Press.

Turing, A. M. (1937). On Computable Numbers, with an Application to the Entscheidungsproblem. Proceedings of the London Mathematical Society, s2-42: 230-265. https://doi.org/10.1112/plms/s2-42.1.230

Turing, A. M. (1950). Computing Machinery and Intelligence. Mind 49: 433-460.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Kaiser, Ł. (2017). *Attention is all you need*. arXiv. https://arxiv.org/abs/1706.03762

von Bertalanffy, L. (1968). *General system theory*. George Braziller.

Vosniadou, S. (1994). Capturing and modeling the process of conceptual change. Learning and Instruction, 4(1), 45–69. https://doi.org/10.1016/0959-4752(94)90018-3

Vosniadou, S. (2018). Initial and scientific understandings and the problem of conceptual change. In T. Amin & O. Levrini (Eds.), Converging perspectives on conceptual change. Routledge.

Vosniadou, S., & Brewer, W. (1992). Mental models of the earth: A study of conceptual change in childhood. Cognitive Psychology.

Vosniadou, S., & Skopeliti, I. (2014). Conceptual change from the framework theory side of the fence. Science & Education, 23(7), 1427–1445.

Wang, Z., Wu, Z., Zhang, J., Guan, X., Jain, N., Lu, S., Gupta, S., & Koshiyama, A. S. (2024). Bias amplification: Large language models as increasingly biased media. arXiv preprint arXiv:2410.15234. https://arxiv.org/abs/2410.15234

Weaver, W. (1948). Science and complexity. *American Scientist, 36*(4), 536–544.

Werbos, P. J. (1974). Beyond regression: New tools for prediction and analysis in the behavioral sciences [Doctoral dissertation, Harvard University].

Wiener, N. (1948). Cybernetics: Or Control and Communication in the Animal and the Machine. MIT Press. (Edizione italiana: La cibernetica. Controllo e comunicazione nell'animale e nella macchina, Il Saggiatore, 1982). d

Winograd, T. (1971). Procedures as a representation for data in a computer program for understanding natural language (AI Technical Report 235). Massachusetts Institute of Technology, Artificial Intelligence Laboratory. https://dspace.mit.edu/handle/1721.1/7095

Wolfram, S. (2023). "What Is ChatGPT Doing ... and Why Does It Work?," Stephen Wolfram Writings. writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work. doi: 10.31855/bc47ee6b-75c

Zahavy, T. (2026). *LLMs can't jump*. PhilSci-Archive. https://philsci-archive.pitt.edu/28024/

Zhao, J., Chu, F., Xie, L., Che, Y., Wu, Y., & Burke, A. F. (2026). A survey of transformer networks for time series forecasting. *Computer Science Review*, *60*, Article 100883. https://doi.org/10.1016/j.cosrev.2025.100883

Acknowledgement

I would like to express my deepest gratitude to my supervisor, Professor Olivia Levrini, for her constant availability and invaluable guidance not only throughout this thesis work but also during my entire academic journey. More than anyone else, she opened my eyes during these years, giving me the opportunity to better discover my own way of understanding physics and to contribute to research through my abilities and gifts. She has been a mentor in the deepest sense of the word.

I would also like to thank Dr. Orso Peruzzi from IFAB International Foundation for his technical expertise and his help with the revision of this work.